Benchmarking mortality risk prediction from electrocardiograms

Platon Lukyanenko[a,c], **Joshua Mayourian**[b,c], **Mingxuan Liu**[a,d]
John K. Triedman[b,c], **Sunil J. Ghelani**[b,c], **William G. La Cava**[a,c,*]
[a]Computational Health Informatics Program, Boston Children's Hospital
[b]Department of Cardiology, Boston Children's Hospital
[c]Harvard Medical School
[d]National University of Singapore

Abstract

Several recent high-impact studies leverage large hospital-owned electrocardiographic (ECG) databases to model and predict patient mortality. MIMIC-IV, released September 2023, is the first comparable public dataset and includes 800,000 ECGs from a U.S. hospital system. Previously, the largest public ECG dataset was Code-15, containing 345,000 ECGs collected during routine care in Brazil. These datasets now provide an excellent resource for a broader audience to explore ECG survival modeling. Here, we benchmark survival model performance on Code-15 and MIMIC-IV with two neural network architectures, compare four deep survival modeling approaches to Cox regressions trained on classifier outputs, and evaluate performance at one to ten years. Our results yield AUROC and concordance scores comparable to past work (circa 0.8) and reasonable AUPRC scores (MIMIC-IV: 0.4-0.5, Code-15: 0.05-0.13) considering the fraction of ECG samples linked to a mortality (MIMIC-IV: 27%, Code-15: 4%). When evaluating models on the opposite dataset, AUROC and concordance values drop by 0.1-0.15, which may be due to cohort differences. All code and results are made public.

1 Introduction

Electrocardiography (ECG) is a measure of the electric activity of the heart and is a staple of cardiac diagnostics. There has been a recent surge of research applying deep learning techniques to ECG data to predict a broader set of patient outcomes than its standard use, including risk of mortality [1–5], or ventricular hypertrophy [6] and dysfunction in adults [7] and children [8].

These studies demonstrate the potential value ECG modeling may have in patient care. However, there is no well-established consensus on the appropriate modeling approach for these tasks, e.g. the choice of model architecture, the survival modeling approach, the evaluation time horizons, and the methods for evaluating a model's ability to generalize to new and diverse patient populations.

Answering many of these technical questions regarding ECG modeling has historically been limited by a lack of access to large, publicly available ECG databases that are linked to high-quality patient-level data. Many studies instead rely on private data resources and provide details on results for only a limited set of modeling approaches.

Fortunately, access to data of this type has recently changed with the public releases of the MIMIC-IV ECG dataset (n $\approx$ 800k) in 2023 [9–12] and the Code-15 dataset (n $\approx$ 345k) in 2020 [13].

*Correspondence: william.lacava@childrens.harvard.edu. Lab page: https://cavalab.org

Preprint. Under review.

Given these new data resources and the high interest in risk prediction from ECGs, a benchmark resource that establishes the baseline performance of several deep learning architectures, survival approaches and experimental conditions could help support future research. To this end, our work establishes a benchmark of survival prediction models using ECG data. We measure the ability of these models, that vary by architecture and survival task definition, to estimate mortality risk at several time horizons and across care centers. We make this resource publicly available at github.com/cavalab/ecg-survival-benchmark.

2 Background

ECG ECG was developed in the 19th and early 20th century. Abnormal ECGs often indicate cardiovascular pathology and are thus a marker for disease and mortality. ECGs are recorded as time series and modern ECG machines provide standard measures of timing and voltage, often proposing algorithmically predicted diagnoses. ECG typically refers to 12-channel ECG, where the first six channels come from electrodes placed on the limbs and the second six come from electrodes placed circularly around the left half of a patient's chest. The first six channels are typically calculated from leads I and II [14]. ECG signals are typically in the 40-150Hz range, and are often affected by baseline drift, muscle activity, or sensor misplacement; 60Hz filters are often employed to reduce noise. ECGs can be recorded in many settings, so a dataset's context can be an important indicator of the likelihood of various outcomes (e.g., mortality is likely more prevalent in high-acuity settings like the intensive care unit). AI-ECG is the common term for the application of AI or machine learning to ECG [3]. A common AI-ECG task is risk stratification, which is equivalent to predicting event occurrence (e.g., mortality or passing a diagnostic threshold).

ECG databases Many hospital systems maintain their own ECG banks, and recent papers report training on hundreds of thousands to millions of privately-held samples (ex: 2.3M ECGs from a health system in Pennsylvania [4]; 1.5M ECGs from a hospital system in Alberta, Canada [5]; or a 600k ECGs from the Neihu and Zhonzheng Districts in China [15]). As these datasets are proprietary, public datasets are important for research and creating models that can then be fine-tuned for niche applications [16, 1, 17]. Since 2021, several large publicly accessible annotated ECG datasets have become available: the public portion of the Code dataset with 345,000 ECGs [13], and just recently the MIMIC-IV ECG dataset with 800,000 ECGss [9]. To our knowledge, these are the largest public ECG datasets. See [18] for a list of public datasets pre-2020.

Survival modeling Survival modeling builds survival functions, $S(t)$, that denote the probability of not having experienced an event by a particular time, t. Sometimes the cumulative hazard function, $H(t) = 1 - S(t)$, or hazard function, $h(t) = \frac{d}{dt}\left(H(t)\right)$, is modeled instead. Survival models use data in the format of {Time-To-Event, Event}, where an 'Event' equals 1 when a subject (patient, device, etc.) experiences an outcome (fatality, breakdown, etc.) and equals 0 if the subject leaves the study prior to an event (censorship). Survival modeling differs from classification in that it uses information from subjects that leave the study before (or after) a time horizon of interest: a device that works for two years before being lost (censored) still provides two years of evidence of non-breakage.

The most common survival functions are Kaplan-Meier curves which display survival over time from a tracked population [19]. If the population is clustered into groups, an individual's trajectory can be estimated from their cluster's survival function. Otherwise, survival is usually approximated by exponential decay or Weibull function fitted by regressors (demographics, measurements, etc.).

The Cox proportional hazards model [20] (hence, CoxPH) models hazard as an unknown base function scaled by exponential decays, i.e.

$$h(t) = h_0(t)e^{(\beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_m x_m)}.$$

Here, $\{x_1, \ldots, x_m\}$ are regressors (i.e. features) and $\{\beta_1, \ldots, \beta_m\}$ are learned weights. The cumulative hazard for two inputs is then 'proportional' in that their ratio is fixed over time. This allows the comparison of subject groups with different regressors without explicitly modeling the survival function and the easy addition (or accounting-for) of factors like demographics. The Cox baseline hazard, $h_0(t)$, can be fit from training data via the Breslow estimator. See [21] for more discussion on Breslow estimators and [22] for a thorough guide to survival modeling.

Table 1: Population characteristics of the Code-15 (top) and MIMIC-IV (bottom) cohorts, grouped by all-time mortality. p-values denote Student's t-tests for continuous variables and χ^2 tests for categorical variables, with Bonferroni correction for multiple comparisons.

Code-15		Overall	False	True	p-value
ECGs, n		345779	331949	13830	
Patients, n		233770	225429	8341	
Age in years, mean (SD)		53.2 (19.7)	52.5 (19.6)	71.5 (13.2)	<0.001
Sex, n (%)	F	206576 (59.7)	200228 (60.3)	6348 (45.9)	<0.001
	M	139203 (40.3)	131721 (39.7)	7482 (54.1)	
Follow-up years, mean (SD)		3.7 (1.9)	3.7 (1.9)	2.0 (1.6)	<0.001
MIMIC-IV		Overall	False	True	p-value
ECGs, n		795546	576898	218648	
Patients, n		159608	134428	25180	
Age, mean (SD)		64.3 (17.1)	61.1 (17.1)	72.5 (14.1)	<0.001
Sex, n (%)	F	389335 (48.9)	288676 (50.0)	100659 (46.0)	<0.001
	M	406211 (51.1)	288222 (50.0)	117989 (54.0)	
Follow-up years, mean (SD)		2.6 (2.6)	2.9 (2.6)	2.0 (2.4)	<0.001

Deep survival models A straightforward way to integrate neural networks and survival models is to build a neural network classifier and then to feed the output through a Cox regression to generate a survival function. In contrast, deep survival models run "end-to-end", optimizing survival-specific loss functions, and yield survival functions without an intermediate classification output. In this work we use four deep survival modeling approaches from PyCox [23]: DeepSurv [24], LogisticHazard (also known as nnet-survival) [25], Multi-Task Logistic Regression [26], and DeepHit [27]. The first does not discretize hazard per time bin, while the latter three do.

DeepSurv minimizes a regularized Cox partial likelihood, which depends on which patients are at risk at each time point. LH minimizes the negative log likelihood of each subject experiencing an event at their Time-to-Event, eliminating the loss function's dependence on other individuals and, theoretically, allowing independence from batch size effects. DeepHit is built for multi-event prediction and optimizes the joint distribution of the first time and event across all subjects per time point. MTLR fits a logistic regression per time bin. These models were chosen for their variety of approaches and standardized implementation.

AI-ECG survival models Several recent studies have predicted all-cause mortality from ECGs. These are either built completely on large ECG banks [4, 28, 15, 5, 2] or are fine-tuned in the context of a particular setting or population [16, 1]. There are two general approaches: modeling survival functions directly with deep survival models [15, 2], or training a classifier and then following with a Cox regression fitted to classifier output [28, 4, 5]. Most studies use a convolutional network, with a slightly-more-popular choice being the architecture from Ribeiro et al. [13, 28, 5].

3 Methods

In this work we benchmark mortality prediction approaches on two ECG datasets (Code-15 and MIMIC-IV). We use two architectures: InceptionTime [29] and a modified Resnet architecture [13]. For each architecture, we train four models that directly build survival functions with the PyCox package [23], as well as four binary classifiers predicting mortality at the one, two, five, and ten-year horizons. The classifiers are then converted to survival functions by fitting Cox regressions. Evaluation is based on survival function performance at the one, two, five, and ten-year horizons.

3.1 Benchmark datasets

We use two publicly available ECG databases to benchmark mortality prediction: Code-15 [13] and MIMIC-IV [9] datasets. The population characteristics for both databases are shown in Table 1.

Code-15 Code-15 is a collection of ECGs obtained by the Telehealth Network of Minas Gerais, a Brazilian public agency that provides telehealth services to the state Minas Gerais as well as Amazonian and Northeast states. Patient ECGs were recorded in primary care facilities by technicians and the resulting tests were examined by a cardiologist remotely. The portion of data made available [13] consists of roughly 345,000 ECGs collected from April to September of 2018. Code15 ECGs are a 7-10 second signal sampled at 400Hz, centered and padded with zeros to total a length of 4096. On visual inspection, many ECGs show baseline drift or movement artifacts.

MIMIC-IV The MIMIC-IV dataset includes roughly 800,000 ECGs from 160,000 patients collected between 2008-2019 at the Beth Israel Deaconess Medical Center in Boston, Massachusetts. Patient ECGs were recorded at various settings, including emergency settings, hospitals, and outpatient care centers. MIMIC-IV ECGs are 10 second signals sampled at 500Hz. Many samples include details on noise filtering such as 60Hz and harmonic filters.

3.2 Dataset processing

Dataset preparation is illustrated in Fig. 1 and detailed for the two care centers below. Data is limited to [Patient ID, ECG ID, ECG, Time-To-Event, Event], where Event = 1 if the patient dies (PyCox model: at all; Classifier: within the time horizon), else 0. Time-To-Event is set to a minimum of 1 day. The data is split 64/16/20 into train/validation/test sets randomly by patient ID.

Code-15 data selection and labels The Code-15 Dataset provides multiple ECGs per subject, indicating 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. While one could bound the Time-To-Event for the subject's other ECGs, labeling these points is not trivial and introduces uncertainty in event timing relative to the other samples. Overall, we kept 233,647 of the original 345,779 ECGs; 1.23% / 2.07% / 3.36% / 3.61% of ECGs are associated with a mortality by year 1 / 2 / 5 / 7.67(max).

MIMIC-IV data selection and labels MIMIC-IV date-of-death is tracked using state and hospital records, and deaths one year after a final recorded hospital visit are censored. ECGs with `nan` entries were excluded. Overall, we kept 785,035 of the original 795,546 ECGs; 14.7% / 18.3% / 24.5% / 27.6% / 27.8% of ECGs are associated with a mortality by year 1 / 2 / 5 / 10 / 12.97(max).

ECG processing MIMIC-IV ECGs were resampled to match the Code-15 format: 10-seconds at 400-Hz, padded with 48 starting/trailing zeros for a total length of 4096. All ECGs were z-score normalized per ECG channel based on the model's training set. All 12 ECG channels, despite the four channel redundancy in the first six, were kept. While ECGs may have been filtered (e.g. 60Hz) during acquisition, we did not apply any filtering. The end ECG shape per sample was 4096 x 12.

3.3 Modeling Process

The modeling process (Fig. 1 and Table 2) differs between the four PyCox models and the four classifiers: classifiers compress data to a binary label relevant to their time horizon and classifier outputs are fed into a Cox regression to generate survival functions before evaluation. PyCox models train on the [Time-To-Event, Event] format and yield survival functions.

Architectures We benchmark two architectures that have shown good performance for time series classification. The first is a modified residual network ("Resnet") that adapts the original ResNet from [30] to ingest multi-channel time series [13, 28]. The Resnet architecture opens with a convolutional layer and then five residual blocks containing two convolution layers, the first followed by a ReLu and the second by a dropout layer and skip connections, before ending in a fully connected layer. Batch-normalization follows each convolutional layer. Architecture parameters such as starting kernel size and filter counts were kept at the defaults tuned to the larger Code dataset.

The second architecture is InceptionTime [29] ("Inception"), which performs well on a broad set of small time series classification benchmarks and has recently been used successfully for fetal heart rate monitoring [31]. InceptionTime is modeled after AlexNet [32], including convolutional layers and skip connections, but differs in that it stacks several convolutional layers with very wide kernel widths, assisted by channel-wise bottlenecks that control overall model complexity.

Table 2: Experimental parameters used to benchmark mortality risk prediction.

Setting	Values
Architectures	InceptionTime [29] (kernel width: 11, 21, 41; no. parameterss: 510-530k), Resnet [13] (kernel width: 17; no. parameters: 6.9-7.5M)
Normalization	z-score per channel, based on model's training data
Classifier horizon (years)	1, 2, 5, 10
Deep survival method	DeepSurv [24], LogisticHazard [25], Multi-Task Logistic Regression [26], DeepHit [27]
Time segmentation	100 segments set by validation data

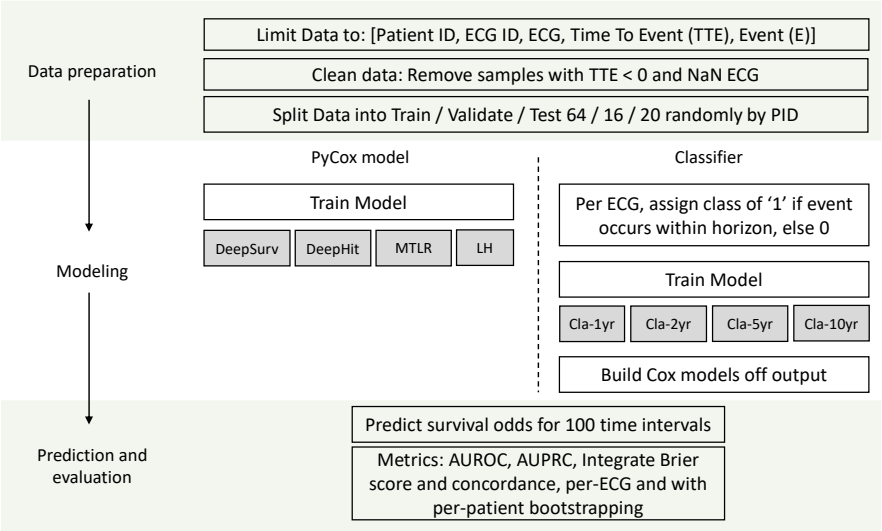


Figure 1: A flow diagram shows the benchmarking process from data preparation to model evaluation.

Training Models train for up to 200 epochs and stop early if their performance on the validation set does not improve for 25 consecutive epochs. We use the AdamW optimizer, start with a 1e-3 learning rate, and scale learning rate by 0.1x if no improvement is seen for 10 consecutive epochs to a minimum of 1e-8[13]. This mimics training regimes in other work[3]. We run five random seeds per model. Models were trained on a cluster with a combination of A40, A100, L40 and Nvidia Quadro GPUs. All data was loaded to RAM, requesting 100GB for Code-15 and 300GB for MIMIC-IV. Training was done on batches of 512 shuffled data samples with 256 loaded onto a GPU at a time. Some PyCox loss functions require batches to contain events for loss calculations; when training PyCox (not classifier) models, GPU-batches that contained no positive-event samples had their last sample replaced with a positive-event sample chosen randomly from the training set. For evaluation, we use the model with maximum validation performance.

Survival time segments Survival functions ultimately predict patient survival at fixed time points. We divided the time-to-event range for each validation dataset into 100 uniformly distributed time points. When measuring model performance, we use the time point closest to the time horizon of interest. (e.g. MIMIC-IV data has a maximum follow-up time of almost 13 years, and to evaluate model performance at the one-year mark we use the survival function output at the 1.008-year mark).

Cox models While PyCox models readily provide survival functions from input ECGs, classifiers need an additional step to transform their output to a survival function. We fit CoxPH models from classifier outputs on the validation set with the scikit-survival package [33–35].

Measures We report the concordance index (concordance or C-Index), integrated Brier score, AUROC, and AUPRC. Concordance evaluates subject (patient) risk ordering: at the time of an event (death), a subject should be at a higher risk than any other subject still under observation; the C-index is the fraction of correctly ranked subject-subject comparisons. For classifiers and DeepSurv, which

are proportional hazard models, a subject's risk rank does not change over time. For DeepHit, LH, and MTLR, it can, and studies typically use Antolini concordance [36]. As Antolini concordance is equivalent to the standard c-index in proportional cases, we use it for all concordance calculations. Concordance is usually measured across all time points, although some authors suggest that an end-point is essential to properly interpret model efficacy [37]. Along with measurements over all time points, we measure the concordance and integrated Brier score with event times censored to the one, two, five, and ten-year marks in the Supplement. The censorship determines which pairs are considered comparable at the evaluation time-point. The Brier score is only integrated up to the censored time, after which it becomes unstable.

The concordance index and integrated Brier score (a weighted MSE) are computed with PyCox. Since these are summary scores over a time when patients may leave the study, it is common to weight intermediate values per time point based on censoring, which is estimated through a Kaplan-Meier fit [23]. We also look at AUPRC and AUROC, with the 'correct' labels set just as in our classifier labeling (1 if a patient experiences an event by time T, else 0, even if censored). For these measures we compare the cumulative hazard $H(t) = (1 - S(t))$, to the label.

External evaluation We also measured the ability of models trained on one dataset (MIMIC-IV or Code-15) to generalize to the other. For each dataset and approach (CoxPH or Classifier), the model with the highest concordance was evaluated on the test portion of the other dataset. This process gives a realistic estimate of how a final model might behave when deployed to new environment.

4 Results

Concordance indices and integrated Brier scores over all time are reported in Table 3, while right-censored values are reported in supplementary Tables 6 and 7. The external evaluation of models across the care centers is presented in Table 4. Fig. 2 shows the Kaplan-Meier and Model survival functions for the top-concordance Code-15 and MIMIC-IV models. Model survival functions follow the Kaplan-Meier curves well before diverging at the end of the study time. AUROC and AUPRC are reported for specific follow-up times in supplementary Tables 5 and 8.

Table 3: Integrated Brier score and C-Index (median, inter-quartile range) across the full study time and all test-set ECGs, grouped by task, architecture, and dataset. Bold marks the best median values.

	Dataset	Code-15		MIMIC-IV	
	Metric	Int. Brier (↓)	Concordance (↑)	Int. Brier (↓)	Concordance (↑)
Task	Architecture				
Cla-1	Inception	0.04 (0.04-0.04)	0.75 (0.75-0.76)	0.18 (0.18-0.18)	0.77 (0.77-0.77)
	Resnet	0.04 (0.04-0.04)	0.72 (0.67-0.77)	0.18 (0.18-0.18)	0.77 (0.77-0.77)
Cla-10	Inception	0.04 (0.04-0.04)	0.78 (0.78-0.78)	0.18 (0.18-0.18)	0.76 (0.75-0.76)
	Resnet	0.04 (0.04-0.04)	0.79 (0.78-0.79)	0.18 (0.18-0.18)	0.76 (0.76-0.76)
Cla-2	Inception	0.04 (0.04-0.04)	0.77 (0.77-0.77)	0.18 (0.18-0.18)	0.77 (0.77-0.77)
	Resnet	0.04 (0.04-0.04)	0.77 (0.73-0.78)	0.18 (0.18-0.18)	0.77 (0.77-0.77)
Cla-5	Inception	0.04 (0.04-0.04)	0.78 (0.78-0.79)	0.18 (0.18-0.18)	0.76 (0.76-0.76)
	Resnet	0.04 (0.04-0.04)	0.78 (0.78-0.79)	0.18 (0.18-0.18)	0.76 (0.76-0.76)
DeepHit	Inception	**0.03 (0.03-0.03)**	0.79 (0.78-0.79)	0.18 (0.18-0.19)	**0.78 (0.78-0.78)**
	Resnet	0.04 (0.04-0.04)	0.59 (0.55-0.60)	0.18 (0.18-0.18)	**0.78 (0.77-0.78)**
DeepSurv	Inception	0.04 (0.04-0.04)	0.79 (0.78-0.79)	**0.17 (0.17-0.17)**	0.77 (0.77-0.77)
	Resnet	0.04 (0.04-0.04)	**0.80 (0.79-0.80)**	0.18 (0.18-0.18)	0.77 (0.77-0.77)
LH	Inception	0.04 (0.04-0.04)	0.78 (0.78-0.78)	**0.17 (0.17-0.17)**	0.77 (0.77-0.77)
	Resnet	0.04 (0.04-0.04)	0.78 (0.77-0.79)	0.18 (0.18-0.18)	0.77 (0.77-0.78)
MTLR	Inception	**0.03 (0.03-0.04)**	0.77 (0.77-0.78)	**0.17 (0.17-0.17)**	0.77 (0.77-0.77)
	Resnet	0.04 (0.04-0.04)	0.00 (0.00-0.00)	**0.17 (0.17-0.17)**	0.77 (0.77-0.78)

Code-15 models report comparable AUROCs and c-index values of around 0.78 and AUPRCs around 0.06 / 0.08 / 0.12 at years 1 / 2 / 5, which is substantially higher than the 1.23% / 2.07% / 3.36% chance rates or overall event rate of 3.6%. Integrated Brier scores are all around 0.03-0.04. Overall, Resnet DeepSurv models are have slightly better metrics than the rest, with AUROCs at 0.81-0.82

Table 4: External evaluation of models on new cohorts. 1-year AUROC, 1-year AUPRC, integrated Brier score and concordance indices, grouped by train/test dataset, and task specification. Best architecture on internal validation is shown. Bold denotes best median values on external validation.

Trained on: Code-15

	1-yr AUROC (↑)		1-yr AUPRC (↑)		Concordance (↑)		Brier (↓)	
Test Data	Code15	MIMIC	Code15	MIMIC	Code15	MIMIC	Code15	MIMIC
Task								
Cla-1	0.80	0.72	0.06	0.28	0.77	0.69	0.04	0.24
Cla-10	0.82	**0.73**	0.07	0.29	0.80	**0.71**	0.04	0.25
Cla-2	0.80	0.69	0.06	0.26	0.79	0.67	0.04	0.25
Cla-5	0.81	0.72	0.06	0.28	0.79	0.70	0.04	0.26
DeepHit	0.80	0.66	0.06	0.25	0.79	0.64	0.03	0.23
DeepSurv	0.82	0.72	0.07	0.28	0.80	0.70	0.04	**0.22**
LH	0.81	**0.73**	0.07	**0.30**	0.79	**0.71**	0.04	0.24
MTLR	0.80	0.68	0.06	0.25	0.78	0.66	0.04	0.23

Trained on: MIMIC-IV

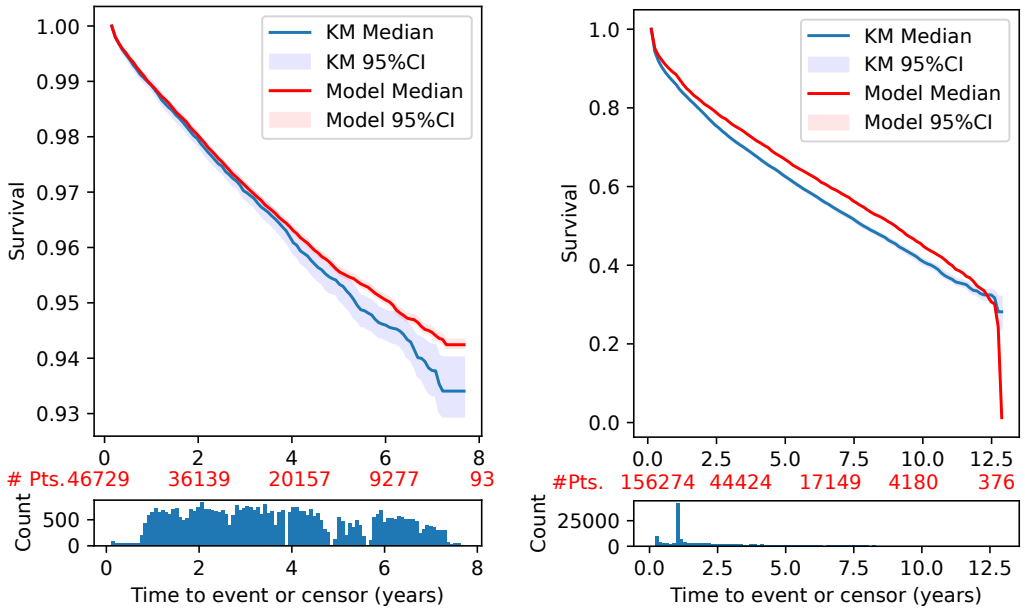
	1-yr AUROC (↑)		1-yr AUPRC (↑)		Concordance (↑)		Brier (↓)	
Test Data	Code15	MIMIC	Code15	MIMIC	Code15	MIMIC	Code15	MIMIC
Task								
Cla-1	0.65	0.81	**0.04**	0.43	0.65	0.78	0.39	0.18
Cla-10	0.69	0.79	0.03	0.40	0.66	0.76	0.38	0.18
Cla-2	0.66	0.81	0.03	0.43	0.63	0.77	0.28	0.18
Cla-5	0.69	0.80	**0.04**	0.40	0.66	0.77	0.34	0.18
DeepHit	**0.71**	0.81	0.03	0.43	0.38	0.78	0.94	0.18
DeepSurv	0.69	0.80	0.03	0.42	0.68	0.77	0.31	0.18
LH	0.64	0.81	**0.04**	0.43	0.63	0.78	**0.05**	0.18
MTLR	**0.71**	0.81	0.03	0.43	**0.67**	0.78	0.61	0.17

and AUPRCs of 0.07/0.10/0.13. DeepHit and MTLR exhibit sensitivity to the choice of architecture: median AUROC performance was 19-22 points higher when using the InceptionTime architecture relative to Resnet. Surprisingly, InceptionTime does as well as Resnet aside from the DeepSurv case, despite Resnet being originally designed for a larger version of the Code dataset in [13].

MIMIC-IV models performed similarly with AUROCs at roughly 0.8. AUPRC values are roughly 0.42 / 0.46 / 0.52 / 0.53 for 1 / 2 / 5 / 10-yr measures, substantially higher than the chance rates of 14.7% / 18.3% / 24.5% / 27.6% or the overall event rate of 27.8%. Classifiers evaluated before their time horizon do slightly worse than other models, and DeepHit models, despite maintaining 0.78 concordance, drop behind other models in AUPRC at later time points. Architecture does not appear to affect outcomes. Concordance and integrated Brier scores evaluated across all-time Table 3 show better scores when evaluated with one ECG per-patient (C-Index around 0.82) rather than across all ECGs simultaneously (C-index around 0.77).

We were surprised to see limited differences between Cox proportional hazards models built on classifier outputs and deep survival models. The largest differences appeared in Code-15 when comparing DeepSurv with other task definitions; however these differences in performance were in the 1-3% range. Classifiers predicting 1-year mortality were competitive at the 2, 5, and 10-year marks, which can be partially explained by drop in patient counts at longer time horizons (Fig. 2).

Model training time On average, Code-15 models took 4.5 hours to train and achieved maximum validation performance by 30 epochs. MIMIC-IV models took 7.2 hours and achieved maximum validation performance by 12 epochs. Training times depend on GPU, with faster GPUs taking 600s/epoch and slower ones up to 2200s/epoch. Batch sizes were set based on slower GPUs.



(a) Code-15 (b) MIMIC-IV

Figure 2: Kaplan-Meier and survival functions for the highest-concordance-index Code-15 (left) and MIMIC-IV (right) models. Bottom subplots show sample count at each time point. Plots show median and 95% confidence intervals per time point (100) over 100 bootstraps. Red values indicate the number of patients under observation. Note differences in scale.

5 Discussion and Conclusions

Comparison to other studies Most papers report AUROC and C-Index scores around 0.81-0.85 for the one year mark [4, 5], and values around 0.78-0.83 for the five-year mark[3, 5]. Notably, these studies typically also model age/sex factors and adjust architectures to improve performance on their datasets, whereas we did not include demographics and use the same architectures for both datasets.

Model external evaluation Model AUROC and concordance drop by 0.1-0.15 on switching evaluation datasets (Table 4), while AUPRC approaches dataset chance rates (MIMIC-IV: 27%, Code-15: 4%). We expect this is mostly due to different ECG sources. Code-15 was mostly collected in routine care whereas MIMIC-IV is often from an in-patient hospital setting, which reflects in cohort mortality rates. Additionally, a cursory look at ECGs suggests that Code-15 is more non-stationary than MIMIC-IV, perhaps indicating different ECG filtering for e.g. motion artifacts. Our data normalization procedures have a limited ability to account for such effects. A prior study evaluating ECG survival models across US hospital systems did not report performance drops [3].

On labeling classifier data When training classifiers, we label events as a 1 if a patient experiences an event by time t, else 0, even if censored. This same labeling applied to AUROC and AUPRC estimates, but not to concordance or integrated Brier scores. While there is no standard approach for integrating censored patients with time-dependent measures [38], past work tends to exclude patients that have not reached the horizon-of-interest [4, 5], which inflates the event rate (imagine a 1000-person study with an annual 90 percent drop-out rate and 1 percent event rate: after two years there are 11 events and 8 remaining controls). We chose to keep censored patients as controls for classifiers and AUROC/PRC to avoid inflating the event rate. This is not unreasonable for MIMIC-IV: many controls are censored at the one-year mark, leaving classifiers only with patients that have repeat ECGS or experience events; the censoring here is clearly informative. In Code-15's case, patients leave the study roughly uniformly after the first year. With only a 3.6 percent overall event rate, the drop-out rate substantially exceeds the event rate.

Right-censoring concordance and integrated Brier score The supplement includes right-censored concordance (Table 6) and right-censored integrated Brier scores (Table 7), with data censored to 1/2/5/10-years. Concordance drops by 0.01-0.02 for both datasets over time, while Brier changes more substantially, similar to AURPC and 'chance' rates. This implies that metrics and models can be quite sensitive to time horizons (see model survival estimates in Fig. 2. B).

Deep survival models vs classifiers Deep survival models tend to have a simpler evaluation as they are not trained with a particular time point in mind and, at least theoretically, should be more applicable to survival modeling. In contrast, classifier approaches are more flexible: since the survival model is ultimately built through a Cox regression, additional demographic factors can be added (or accounted for) in outcome modeling. This has previously shown improved performance[4].

Limitations This study has several limitations. First, not all survival modeling techniques are represented: random forests or approaches such as XGBoost on ECG machine outputs could provide additional valuable baselines for future work. Second, while many papers tune model architectures to perform better on their datasets, we did not do so and cannot provide guidance for architecture modifications. Third, we do not include saliency maps (e.g. GradCam) to indicate which parts of the ECG signal most strongly affect model values. Last, past work typically includes demographic variables in their modeling or evaluates model performance per pathology. Doing so requires expert input and dataset standardization across multiple public databases, which is beyond the scope of this work. The ethical deployment of survival models in a clinical setting would require that model performance be evaluated in those contexts with appropriate subgroup analyses and a focus on how the model's use in downstream-decision making would impact patient care.

Conclusions We benchmark mortality prediction from ECGs on the two largest public ECG datasets using four deep learning survival models and four classifiers followed by Cox regressions to predict mortality at one, two, five, and ten-year horizons. Our results highlight some models as lacking (Resnet + DeepHit or MTLR with the Code-15 dataset), with most models otherwise performing at comparable levels. Cross-dataset evaluation shows performance drops, not unexpectedly given cohort differences, and implies that model generalization to different settings is not guaranteed. Our results are reasonable given other papers' results and the event rates per dataset. We hope this provides a valuable resource for future ECG and ECG-mortality studies.

Acknowledgments and Disclosure of Funding

Computing Sources The authors would like to acknowledge Boston Children's Hospital's High-Performance Computing Resources Clusters Enkefalos 2/3 (E2/3) made available for conducting the research reported in this publication.

Funding Sources Funding support received from the Thrasher Research Fund Early Career Award (J.M.), Boston Children's Hospital Electrophysiology Research Education Fund (J.M., J.K.T.), NIH grant R00-LM012926 from the National Library of Medicine (W.G.L.), NIH grant T32HD040128 from the National Institute of Childhood Diseases and Human Development (P.L.), and the Kostin Innovation Fund (W.G.L.).

References

[1] Salah S. Al-Zaiti, Christian Martin-Gill, Jessica K. Zègre-Hemsey, Zeineb Bouzid, Ziad Faramand, Mohammad O. Alrawashdeh, Richard E. Gregg, Stephanie Helman, Nathan T. Riek, Karina Kraevsky-Phillips, Gilles Clermont, Murat Akcakaya, Susan M. Sereika, Peter Van Dam, Stephen W. Smith, Yochai Birnbaum, Samir Saba, Ervin Sejdic, and Clifton W. Callaway. Machine learning for ECG diagnosis and risk stratification of occlusion myocardial infarction. *Nature Medicine*, 29(7):1804–1813, July 2023. 1, 2, 3

[2] Shaan Khurshid, Samuel Friedman, Christopher Reeder, Paolo Di Achille, Nathaniel Diamant, Pulkit Singh, Lia X. Harrington, Xin Wang, Mostafa A. Al-Alusi, Gopal Sarma, Andrea S. Foulkes, Patrick T. Ellinor, Christopher D. Anderson, Jennifer E. Ho, Anthony A. Philippakis, Puneet Batra, and Steven A. Lubitz. ECG-Based Deep Learning and Clinical Risk Factors to Predict Atrial Fibrillation. *Circulation*, 145(2):122–133, January 2022. 3

[3] J. Weston Hughes, James Tooley, Jessica Torres Soto, Anna Ostropolets, Tim Poterucha, Matthew Kai Christensen, Neal Yuan, Ben Ehlert, Dhamanpreet Kaur, Guson Kang, Albert Rogers, Sanjiv Narayan, Pierre Elias, David Ouyang, Euan Ashley, James Zou, and Marco V. Perez. A deep learning-based electrocardiogram risk score for long term cardiovascular death and disease. *npj Digital Medicine*, 6(1):169, September 2023. 2, 5, 8

[4] Sushravya Raghunath, Alvaro E. Ulloa Cerna, Linyuan Jing, David P. vanMaanen, Joshua Stough, Dustin N. Hartzel, Joseph B. Leader, H. Lester Kirchner, Martin C. Stumpe, Ashraf Hafez, Arun Nemani, Tanner Carbonati, Kipp W. Johnson, Katelyn Young, Christopher W. Good, John M. Pfeifer, Aalpen A. Patel, Brian P. Delisle, Amro Alsaid, Dominik Beer, Christopher M. Haggerty, and Brandon K. Fornwalt. Prediction of mortality from 12-lead electrocardiogram voltage data using a deep neural network. *Nature Medicine*, 26(6):886–891, June 2020. 2, 3, 8, 9

[5] Weijie Sun, Sunil Vasu Kalmady, Nariman Sepehrvand, Amir Salimi, Yousef Nademi, Kevin Bainey, Justin A. Ezekowitz, Russell Greiner, Abram Hindle, Finlay A. McAlister, Roopinder K. Sandhu, and Padma Kaul. Towards artificial intelligence-based learning health system for population-level mortality prediction using electrocardiograms. *npj Digital Medicine*, 6(1):21, February 2023. 1, 2, 3, 8

[6] Takahiro Kokubo, Satoshi Kodera, Shinnosuke Sawano, Susumu Katsushika, Mitsuhiko Nakamoto, Hirotoshi Takeuchi, Nisei Kimura, Hiroki Shinohara, Ryo Matsuoka, Koki Nakanishi, Tomoko Nakao, Yasutomi Higashikuni, Norifumi Takeda, Katsuhito Fujiu, Masao Daimon, Hiroshi Akazawa, Hiroyuki Morita, Yutaka Matsuyama, and Issei Komuro. Automatic detection of left ventricular dilatation and hypertrophy from electrocardiograms using deep learning. *International Heart Journal*, 63(5):939–947, 2022. 1

[7] Demilade Adedinsewo, Rickey E. Carter, Zachi Attia, Patrick Johnson, Anthony H. Kashou, Jennifer L. Dugan, Michael Albus, Johnathan M. Sheele, Fernanda Bellolio, Paul A. Friedman, Francisco Lopez-Jimenez, and Peter A. Noseworthy. Artificial Intelligence-Enabled ECG Algorithm to Identify Patients With Left Ventricular Systolic Dysfunction Presenting to the Emergency Department With Dyspnea. *Circulation: Arrhythmia and Electrophysiology*, 13(8):e008437, August 2020. Publisher: American Heart Association. 1

[8] Joshua Mayourian, William G. La Cava, Akhil Vaid, Girish N. Nadkarni, Sunil J. Ghelani, Rebekah Mannix, Tal Geva, Audrey Dionne, Mark E. Alexander, Son Q. Duong, and John K. Triedman. Pediatric ECG-Based Deep Learning to Predict Left Ventricular Dysfunction and Remodeling. *Circulation*, 149(12):917–931, March 2024. Publisher: American Heart Association. 1

[9] Brian Gow, Tom Pollard, Larry A Nathanson, Alistair Johnson, Benjamin Moody, Chrystinne Fernandes, Nathaniel Greenbaum, Jonathan W Waks, Parastou Eslami, Tanner Carbonati, Ashish Chaudhari, Elizabeth Herbst, Dana Moukheiber, Seth Berkowitz, Roger Mark, and Steven Horng. MIMIC-IV-ECG: Diagnostic Electrocardiogram Matched Subset, 2023. 1, 2, 3

[10] A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. Ch. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220, 2000 (June 13). Circulation Electronic Pages: http://circ.ahajournals.org/content/101/23/e215.full PMID:1085218; doi: 10.1161/01.CIR.101.23.e215.

[11] Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. Mimic-iv (version 2.2), 2023.

[12] Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J. Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, Li-wei H. Lehman, Leo A. Celi, and Roger G. Mark. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10(1):1, January 2023. Publisher: Nature Publishing Group. 1

[13] Antônio H. Ribeiro, Manoel Horta Ribeiro, Gabriela M. M. Paixão, Derick M. Oliveira, Paulo R. Gomes, Jéssica A. Canazart, Milton P. S. Ferreira, Carl R. Andersson, Peter W. Macfarlane,

Wagner Meira, Thomas B. Schön, and Antonio Luiz P. Ribeiro. Automatic diagnosis of the 12-lead ECG using a deep neural network. *Nature Communications*, 11(1):1760, April 2020. 1, 2, 3, 4, 5, 7

[14] Harold Smulyan. The Computerized ECG: Friend and Foe. *The American Journal of Medicine*, 132(2):153–160, February 2019. 2

[15] Dung-Jang Tsai, Yu-Sheng Lou, Chin-Sheng Lin, Wen-Hui Fang, Chia-Cheng Lee, Ching-Liang Ho, Chih-Hung Wang, and Chin Lin. Mortality risk prediction of the electrocardiogram as an informative indicator of cardiovascular diseases. *DIGITAL HEALTH*, 9:20552076231187247, January 2023. 2, 3

[16] R. R. Van De Leur, H. Bleijendaal, K. Taha, T. Mast, J. M. I. H. Gho, M. Linschoten, B. Van Rees, M. T. H. M. Henkens, S. Heymans, N. Sturkenboom, R. A. Tio, J. A. Offerhaus, W. L. Bor, M. Maarse, H. E. Haerkens-Arends, M. Z. H. Kolk, A. C. J. Van Der Lingen, J. J. Selder, E. E. Wierda, P. F. M. M. Van Bergen, M. M. Winter, A. H. Zwinderman, P. A. Doevendans, P. Van Der Harst, Y. M. Pinto, F. W. Asselbergs, R. Van Es, F. V. Y. Tjong, and the CAPACITY-COVID collaborative consortium. Electrocardiogram-based mortality prediction in patients with COVID-19 using machine learning. *Netherlands Heart Journal*, 30(6):312–318, June 2022. 2, 3

[17] Kuba Weimann and Tim O. F. Conrad. Transfer learning for ECG classification. *Scientific Reports*, 11(1):5251, March 2021. Publisher: Nature Publishing Group. 2

[18] Elena Merdjanovska and Aleksandra Rashkovska. Comprehensive survey of computational ECG analysis: Databases, methods and applications. *Expert Systems with Applications*, 203:117206, October 2022. 2

[19] Jason T Rich, J Gail Neely, Randal C Paniello, Courtney CJ Voelker, Brian Nussenbaum, and Eric W Wang. A practical guide to understanding kaplan-meier curves. *Otolaryngology—Head and Neck Surgery*, 143(3):331–336, 2010. 2

[20] David R Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972. 2

[21] D. Y. Lin. On the Breslow estimator. *Lifetime Data Analysis*, 13(4):471–480, December 2007. 2

[22] Terry M. Therneau. Modeling survival data: Extending the Cox Model. In P. Bickel, P. Diggle, S. Fienberg, K. Krickeberg, I. Olkin, N. Wermuth, S. Zeger, D. Y. Lin, and T. R. Fleming, editors, *Proceedings of the First Seattle Symposium in Biostatistics*, volume 123, pages 51–84. Springer US, New York, NY, 1997. 2

[23] Håvard Kvamme, Ørnulf Borgan, and Ida Scheel. Time-to-Event Prediction with Neural Networks and Cox Regression. *Journal of Machine Learning Research*, 20(129):1–30, 2019. 3, 6

[24] Jared L. Katzman, Uri Shaham, Alexander Cloninger, Jonathan Bates, Tingting Jiang, and Yuval Kluger. DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC Medical Research Methodology*, 18(1):24, December 2018. 3, 5

[25] Michael F. Gensheimer and Balasubramanian Narasimhan. A scalable discrete-time survival model for neural networks. *PeerJ*, 7:e6257, January 2019. 3, 5

[26] Stephane Fotso. Deep Neural Networks for Survival Analysis Based on a Multi-Task Framework, January 2018. arXiv:1801.05512 [cs, stat]. 3, 5

[27] Changhee Lee, William Zame, Jinsung Yoon, and Mihaela Van Der Schaar. DeepHit: A Deep Learning Approach to Survival Analysis With Competing Risks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), April 2018. 3, 5

[28] Emilly M. Lima, Antônio H. Ribeiro, Gabriela M. M. Paixão, Manoel Horta Ribeiro, Marcelo M. Pinto-Filho, Paulo R. Gomes, Derick M. Oliveira, Ester C. Sabino, Bruce B. Duncan, Luana Giatti, Sandhi M. Barreto, Wagner Meira Jr, Thomas B. Schön, and Antonio Luiz P. Ribeiro. Deep neural network-estimated electrocardiographic age as a mortality predictor. *Nature Communications*, 12(1):5117, August 2021. 3, 4

[29] Hassan Ismail Fawaz, Benjamin Lucas, Germain Forestier, Charlotte Pelletier, Daniel F. Schmidt, Jonathan Weber, Geoffrey I. Webb, Lhassane Idoumghar, Pierre-Alain Muller, and François Petitjean. InceptionTime: Finding AlexNet for Time Series Classification. *Data Mining and Knowledge Discovery*, 34(6):1936–1962, November 2020. arXiv:1909.04939 [cs, stat]. 3, 4, 5

[30] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, June 2016. 4

[31] Jennifer A. Mccoy, Lisa D. Levine, Guangya Wan, Corey Chivers, Joseph Teel, and William G. La cava. Intrapartum electronic fetal heart rate monitoring to predict acidemia at birth with the use of deep learning. *American Journal of Obstetrics and Gynecology*, April 2024. 4

[32] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. 4

[33] Sebastian Pölsterl, Nassir Navab, and Amin Katouzian. Fast training of support vector machines for survival analysis. In Annalisa Appice, Pedro Pereira Rodrigues, Vítor Santos Costa, João Gama, Alípio Jorge, and Carlos Soares, editors, *Machine Learning and Knowledge Discovery in Databases*, Lecture Notes in Computer Science, pages 243–259, 2015. 5

[34] Sebastian Pölsterl, Gupta Pankaj, Lichao Wang, Sailesh Conjeti, Amin Katouzian, and Nassir Navab. Heterogeneous Ensembles for Predicting Survival of Metastatic, Castrate-Resistant Prostate Cancer Patients. *F1000Research*, 5(2676), November 2016.

[35] Sebastian Pölsterl, Nassir Navab, and Amin Katouzian. An Efficient Training Algorithm for Kernel Survival Support Vector Machines. In *3rd Workshop on Machine Learning in Life Sciences*, September 2016. 5

[36] Laura Antolini, Patrizia Boracchi, and Elia Biganzoli. A time-dependent discrimination index for survival data. *Statistics in Medicine*, 24(24):3927–3944, December 2005. 6

[37] Terry M Therneau. *A Package for Survival Analysis in R*, 2024. R package version 3.6-4. 6

[38] Adina Najwa Kamarudin, Trevor Cox, and Ruwanthi Kolamunnage-Dona. Time-dependent ROC curve analysis in medical research: current methods and applications. *BMC Medical Research Methodology*, 17:53, April 2017. 8

A Appendix

Dataset Licenses The MIMIC-IV ECG dataset is available from https://physionet.org/content/mimic-iv-ecg/1.0/ under the Open Data Commons Open Database License v1.0. The Code-15 dataset is available from https://zenodo.org/records/4916206 under Creative Commons license CC-BY 4.0. The datasets use de-identified patient information. Code to reproduce the experiments is available from https://github.com/cavalab/ecg-survival-benchmark under a GNU Public License version 3.0.

B Supplemental Results

Table 5: AUROC (median, inter-quartile range) at different prediction horizons, grouped by dataset, task specification, and architecture.

Code-15					
Task	Architecture	1-yr AUROC	2-yr AUROC	5-yr AUROC	
Cla-1	Inception	0.77 (0.76-0.77)	0.76 (0.75-0.76)	0.76 (0.75-0.76)	
	Resnet	0.73 (0.69-0.79)	0.72 (0.68-0.78)	0.72 (0.68-0.77)	
Cla-2	Inception	0.79 (0.78-0.79)	0.78 (0.78-0.78)	0.78 (0.78-0.78)	
	Resnet	0.78 (0.74-0.80)	0.77 (0.73-0.79)	0.77 (0.74-0.78)	
Cla-5	Inception	0.79 (0.79-0.80)	0.79 (0.79-0.79)	0.79 (0.79-0.79)	
	Resnet	0.80 (0.79-0.81)	0.79 (0.79-0.80)	0.79 (0.79-0.79)	
Cla-10	Inception	0.79 (0.79-0.80)	0.79 (0.79-0.79)	0.79 (0.79-0.79)	
	Resnet	0.81 (0.80-0.81)	0.80 (0.79-0.80)	0.79 (0.78-0.80)	
DeepHit	Inception	0.80 (0.80-0.80)	0.79 (0.79-0.79)	0.79 (0.79-0.79)	
	Resnet	0.58 (0.54-0.59)	0.58 (0.54-0.59)	0.60 (0.55-0.61)	
DeepSurv	Inception	0.80 (0.80-0.80)	0.79 (0.79-0.79)	0.79 (0.79-0.79)	
	Resnet	**0.82 (0.80-0.82)**	**0.81 (0.80-0.81)**	**0.81 (0.79-0.81)**	
LH	Inception	0.79 (0.78-0.79)	0.78 (0.78-0.78)	0.79 (0.78-0.79)	
	Resnet	0.79 (0.78-0.81)	0.78 (0.77-0.80)	0.78 (0.77-0.79)	
MTLR	Inception	0.78 (0.78-0.79)	0.78 (0.77-0.79)	0.78 (0.78-0.78)	
	Resnet	0.50 (0.50-0.50)	0.50 (0.50-0.50)	0.50 (0.50-0.50)	

MIMIC-IV					
Task	Architecture	1-yr AUROC	2-yr AUROC	5-yr AUROC	10-yr AUROC
Cla-1	Inception	**0.81 (0.80-0.81)**	0.79 (0.79-0.79)	0.77 (0.77-0.77)	0.76 (0.75-0.76)
	Resnet	**0.81 (0.81-0.81)**	**0.80 (0.80-0.80)**	0.78 (0.78-0.78)	0.76 (0.76-0.76)
Cla-2	Inception	0.80 (0.80-0.80)	0.79 (0.79-0.79)	0.78 (0.77-0.78)	0.76 (0.76-0.77)
	Resnet	0.80 (0.80-0.81)	0.79 (0.79-0.80)	0.78 (0.78-0.78)	0.77 (0.76-0.77)
Cla-5	Inception	0.79 (0.79-0.80)	0.79 (0.79-0.79)	0.78 (0.78-0.78)	0.77 (0.76-0.77)
	Resnet	0.79 (0.79-0.80)	0.79 (0.78-0.79)	0.78 (0.77-0.78)	0.77 (0.76-0.77)
Cla-10	Inception	0.78 (0.78-0.79)	0.78 (0.78-0.78)	0.77 (0.77-0.77)	0.77 (0.77-0.77)
	Resnet	0.79 (0.79-0.79)	0.78 (0.78-0.79)	0.78 (0.78-0.78)	0.77 (0.77-0.77)
DeepHit	Inception	0.80 (0.80-0.80)	0.79 (0.79-0.79)	0.75 (0.75-0.76)	0.73 (0.72-0.73)
	Resnet	**0.81 (0.80-0.81)**	0.79 (0.79-0.80)	0.77 (0.77-0.78)	0.74 (0.74-0.76)
DeepSurv	Inception	0.80 (0.80-0.80)	0.79 (0.79-0.79)	0.78 (0.77-0.78)	0.76 (0.76-0.76)
	Resnet	0.80 (0.80-0.80)	0.79 (0.79-0.79)	0.78 (0.78-0.78)	0.77 (0.77-0.77)
LH	Inception	0.80 (0.80-0.81)	0.79 (0.79-0.79)	0.77 (0.77-0.77)	0.75 (0.75-0.76)
	Resnet	0.80 (0.80-0.81)	0.79 (0.79-0.80)	0.78 (0.78-0.78)	0.76 (0.76-0.76)
MTLR	Inception	**0.81 (0.80-0.81)**	0.79 (0.79-0.79)	0.77 (0.77-0.78)	0.76 (0.76-0.76)
	Resnet	**0.81 (0.81-0.81)**	0.79 (0.79-0.80)	0.78 (0.78-0.78)	0.76 (0.76-0.76)

Table 6: Concordance with data right-censored to evaluation time. (median, interquartile range over 20 bootstraps, picking a random ECG per patient)

Dataset	Task	Architecture	1-yr Concordance	2-yr Concordance	5-yr Concordance	10-yr Concordance
Code15	DeepSurv	Inception	0.80 (0.79-0.80)	0.79 (0.79-0.79)	0.79 (0.78-0.79)	-
		Resnet	0.82 (0.80-0.82)	0.81 (0.79-0.81)	0.80 (0.79-0.80)	-
	DeepHit	Inception	0.80 (0.80-0.80)	0.79 (0.79-0.79)	0.79 (0.78-0.79)	-
		Resnet	0.58 (0.54-0.58)	0.58 (0.54-0.58)	0.59 (0.55-0.60)	-
	LH	Inception	0.78 (0.78-0.78)	0.78 (0.78-0.78)	0.78 (0.78-0.78)	-
		Resnet	0.79 (0.78-0.81)	0.78 (0.77-0.80)	0.78 (0.77-0.79)	-
	MTLR	Inception	0.78 (0.78-0.79)	0.77 (0.77-0.78)	0.77 (0.77-0.78)	-
		Resnet	0.00 (0.00-0.00)	0.00 (0.00-0.00)	0.00 (0.00-0.00)	-
	Cla-1	Inception	0.77 (0.76-0.77)	0.75 (0.75-0.76)	0.75 (0.75-0.76)	-
		Resnet	0.73 (0.68-0.79)	0.72 (0.68-0.78)	0.72 (0.67-0.77)	-
	Cla-2	Inception	0.79 (0.78-0.79)	0.78 (0.78-0.78)	0.77 (0.77-0.77)	-
		Resnet	0.78 (0.74-0.79)	0.77 (0.72-0.78)	0.77 (0.73-0.78)	-
	Cla-5	Inception	0.79 (0.79-0.80)	0.78 (0.78-0.79)	0.78 (0.78-0.79)	-
		Resnet	0.80 (0.79-0.81)	0.79 (0.78-0.80)	0.78 (0.78-0.79)	-
	Cla-10	Inception	0.79 (0.79-0.80)	0.79 (0.79-0.79)	0.78 (0.78-0.78)	-
		Resnet	0.80 (0.80-0.81)	0.79 (0.79-0.80)	0.79 (0.78-0.79)	-
MIMICIV	DeepSurv	Inception	0.83 (0.83-0.83)	0.83 (0.83-0.83)	0.83 (0.82-0.83)	0.83 (0.82-0.83)
		Resnet	0.83 (0.83-0.84)	0.83 (0.83-0.83)	0.83 (0.82-0.83)	0.83 (0.82-0.83)
	DeepHit	Inception	0.84 (0.84-0.84)	0.83 (0.83-0.84)	0.83 (0.83-0.83)	0.83 (0.83-0.83)
		Resnet	0.84 (0.83-0.84)	0.83 (0.83-0.84)	0.83 (0.83-0.83)	0.83 (0.83-0.83)
	LH	Inception	0.83 (0.83-0.83)	0.83 (0.83-0.83)	0.83 (0.82-0.83)	0.83 (0.82-0.83)
		Resnet	0.83 (0.83-0.84)	0.83 (0.83-0.83)	0.83 (0.82-0.83)	0.83 (0.82-0.83)
	MTLR	Inception	0.83 (0.83-0.84)	0.83 (0.83-0.83)	0.83 (0.83-0.83)	0.83 (0.83-0.83)
		Resnet	0.83 (0.83-0.84)	0.83 (0.83-0.83)	0.83 (0.83-0.83)	0.83 (0.83-0.83)
	Cla-1	Inception	0.83 (0.83-0.83)	0.83 (0.83-0.83)	0.82 (0.82-0.83)	0.82 (0.82-0.82)
		Resnet	0.84 (0.84-0.84)	0.83 (0.83-0.83)	0.83 (0.83-0.83)	0.83 (0.83-0.83)
	Cla-2	Inception	0.83 (0.83-0.83)	0.83 (0.83-0.83)	0.83 (0.82-0.83)	0.82 (0.82-0.83)
		Resnet	0.83 (0.83-0.83)	0.83 (0.83-0.83)	0.83 (0.82-0.83)	0.83 (0.82-0.83)
	Cla-5	Inception	0.83 (0.83-0.83)	0.83 (0.82-0.83)	0.82 (0.82-0.82)	0.82 (0.82-0.82)
		Resnet	0.83 (0.83-0.83)	0.83 (0.82-0.83)	0.82 (0.82-0.83)	0.82 (0.82-0.83)
	Cla-10	Inception	0.82 (0.82-0.83)	0.82 (0.82-0.82)	0.82 (0.82-0.82)	0.82 (0.82-0.82)
		Resnet	0.83 (0.83-0.83)	0.82 (0.82-0.83)	0.82 (0.82-0.82)	0.82 (0.82-0.82)

Table 7: Integrated Brier scores with data right-censored to evaluation time. (median, interquartile range over 20 bootstraps, picking a random ECG per patient)

Dataset	Task	Architecture	1-yr Brier	2-yr Brier	5-yr Brier	10-yr Brier
Code15	DeepSurv	Inception	0.01 (0.01-0.01)	0.01 (0.01-0.01)	0.02 (0.02-0.02)	-
		Resnet	0.01 (0.01-0.01)	0.01 (0.01-0.01)	0.02 (0.02-0.02)	-
	DeepHit	Inception	0.01 (0.01-0.01)	0.01 (0.01-0.01)	0.02 (0.02-0.02)	-
		Resnet	0.01 (0.01-0.01)	0.01 (0.01-0.01)	0.02 (0.02-0.02)	-
	LH	Inception	0.01 (0.01-0.01)	0.01 (0.01-0.01)	0.02 (0.02-0.02)	-
		Resnet	0.01 (0.01-0.01)	0.01 (0.01-0.01)	0.02 (0.02-0.02)	-
	MTLR	Inception	0.01 (0.01-0.01)	0.01 (0.01-0.01)	0.02 (0.02-0.02)	-
		Resnet	0.01 (0.01-0.01)	0.01 (0.01-0.01)	0.02 (0.02-0.02)	-
	Cla-1	Inception	0.01 (0.01-0.01)	0.01 (0.01-0.01)	0.02 (0.02-0.02)	-
		Resnet	0.01 (0.01-0.01)	0.01 (0.01-0.01)	0.02 (0.02-0.02)	-
	Cla-2	Inception	0.01 (0.01-0.01)	0.01 (0.01-0.01)	0.02 (0.02-0.02)	-
		Resnet	0.01 (0.01-0.01)	0.01 (0.01-0.01)	0.02 (0.02-0.02)	-
	Cla-5	Inception	0.01 (0.01-0.01)	0.01 (0.01-0.01)	0.02 (0.02-0.02)	-
		Resnet	0.01 (0.01-0.01)	0.01 (0.01-0.01)	0.02 (0.02-0.02)	-
	Cla-10	Inception	0.01 (0.01-0.01)	0.01 (0.01-0.01)	0.02 (0.02-0.02)	-
		Resnet	0.01 (0.01-0.01)	0.01 (0.01-0.01)	0.02 (0.02-0.02)	-
MIMICIV	DeepSurv	Inception	0.06 (0.06-0.06)	0.08 (0.08-0.08)	0.11 (0.11-0.11)	0.14 (0.14-0.14)
		Resnet	0.06 (0.06-0.06)	0.08 (0.08-0.08)	0.11 (0.11-0.11)	0.14 (0.14-0.14)
	DeepHit	Inception	0.06 (0.06-0.06)	0.08 (0.08-0.09)	0.12 (0.12-0.13)	0.16 (0.16-0.16)
		Resnet	0.06 (0.06-0.06)	0.08 (0.08-0.08)	0.12 (0.12-0.13)	0.16 (0.15-0.16)
	LH	Inception	0.06 (0.06-0.06)	0.08 (0.08-0.08)	0.11 (0.11-0.11)	0.14 (0.14-0.14)
		Resnet	0.06 (0.06-0.06)	0.08 (0.08-0.08)	0.11 (0.11-0.11)	0.14 (0.14-0.14)
	MTLR	Inception	0.06 (0.06-0.06)	0.08 (0.08-0.08)	0.11 (0.11-0.11)	0.14 (0.14-0.14)
		Resnet	0.06 (0.06-0.06)	0.08 (0.08-0.08)	0.11 (0.11-0.11)	0.14 (0.14-0.14)
	Cla-1	Inception	0.06 (0.06-0.06)	0.08 (0.08-0.08)	0.11 (0.11-0.11)	0.15 (0.15-0.15)
		Resnet	0.06 (0.06-0.06)	0.08 (0.08-0.08)	0.11 (0.11-0.11)	0.15 (0.15-0.15)
	Cla-2	Inception	0.06 (0.06-0.06)	0.08 (0.08-0.08)	0.11 (0.11-0.11)	0.15 (0.15-0.15)
		Resnet	0.06 (0.06-0.06)	0.08 (0.08-0.08)	0.11 (0.11-0.11)	0.15 (0.15-0.15)
	Cla-5	Inception	0.06 (0.06-0.06)	0.08 (0.08-0.08)	0.11 (0.11-0.11)	0.14 (0.14-0.14)
		Resnet	0.06 (0.06-0.06)	0.08 (0.08-0.08)	0.11 (0.11-0.11)	0.14 (0.14-0.15)
	Cla-10	Inception	0.06 (0.06-0.06)	0.08 (0.08-0.08)	0.11 (0.11-0.11)	0.14 (0.14-0.14)
		Resnet	0.06 (0.06-0.06)	0.08 (0.08-0.08)	0.11 (0.11-0.11)	0.14 (0.14-0.15)

Table 8: Area under the precision-recall curve (AUPRC, median, interquartile range) grouped by dataset, task specification, and architecture.

Dataset	Task	Architecture	1-yr AUPRC	2-yr AUPRC	5-yr AUPRC	10-yr AUPRC
Code15	DeepSurv	Inception	0.06 (0.06-0.06)	0.08 (0.08-0.08)	0.12 (0.12-0.13)	
		Resnet	0.07 (0.06-0.07)	0.10 (0.08-0.10)	0.13 (0.12-0.14)	
	DeepHit	Inception	0.06 (0.06-0.06)	0.08 (0.08-0.08)	0.11 (0.11-0.11)	
		Resnet	0.02 (0.02-0.02)	0.03 (0.03-0.04)	0.05 (0.04-0.06)	
	LH	Inception	0.05 (0.05-0.06)	0.08 (0.08-0.08)	0.12 (0.12-0.12)	
		Resnet	0.06 (0.05-0.06)	0.08 (0.07-0.09)	0.11 (0.11-0.12)	
	MTLR	Inception	0.06 (0.05-0.06)	0.09 (0.08-0.09)	0.12 (0.11-0.12)	
		Resnet	0.01 (0.01-0.01)	0.02 (0.02-0.02)	0.03 (0.03-0.03)	
	Cla-1	Inception	0.06 (0.06-0.06)	0.08 (0.08-0.08)	0.12 (0.11-0.12)	
		Resnet	0.05 (0.04-0.06)	0.06 (0.05-0.08)	0.10 (0.08-0.12)	
	Cla-2	Inception	0.06 (0.06-0.06)	0.08 (0.08-0.08)	0.12 (0.12-0.12)	
		Resnet	0.06 (0.04-0.06)	0.08 (0.06-0.08)	0.11 (0.09-0.12)	
	Cla-5	Inception	0.06 (0.06-0.06)	0.08 (0.08-0.08)	0.13 (0.12-0.13)	
		Resnet	0.06 (0.06-0.06)	0.08 (0.08-0.08)	0.12 (0.12-0.12)	
	Cla-10	Inception	0.06 (0.06-0.06)	0.08 (0.08-0.09)	0.13 (0.12-0.13)	
		Resnet	0.06 (0.06-0.07)	0.09 (0.09-0.09)	0.13 (0.13-0.13)	
MIMICIV	DeepSurv	Inception	0.41 (0.41-0.42)	0.46 (0.46-0.46)	0.52 (0.52-0.52)	0.54 (0.53-0.54)
		Resnet	0.42 (0.41-0.42)	0.46 (0.46-0.47)	0.52 (0.52-0.52)	0.54 (0.53-0.54)
	DeepHit	Inception	0.42 (0.41-0.42)	0.45 (0.45-0.45)	0.47 (0.46-0.47)	0.45 (0.44-0.46)
		Resnet	0.42 (0.41-0.43)	0.45 (0.45-0.47)	0.48 (0.47-0.50)	0.47 (0.46-0.50)
	LH	Inception	0.42 (0.42-0.42)	0.46 (0.46-0.46)	0.51 (0.51-0.51)	0.51 (0.51-0.51)
		Resnet	0.42 (0.41-0.43)	0.46 (0.45-0.47)	0.52 (0.51-0.52)	0.52 (0.52-0.52)
	MTLR	Inception	0.42 (0.42-0.43)	0.46 (0.46-0.46)	0.52 (0.51-0.52)	0.53 (0.53-0.53)
		Resnet	0.42 (0.42-0.43)	0.47 (0.46-0.47)	0.52 (0.52-0.52)	0.52 (0.52-0.53)
	Cla-1	Inception	0.42 (0.42-0.42)	0.46 (0.46-0.47)	0.52 (0.51-0.52)	0.53 (0.53-0.53)
		Resnet	0.42 (0.42-0.43)	0.47 (0.47-0.47)	0.52 (0.52-0.52)	0.54 (0.54-0.54)
	Cla-2	Inception	0.42 (0.42-0.42)	0.46 (0.46-0.47)	0.52 (0.52-0.52)	0.54 (0.53-0.54)
		Resnet	0.42 (0.42-0.42)	0.46 (0.46-0.46)	0.52 (0.52-0.52)	0.54 (0.54-0.54)
	Cla-5	Inception	0.40 (0.40-0.40)	0.45 (0.45-0.45)	0.52 (0.52-0.52)	0.54 (0.54-0.54)
		Resnet	0.39 (0.39-0.40)	0.45 (0.44-0.45)	0.51 (0.51-0.52)	0.53 (0.53-0.54)
	Cla-10	Inception	0.38 (0.38-0.38)	0.43 (0.43-0.43)	0.51 (0.51-0.51)	0.53 (0.53-0.53)
		Resnet	0.39 (0.39-0.39)	0.44 (0.44-0.44)	0.51 (0.51-0.51)	0.54 (0.53-0.54)

Table 9: Per-patient Concordance Index and integrated Brier scores (Median, inter-quartile range) for models on the MIMIC-IV dataset. Scores are calculated from 20 bootstrap samples of one random ECG per patient in the test data.

Task	Dataset variable Architecture	MIMICIV Per-Patient Brier	MIMICIV Per-Patient Concordance
Cla-1	Inception	0.17 (0.16, 0.17)	0.82 (0.82, 0.82)
	Resnet	0.17 (0.16, 0.17)	0.83 (0.83, 0.83)
Cla-10	Inception	0.16 (0.15, 0.16)	0.82 (0.82, 0.82)
	Resnet	0.16 (0.16, 0.16)	0.82 (0.82, 0.82)
Cla-2	Inception	0.16 (0.16, 0.16)	0.82 (0.82, 0.83)
	Resnet	0.16 (0.16, 0.17)	0.83 (0.82, 0.83)
Cla-5	Inception	0.16 (0.16, 0.16)	0.82 (0.82, 0.82)
	Resnet	0.16 (0.16, 0.16)	0.82 (0.82, 0.83)
DeepHit	Inception	0.17 (0.17, 0.17)	0.83 (0.83, 0.83)
	Resnet	0.17 (0.17, 0.17)	0.83 (0.83, 0.83)
DeepSurv	Inception	0.15 (0.15, 0.15)	0.83 (0.82, 0.83)
	Resnet	0.16 (0.15, 0.16)	0.83 (0.82, 0.83)
LH	Inception	0.15 (0.15, 0.15)	0.83 (0.82, 0.83)
	Resnet	0.16 (0.15, 0.16)	0.83 (0.82, 0.83)
MTLR	Inception	0.15 (0.15, 0.15)	0.83 (0.83, 0.83)
	Resnet	0.15 (0.15, 0.15)	0.83 (0.83, 0.83)