Heart Failure Hospitalization Risk Models Predict Mortality among Heart Failure Patients and other Groups at Risk

Sergio González

AI Center, Inventec Corporation, Taipei, Taiwan

Abstract

Machine learning based early risk assessment and prognosis prediction of Heart Failure (HF) are beneficial for disease management, but challenging due to the limited availability of large survival datasets. In this paper, we study whether models originally trained to predict the risk of HF hospitalization can be repurposed to estimate mortality. We hypothesize a relationship between hospitalization and mortality risks based on the progressive nature of HF that could be leveraged to unlock data limitations. Using our previously developed models based on 30-second lead I ECG and basic patient information, we evaluate their performance in predicting mortality in multiple cohorts: HF patients in the MUSIC dataset with additional insights into the cause of death and left ventricular ejection fraction, as well as other risk groups in the SaMi-Trop and CODE-15% datasets. Our results demonstrate that our HF hospitalization models are capable of effectively stratifying mortality risk among different populations.

1. Introduction

Heart Failure (HF) [1] is a highly lethal condition often diagnosed at advanced stages, when treatment options are limited and costly. Early detection or identification of individuals at risk would imply a longer and higher quality life for them and a reduction in medical costs. Once diagnosed, usually after first hospitalization, the risk assessment of rehospitalization and mortality also plays a critical role in decision-making for treatment selection and discharge [1, 2].

Survival analysis, especially models based on machine learning (ML), have recently proven useful for these tasks [2–4]. However, they are very data-hungry models that, without sufficient samples, could suffer from overfitting and poor generalization when put into practice. The collection of large survival data is also extremely challenging and expensive, especially for diseases with moderate prevalence such as HF, since a large group of patients needs to be followed for years. Thus, this is a major limitation for their application.

In this paper, we explore the transferability of models trained to assess the risk of HF hospitalization to predict the risk of mortality. Due to the progressive characteristic of HF, there may be a relationship between the risks of HF hospitalization and mortality. By exploiting it, a single model trained with a larger dataset for HF risk prediction could be useful for other tasks, such as mortality prediction, where data limitations may be greater.

For this purpose, we analyzed the performance in predicting mortality of our previously published models trained to assess HF risk from 30-second lead I ECG and basic patient information [3, 4]. We studied their behavior in predicting mortality among HF patients in the MUSIC database [5, 6], giving details related to the cause of death and the Left Ventricular Ejection Fraction (LVEF). Besides, we compared the performance in predicting mortality in other risk groups of the SaMi-Trop [7, 8] and CODE-15% [9, 10] datasets with the previous ECG-based risk score [9]. Our promising results indicate that our models, delivered in a highly accessible manner with *myHeartScore* App [11], would help to manage HF by discriminating patients with a higher mortality risk, in addition to providing early detection.

2. Methods

2.1. Heart Failure Hospitalization Risk Models

Our previously published models for HF risk [3,4] were developed with a cohort of 21,891 subjects, which were followed for HF for a minimum of 2 and up to 11 years, with an average follow-up time of 6.4 years and an average time-to-HF of 1.85 among the 1,805 HF patients (8.2%) at the end of the study. Hospitalization with a primary discharge diagnosis of HF was used as our target event during our study. Neither death nor other clinical events were used as censoring factors. For each patient, a 24-hour 3-channel Holter recording and their basic information were collected prior to follow-up. Additional information includes sex, age, and previous diagnosis of the following

diseases: atrial fibrillation, chronic kidney disease, chronic obstructive pulmonary disease, diabetes mellitus, hyperlipidemia, hypertension, ischemic heart disease, myocardial infarction, stroke, and valvular heart disease.

Before model training, we transformed the 3-channel Holter ECG data to a derived 12-lead ECG data using the EASI linear coefficients [12]. Then, we trained two different models with 30s lead I ECG strips and the aforementioned information to predict the HF hospitalization.

For our feature-based model, we applied bandpass filtering (0.05-45 Hz), delineation, and template extraction pipeline to obtain short-term Heart Rate Variability (HRV) metrics (mean HR, SDNN and SD1 / SD2) and morphological characteristics of the median ECG cycle (PQRST timing and amplitude). These interpretable features, along with patient information, were used to train an XGBoost-based Accelerated Failure Time (AFT) survival model. The model assumed a log-logistic distribution for the time-to-event and was optimized via a regularized negative log-likelihood loss. This approach offers a more interpretable and computationally efficient solution, suitable for deployment in low-resource environments.

Our second model leveraged raw 30s lead I ECG signals to automatically learn relevant representations through a deep ResNet architecture. The extracted features were concatenated with patient data and fed into a discriminator network to predict the probability distribution over 25 discretized time-to-HF intervals. This model is trained using a modified DeepHit loss function that incorporates Focal Loss to address severe class imbalance. In addition, we applied data augmentation through random scaling to improve robustness and generalization.

In our previous work, we proposed the inclusion of approximate long-term HRV to increase the performance of our models [4]. However, we decided to exclude these features from this study, because we did not observe significant differences when including them to estimate mortality risk in the MUSIC dataset. When analyzing the possible reason and the variable distributions from our original dataset and MUSIC dataset, there was a more evident difference in the distributions of some HRV features between hospitalized and non-hospitalized subjects of our dataset compared to survival and dead patients in MUSIC. In addition, SaMi-Trop and CODE-15% lack long-term ECG recordings from which we could extract our HRV features.

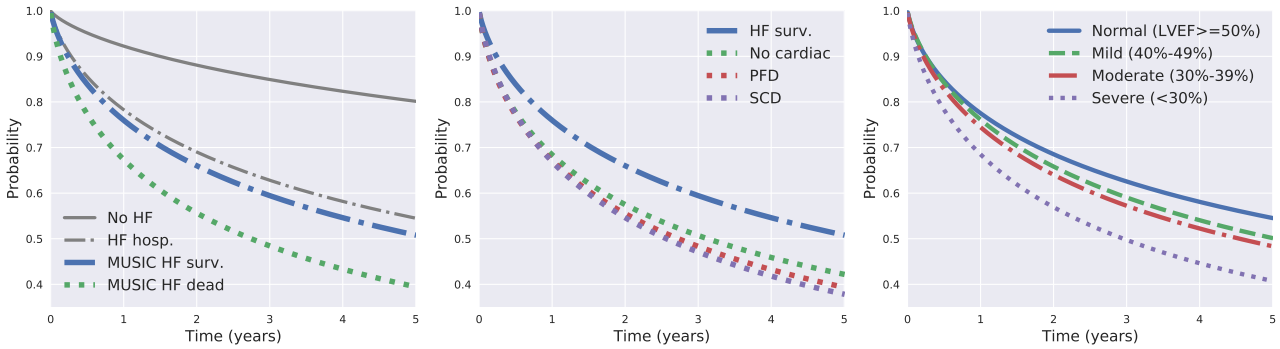
2.2. Data preparation

In our experiments, we used three different datasets with their unique characteristics.

MUSIC dataset [5, 6] was designed to study the risk of cardiac mortality, Sudden Cardiac Death (SCD) or Pump Failure Death (PFD), in patients with chronic HF. The cohort consists of 992 patients followed for an average of 4 years and an average time-to-death of 1.85 years. The dataset includes a 3-orthogonal-lead ECG, 3-lead Holter ECG, and chest X-ray, echocardiography, and blood laboratory features at enrollment. We performed a series of data preparation to match the input of our models. First, we excluded those patients who lost to follow-up and exited the study due to cardiac transplantation. When matching the patient's clinical information, we directly found data for age, sex, diabetes, dyslipidemia, hypertension, and prior myocardial infarction. Atrial fibrillation was assigned according to the rhythms found in their Holter and resting ECGs. Chronic obstructive pulmonary disease was associated with signs of pulmonary venous hypertension. Ischemic heart disease was considered positive if there was prior revascularization. Valvular heart disease was matched with any class of mitral valve insufficiency. We did not find direct or indirect information on prior stroke or chronic kidney disease. Therefore, we set them as negative for each patient in our models. Regarding the ECG signals, we transformed the 3 orthogonal lead ECG into a derived 12 lead ECG using Dower's transform [13] to use the lead I ECG as our input. We randomly selected a 30s lead I ECG strip per patient from the high-resolution ECG data, or the Holter data when the former was missing. The ECG signals were then resampled at 200 Hz to match our original data and preprocessed according to our previous methodology.

SaMi-Trop dataset [7, 8] consists of the first ECG exam of 1631 patients from a prospective study of chronic Chagas cardiomyopathy. ECGs are standard 12-lead ECG tracings of 7 to 10 seconds recorded at 400 Hz. Thus, we resampled them at 200 Hz and padded them with zeros. Then, we preprocessed the lead I ECG signal as usual, considering the original signal length during feature extraction. The average follow-up time and time-to-death were around 2 years and 1 year, respectively. The dataset includes age and sex, but does not contain clinical data or history. Therefore, we assumed that there was no prior disease for each patient when using our models.

CODE-15% dataset [9, 10] contains 345,779 ECG recordings, of which 233647 ECGs of unique patients have survival information. This dataset consists of a more general population with a greater variety of conditions compared to the other two sets. The average follow-up time and time-to-death were 3.7 years and 2 years, respectively. The ECG strips have the same characteristic as the SaMi-Trop and **CODE-15%** also lacks clinical data. Therefore, it has been preprocessed in the same manner.



(a) (b) (c)

Figure 1: Average predicted survival curves of XGBoost-AFT grouped by target (a), cause of death (b) and LVEF (c).

	C-index	iBS	c/d AUC
MUSIC [14]	**0.750**	-	-
XGBoost-AFT	0.653	**0.189**	**0.675**
ResNet	0.637	0.188	0.673
SaMI-Trop [9]	0.532	**0.046**	0.543
XGBoost-AFT	**0.792**	**0.046**	**0.8**
ResNet	0.724	0.051	0.74
CODE-15% [9]	0.541	0.038	0.536
XGBoost-AFT	0.758	0.051	0.759
ResNet	**0.784**	**0.05**	**0.782**

Table 1: Performance of our HF models on mortality risk.

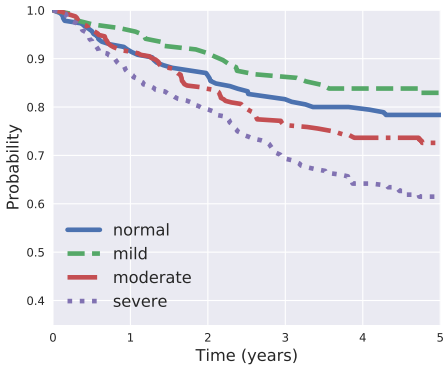


Figure 2: Kaplan-Meier curves for the levels of LVEF.

3. Results

3.1. Mortality risk among Heart Failure Patients: MUSIC dataset

Table 1 collects the results of our models for the MUSIC dataset. Model performance was assessed using C-index, the integrated Brier score (iBS), and the average cumulative/dynamic AUC (c/d AUC) for all time horizons. We include the reported out-of-bag C-index for all death causes of the MUSIC Risk Score [14]. Its iBS and c/d AUC were not reported in the original paper. Our models exhibit decent discrimination and calibration in the MUSIC dataset, with a slight lead from XGBoost-AFT. Considering that our models were not trained on MUSIC data, nor on mortality, and only employ a single-lead ECG and basic patient information, the drop in performance compared to the MUSIC score is reasonable, but it is impressive that they can still perform well on this task.

Figure 1a shows the average predicted survival curves of XGBoost-AFT for the different targets of our original dataset in thin gray lines and the MUSIC dataset in thicker colored lines. The figure visually reflects our model's ability to discriminate between healthier individuals, patients at risk of HF or at a potential early stage, and HF patients at risk of death. The non-hospitalized subjects of our original set (straight gray line) have a significantly higher average survival curve. HF hospitalized individuals in our data (dashed gray line) and MUSIC survival patients (dashed blue line) have similar predicted survival curves, slightly higher in the former group and with a great gap with the non-hospitalized group. Lastly, HF patients who died during the MUSIC study have an average predicted survival curve considerably lower than the rest.

Figure 1b depicts the average predicted survival curves grouped by the different causes of death in the MUSIC dataset. Interestingly, the average survival curves for cardiac death, PFD and SCD, are slightly lower than those for noncardiac death.

Figure 1c shows the average predicted survival curves for the different levels of LVEF or ventricle dysfunction of MUSIC: normal (LVEF>=50%), mild (40%-49%), moderate (30%-39%), and severe (<30%). Our models on av-

erage predict a lower survival for patients with a worse LVEF, which could be reasonable. However, the Kaplan-Meier curves for LVEF levels (Fig. 2) show a higher survival probability for patients with mild dysfunction compared to patients with normal LVEF, which could indicate a possible bias of our model towards HFrEF.

3.2. Mortality risk among other groups at risk: SaMi-Trop and CODE-15%

Table 1 also shows our results in predicting mortality in other population groups: SaMi-Trop and CODE-15%. We included the performance of a previous mortality risk model based on ECG-age prediction with deep learning [9]. Both datasets already included the predicted ages of every data point. Therefore, we reproduced the risk model by training a Cox proportional hazards model with the difference between the given age and the ECG-age as input and the mortality information as target. Our models exhibited considerably better discrimination and calibration than the ECG age predictor for both datasets. XGBoost-AFT performs better in SaMi-Trop and ResNet in CODE-15%. There is a clear performance difference between MUSIC and the other two datasets. It could be caused by additional complexity of HF disease or differences in data distribution, such as a lower mortality rate in SaMi-Trop and CODE-15% compared to MUSIC.

4. Conclusions

In this paper, we explored the transferability of risk models for HF hospitalization to predict mortality. Our results in the MUSIC dataset showed that our models are able to discriminate between healthier individuals, patients at risk or early stages of HF and HF patients at risk of death. Besides, we demonstrated that our models also predict mortality among other groups at risk. Hence, our models, implemented in the *myHeartScore* App [11], would help early detection and management of HF. This outcome would open the door to future studies. A possible advantage to explore is whether these models trained for longer time horizons of HF would maintain their performance in predicting mortality during these periods. Additionally, it would be interesting to study their use as a pre-training for mortality prediction or design a data fusion training, where HF hospitalization and mortality data are simultaneously used during model training.

References

[1] Riccardi M, Sammartino AM, Piepoli M, Adamo M, Pagnesi M, Rosano G, et al. Heart failure: an update from the last years and a look at the near future. ESC Heart Failure 2022;9(6):3667–3693.

[2] Petmezas G, Papageorgiou VE, Vassilikos V, Pagourelias E, Tsaklidis G, Katsaggelos AK, et al. Recent advancements and applications of deep learning in heart failure: A systematic review. Computers in Biology and Medicine 2024; 108557.

[3] González S, Hsieh WT, Burba D, Chen TPC, Wang CL, Wu VCC, et al. Interpretable estimation of the risk of heart failure hospitalization from a 30-second electrocardiogram. In 2022 E-Health and Bioengineering Conference (EHB). IEEE, 2022; 1–4.

[4] González S, Yi AKC, Hsieh WT, Chen WC, Wang CL, Wu VCC, et al. Multi-modal heart failure risk estimation based on short ECG and sampled long-term HRV. Information Fusion 2024;107:102337.

[5] Martín-Yebra A, Bayés de Luna A, Vázquez R, Caminal P, Laguna P, Martínez J. The MUSIC database: Sudden cardiac death in heart failure patients. In Computing in Cardiology, volume 51. 2024; 1–4.

[6] Martin-Yebra A, Martínez JP, Laguna P. MUSIC (sudden cardiac death in chronic heart failure) (1.0.1) [dataset], 2025. *PhysioNet*.

[7] Cardoso CS, Sabino EC, Oliveira CDL, de Oliveira LC, Ferreira AM, Cunha-Neto E, et al. Longitudinal study of patients with chronic chagas cardiomyopathy in brazil (SaMi-Trop project): a cohort profile. BMJ open 2016; 6(5):e011181.

[8] Ribeiro ALP, Ribeiro AH, Paixao GMM, Lima EM, Horta Ribeiro M, Pinto Filho MM, et al. SaMi-Trop: 12-lead ecg traces with age and mortality annotations (1.0.0) [dataset], 2021. *Zenodo*.

[9] Lima EM, Ribeiro AH, Paixão GM, Ribeiro MH, Pinto-Filho MM, Gomes PR, et al. Deep neural network-estimated electrocardiographic age as a mortality predictor. Nature communications 2021;12(1):5117.

[10] Ribeiro AH, Paixao G, Lima EM, Ribeiro MH, Pinto Filho MM, Gomes PR, et al. CODE-15%: a large scale annotated dataset of 12-lead ECGs (1.0.0) [dataset], 2021. *Zenodo*.

[11] Inventec Corporation. myHeartScore (1.2.1) [App]. App Store `https://apps.apple.com/app/myheartscore/id6462642173`, 2025.

[12] Feild DQ, Feldman CL, Hor BM, et al. Improved easi coefficients: their derivation, values, and performance. Journal of electrocardiology 2002;35(4):23–33.

[13] Vondrak J, Penhaker M. Review of processing pathological vectorcardiographic records for the detection of heart disease. Frontiers in Physiology 2022;13:856590.

[14] Vazquez R, Bayes-Genis A, Cygankiewicz I, Pascual-Figal D, Grigorian-Shamagian L, Pavon R, et al. The music risk score: a simple method for predicting mortality in ambulatory patients with chronic heart failure. European heart journal 2009;30(9):1088–1096.

Address for correspondence:

Sergio González
111 No. 166, Sec. 4, Chengde Rd., Shilin Dist., Taipei, Taiwan
gonzalez-vazquez.sergio@inventec.com