


Machine learning algorithms for claims data-based prediction of in-hospital mortality in patients with heart failure

Sebastian König^{1,2,*†} , Vincent Pellissier^{2†}, Sven Hohenstein², Andres Bernal², Laura Ueberham^{1,2}, Andreas Meier-Hellmann³, Ralf Kuhlen⁴, Gerhard Hindricks^{1,2} and Andreas Bollmann^{1,2}

¹Department of Electrophysiology, Heart Center Leipzig at University of Leipzig, Strümpellstraße 39, Leipzig, 04289, Germany; ²Leipzig Heart Institute, Leipzig, Germany; ³Helios Hospitals, Berlin, Germany; and ⁴Helios Health, Berlin, Germany

Abstract

Aims Models predicting mortality in heart failure (HF) patients are often limited with regard to performance and applicability. The aim of this study was to develop a reliable algorithm to compute expected in-hospital mortality rates in HF cohorts on a population level based on administrative data comparing regression analysis with different machine learning (ML) models.

Methods and results Inpatient cases with primary International Statistical Classification of Diseases and Related Health Problems (ICD-10) encoded discharge diagnosis of HF non-electively admitted to 86 German Helios hospitals between 1 January 2016 and 31 December 2018 were identified. The dataset was randomly split 75%/25% for model development and testing. Highly unbalanced variables were removed. Four ML algorithms were applied, and all algorithms were tuned using a grid search with multiple repetitions. Model performance was evaluated by computing receiver operating characteristic areas under the curve. In total, 59 125 cases (69.8% aged 75 years or older, 51.9% female) were investigated, and in-hospital mortality was 6.20%. Areas under the curve of all ML algorithms outperformed regression analysis in the testing dataset with values of 0.829 [95% confidence interval (CI) 0.814–0.843] for logistic regression, 0.875 (95% CI 0.863–0.886) for random forest, 0.882 (95% CI 0.871–0.893) for gradient boosting machine, 0.866 (95% CI 0.854–0.878) for single-layer neural networks, and 0.882 (95% CI 0.872–0.893) for extreme gradient boosting. Brier scores demonstrated a good calibration especially of the latter three models.

Conclusions We introduced reliable models to calculate expected in-hospital mortality based only on administrative routine data using ML algorithms. A broad application could supplement quality measurement programs and therefore improve future HF patient care.

Keywords Mortality prediction; Prediction models; Machine learning; In-hospital mortality; Heart failure

Received: 26 February 2021; Revised: 30 March 2021; Accepted: 21 April 2021

*Correspondence to: Sebastian König, Department of Electrophysiology, Heart Center Leipzig at University of Leipzig, Strümpellstraße 39, 04289 Leipzig, Germany. Tel: +49-341-865-252-613; Fax: +49-341-865-1460. Email: sebastian.koenig@helios-gesundheit.de

†These authors contributed equally to this paper.

Introduction

Heart failure (HF) is still of highest socio-economic relevance and despite increased survival rates one leading cause of overall mortality and hospitalizations.^{1–3} To further improve HF treatment calls for comparative value-based patient care programs and standardized quality metrics.^{4,5} Numerous scores predicting HF-related mortality or hospital readmissions have

been introduced but were still rarely used in clinical practice due to modest performance, insufficient applicability, or lacking external validation.⁶ Moreover, many prediction tools were developed using highly selected datasets collected for other purposes that hinder generalizability.⁷ The evaluation of large datasets and the application of machine learning (ML) algorithms have the potential to enhance model performance, with already encouraging results in the context of HF

hospitalization.^{8–10} Most existing tools predicting mortality either failed to implement ML models or required variables that were not easily available on a population level.^{6,11} This is, however, relevant in order to retrospectively compare disease-related outcomes between cohorts and health care facilities in a standardized way rather than on an individual basis at hospital admission to evaluate influencing factors and facilitate quality care management. Therefore, the aim of this study was to develop algorithms computing expected in-hospital mortality rates in HF patient cohorts using a nationwide, real-world administrative dataset that also includes variables that could only be gathered retrospectively. Doing so, we compared different ML methods with each other and with a classic regression approach regarding their predictive performance.

Methods

Data source

Administrative data of 86 Helios hospitals in Germany were retrospectively analysed. Patient cases with inpatient treatment within 1 January 2016 to 31 December 2018 and a main discharge diagnosis of HF defined in accordance to prior publications were identified.¹² Types of admission and discharge were gathered from administrative data, and only cases with both urgent (non-elective) hospital admission and hospital discharge type other than hospital transfer were further studied. In-hospital death as the outcome of interest has been defined based on hospital discharge type. Discharge diagnoses were encoded by the International Statistical Classification of Diseases and Related Health Problems [ICD-10-GM (German Modification)]. Relevant co-morbidities were identified from encoded secondary diagnoses within hospital discharge data according to the Elixhauser co-morbidity score as defined previously without a distinction being made between pre-existing co-morbidities and new medical conditions.^{13,14} Cases with missing information for New York Heart Association (NYHA) class ($n = 5315$ cases) were discarded due to adequate calibration of ML models. Detailed information regarding used ICD codes and a comparison of datasets with and without the exclusion of cases with missing NYHA class is provided in the Supporting Information, Tables S1–S3. The investigation conforms with the principles outlined in the Declaration of Helsinki. Given the anonymized data analysis of administrative data, ethics committee approval was determined not to be required in accordance with German law [Professional Code for Physicians (Saxony) §15]. Due to the retrospective study of anonymized data, informed consent has not been obtained.

Statistical analysis

All analyses were performed within the R environment for statistical computing (Version 3.6.1, 64-bit build).¹⁵ The dataset has been randomly split using 75% for model development and 25% for model testing. Within the development dataset, models were evaluated using three concurrent variable settings with all of them implementing baseline variables (age, gender, admission year, NYHA class, length of hospital stay, and length of intensive care unit stay) and each one set containing Elixhauser co-morbidities as separate variables, Elixhauser weighted co-morbidity scores (Elixhauser score), or Elixhauser weighted co-morbidity score quintiles (Elixhauser index). Because no cross-linking of patients' cases between different hospitals was possible, no variables based on historic information (e.g. previous HF-associated hospitalizations) were used. Variables that are only known at hospital discharge (e.g. length of stay) were included because the aim of this project was to develop models for comparing expected and observed mortality on a population basis under different circumstances rather than creating a tool for individual outcome prediction. Variables being highly sparse and unbalanced (near-zero variance) were removed prior to the analysis. Near-zero variance variables were defined as variables with a per cent of unique values (number of unique value/number of samples * 100) below 10% and a frequency ratio (frequency of most prevalent value over the second most prevalent value) over 95/5. All variables were scaled and centred before the analyses.

The three development data subsets were evaluated using five algorithms: logistic regression [generalized linear models (GLM)], random forest (RF), gradient boosting machine (GBM), single-layer neural network (NNET), and extreme gradient boosting (XGBoost). Even though ML algorithms can implicitly take into account interactions between variables and non-linearity, we decided not to specify interactions or non-linear effects in the GLM for several reasons: (i) to keep the model specification as simple as possible for comprehensibility, (ii) to reflect the approaches already existing in the literature, and (iii) because non-linearity was not expected from a clinical perspective.^{16,17} The algorithms were tuned using a grid search with a k-fold approach, using three repetitions of 10-fold each. During each repetition, the training dataset was split into 10 equal chunks, and each model was run with 9/10 of the training dataset for every combination of hyper-parameters and evaluated on the remaining 1/10. This has been performed 10 times for each repetition so that each 10th was used once for validation. After all iterations, the hyper-parameter values maximizing the area under the precision-recall curve (AUPRC) during cross-validation process were selected, and the model was run on the whole training set. To evaluate the performance of the models trained, the values predicted during the cross-validation process were used to compute receiver operating characteristic

(ROC) areas under the curve (AUCs), and the model with the highest AUPRC was considered the best. Accuracy has not been chosen because of the outcome imbalance (less than 10% of the admissions resulted in in-hospital mortality). The relative importance of variables was assessed after the training, and values were scaled to 100 in order to make them comparable across algorithms. The methods used to assess the relative importance of variable were algorithm specific (GLM: absolute value of the *t*-statistic; RF: average across all trees of the difference between out-of-bag accuracy and out-of-bag accuracy after prediction permutation, normalized by the standard error; GBM and XGBoost: sum across all boosting iteration of the reduction in squared error attributed to a variable in each split based on this variable; and NNET: combination of the absolute values of connecting weights).

Two final steps were carried out before the final evaluation of each model with the test dataset:

- The algorithm not based on likelihood or non-linear procedure approximated probabilities only, and hence, the probabilities had to be recalibrated. Here, we trained a logistic generalized additive model to be monotonically increasing, predicting the outcome with the probabilities computed by the initial model using the training set. We then derive probabilities from the test set with our initial algorithm and scale these probabilities with generalized additive model.^{18,19} Improvement or degradation of the prediction with the recalibration was estimated with the Brier score.
- Due to high outcome imbalances, expected mortality probabilities were likely to be low and the classification threshold had to be changed (i.e. the threshold above which an event is classified positive). Here, threshold has been chosen based on the ROC curve, selecting the threshold that maximizes the F1 statistic.

Using the probabilities predicted within the test data and the optimal threshold, the predictive abilities of each algorithm were assessed by ROC, the precision-recall curve, calibration-in-the-large (overall expected and observed mortality rate), weak calibration (intercept and slope of the calibration curve), calibration plots, and AUCs.²⁰ A continuous net reclassification improvement index has been computed comparing GLM with each ML algorithm in order to examine differences between model performance in addition to the comparison of the aforementioned metrics with values above zero indicating superiority.

Results

We included 59 125 patient cases from 69 German Helios hospitals into our analysis. The majority of patients were

75 years or older (69.8%), 48.1% were male, and most patients were highly symptomatic with 42.0% and 47.4% presenting with NYHA classes III and IV, respectively. In-hospital mortality was 6.20% overall. Patient cohorts used for model training and model testing were well balanced with regard to age, gender, NYHA class, and Elixhauser co-morbidities. Baseline characteristics as well as a comparison of training and testing dataset are summarized in the Supporting Information, *Table S4*. In univariable regression analysis, higher age, longer stay at the intensive care unit, male sex, higher NYHA class, and several variables from the Elixhauser co-morbidity score were predictors of in-hospital mortality (*Table 1*).

Model training

During the training process, the hyper-parameters of each algorithm (except for GLM) were tuned keeping the following values (three values are specified for variable sets containing the Elixhauser co-morbidities, the Elixhauser score, or the Elixhauser index):

- RF: number of variables randomly selected at each split = 3/2/2, number of trees = 500;
- GBM: number of trees = 300; maximum depth = 3, learning rate = 0.1, minimum number of observations in each node = 2;
- NNET: number of units in the hidden layer = 8/11/11, learning rate = 0.6/0.4/0.3; and
- XGBoost: maximum number of boosting iterations = 100, maximum depth = 5; learning rate = 0.1, minimum loss reduction = 0.2; proportion of columns sampled per tree = 0.5; minimum child weight = 5; proportion of rows sampled per tree = 0.8.

Comparing the three aforementioned variable sets, ROC AUCs for the GLM were 0.820, 0.812, and 0.809. AUCs for all ML models were higher including RF (0.858, 0.858, and 0.857), GBM (0.867, 0.863, and 0.863), NNET (0.859, 0.859, and 0.858), and XGBoost (0.867, 0.865, and 0.863). For further model testing, the variable set containing the single Elixhauser co-morbidities was used because the corresponding models showed similar or superior ROC AUC values compared with both other variable sets in all models. Specific variable importance values for the baseline variables and Elixhauser co-morbidities are listed in the Supporting Information, *Table S5*.

Model testing

All ML models performed better than GLM with respect to the ROC AUC. Within ML models, GBM and XGBoost showed

Table 1 Event rates and univariable regression analysis results for in-hospital mortality

Variable ^a	Event rate ^{b,c} [% (n/N)]	Univariable analysis	
		OR (95% CI)	P-value
Age			
<65 years	2.3 (175/7459)		
65–74 years	3.7 (385/10 377)	1.60 (1.34–1.92)	<0.001
>74 years	7.5 (3117/41 289)	3.39 (2.91–3.97)	<0.001
Length of stay			
<5 days	7.8 (1713/21 981)		
5–9 days	3.8 (740/19 392)	0.47 (0.43–0.51)	<0.001
>9 days	6.9 (1224/17 752)	0.88 (0.81–0.95)	<0.001
Length of ICU stay			
0 days	4.2 (2008/47 485)		
>0 days	14.3 (1669/11 640)	3.79 (3.54–4.06)	<0.001
NYHA class			
NYHA class II	0.5 (24/5289)		
NYHA class III vs. NYHA class I/II	2.2 (542/24 842)	4.89 (3.25–7.37)	<0.001
NYHA class IV vs. NYHA class I/II	11.1 (3111/28 027)	27.4 (18.3–40.9)	<0.001
Gender			
Female	6.0 (1850/30 689)		
Male	6.4 (1827/28 436)	1.07 (1.00–1.14)	0.046
Elixhauser co-morbidity score			
Cardiac arrhythmias	6.6 (2451/36 921)	1.22 (1.13–1.31)	<0.001
Chronic pulmonary disease	6.5 (754/11 515)	1.07 (0.99–1.16)	0.103
Chronic renal failure	6.7 (2495/37 250)	1.26 (1.17–1.35)	<0.001
Deficiency anaemia	6.2 (197/3195)	0.99 (0.85–1.15)	0.898
Depression	5.1 (158/3121)	0.79 (0.68–0.94)	0.006
Diabetes, complicated	6.6 (864/13 028)	1.09 (1.01–1.18)	0.027
Diabetes, uncomplicated	5.9 (628/10 642)	0.93 (0.86–1.02)	0.134
Fluid and electrolyte disorders	10.9 (2010/18 504)	2.85 (2.66–3.05)	<0.001
Hypertension, complicated	4.2 (1246/29 343)	0.50 (0.47–0.54)	<0.001
Hypertension, uncomplicated	6.1 (1087/17 800)	0.97 (0.90–1.05)	0.458
Hypothyroidism	4.8 (381/7935)	0.73 (0.66–0.82)	<0.001
Obesity	4.1 (567/13 759)	0.58 (0.53–0.64)	<0.001
Peripheral vascular disease	7.2 (550/7633)	1.20 (1.09–1.32)	<0.001
Pulmonary circulation disorder	6.2 (705/11 357)	0.99 (0.92–1.09)	0.955
Valvular heart disease	5.8 (1289/22 269)	0.89 (0.83–0.95)	<0.001
Weight loss	15.3 (542/3548)	3.02 (2.73–3.33)	<0.001

CI, confidence interval; ICU, intensive care unit; NYHA, New York Heart Association; OR, odds ratio.

^aOnly variables after feature selection are shown.

^bFor all variables, event rates and proportions of events/patients were given.

^cFor all Elixhauser co-morbidities, event rates and proportions of events/patients were given for the group of present co-morbidity.

the highest ROC AUC values, but 95% confidence intervals (CIs) overlapped with those of RF and NNET. In detail, AUCs were 0.829 (95% CI 0.814–0.843) for GLM, 0.875 (95% CI 0.863–0.886) for RF, 0.882 (95% CI 0.871–0.893) for GBM, 0.866 (95% CI 0.854–0.878) for NNET, and 0.882 (95% CI 0.872–0.893) for XGBoost. This was confirmed by net reclassification improvement index calculation for the comparison of GLM with the different ML models: GLM vs. RF 0.372 (95% CI 0.306–0.438, $P < 0.001$), GLM vs. GBM 0.469 (95% CI 0.404–0.534, $P < 0.001$), GLM vs. NNET 0.569 (95% CI 0.504–0.634, $P < 0.001$), and GLM vs. XGBoost 0.685 (0.621–0.749, $P < 0.001$). Corresponding AUPRCs were 0.309 (95% CI 0.280–0.339), 0.456 (95% CI 0.424–0.489), 0.476 (95% CI 0.444–0.508), 0.446 (95% CI 0.414–0.478), and 0.477 (95% CI 0.445–0.509), respectively. Results of AUPRCs are illustrated in *Figure 1*.

Calibration metrics are shown in *Table 2*, and calibration plots are illustrated in *Figure 2*. Brier scores (uncalibrated)

were 0.050 for GLM, 0.045 for RF, 0.043 for both GBM and XGBoost, and 0.044 for NNET. A recalibration of probabilities did not improve Brier scores relevantly. Expected and observed event rates were compared overall and within subsets of different age groups and the year of admission (*Figure 3*).

Discussion

With the present study, we showed that the calculation of reliable models for the calculation of expected in-hospital mortality rates on a population basis is possible by using widely available administrative data only. Model performance was improved when implementing ML algorithms compared with a classic regression analysis, with GBM and XGBoost providing the most encouraging results. These

Figure 1 Summarized areas under the precision-recall curve from the testing dataset for all investigated models. GBM, gradient boosting machine; GLM, generalized linear models; NNET, single-layer neural network; RF, random forest; XGBoost, extreme gradient boosting.

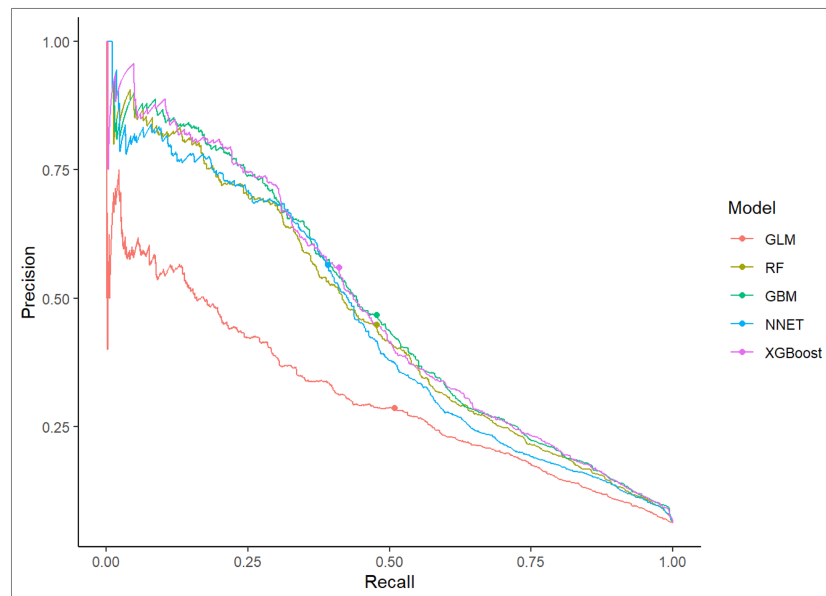


Table 2 Calibration metrics for all investigated models within the test dataset

	Calibration-in-the-large	Calibration intercept	Calibration slope
GLM	6.2% vs. 6.2%	0.00 (−0.068 to 0.074)	1.05 (0.981–1.111)
RF	6.2% vs. 6.4%	−0.03 (−0.100 to 0.044)	1.35 (1.274–1.427)
GBM	6.2% vs. 6.4%	−0.05 (−0.125 to 0.030)	1.06 (1.008–1.119)
NNET	6.2% vs. 6.3%	−0.01 (−0.086 to 0.069)	1.01 (0.951–1.061)
XGBoost	6.2% vs. 6.3%	−0.03 (−0.102 to 0.051)	1.10 (1.041–1.158)

GBM, gradient boosting machine; GLM, generalized linear models; NNET, single-layer neural network; RF, random forest; XGBoost, extreme gradient boosting.

kinds of models could be used in a second step to compute a standardized mortality ratio for further scientific evaluations and quality management programs by comparing temporal or regional disparities between expected and observed event rates in specific patient cohorts retrospectively, which could also be interesting with respect to hospital benchmarking.

To date, prediction tools for mortality in HF patients are rarely used in routine practice outside clinical studies, both overall and especially in the setting of mortality prediction on a population level rather than on an individual basis. This is most likely caused by either the lacking availability of required variables on a population level or poor predictive performance of less complex scores.^{6,11} Existing risk stratification models for mortality of individual HF patients (with most of them relating to long-term mortality risk) showed a wide range with regard to C-statistics ranging from 0.60 to 0.89 for different follow-up periods and HF subgroups with all models with values >0.80 either using

data of laboratory results, clinical metrics, or medication data except from one.¹¹ Reflecting this fact, N-terminal pro-brain natriuretic peptide, serum creatinine, blood urea nitrogen, systolic blood pressure, heart rate, and left ventricular ejection fraction are under the most commonly reported variables used for patient-based outcome prediction.⁶ However, implementing those or even more disease-specific values did not result in a further improvement of statistic performance neither in established HF risk scores for long-term risk prediction nor in newly introduced models.^{21–25} This is also true for common clinical risk prediction models of individual in-hospital mortality in HF patients.^{26,27} All of those prediction tools estimate the individual, patient-based risk, which was not the intention of our project, and referencing them is therefore not meant to be as a comparison with our results. However, it shows an existing conflict between variable availability and model integration to a larger scale. Only two prior studies solely focused on administrative data computing

Figure 2 Calibration plots from the testing dataset for all investigated models. GBM, gradient boosting machine; GLM, generalized linear models; NNET, single-layer neural network; RF, random forest; XGBoost, extreme gradient boosting.

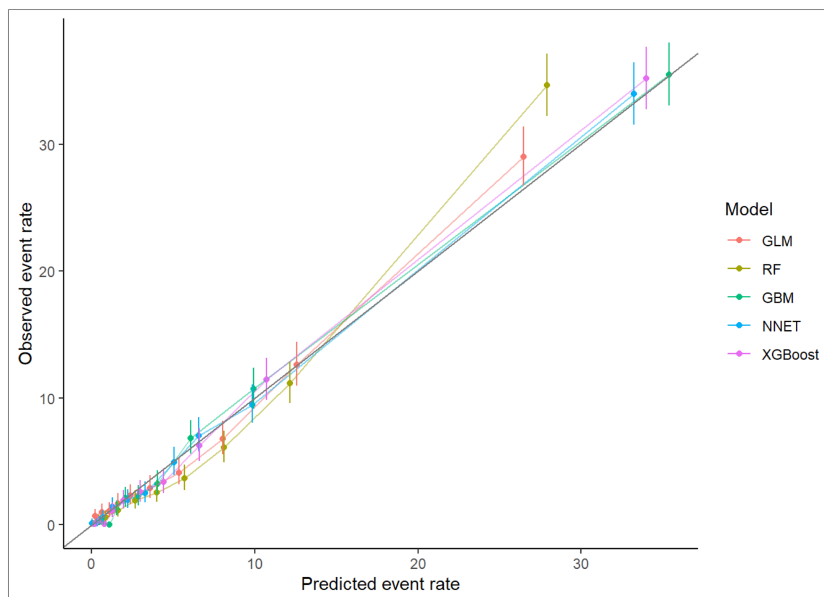
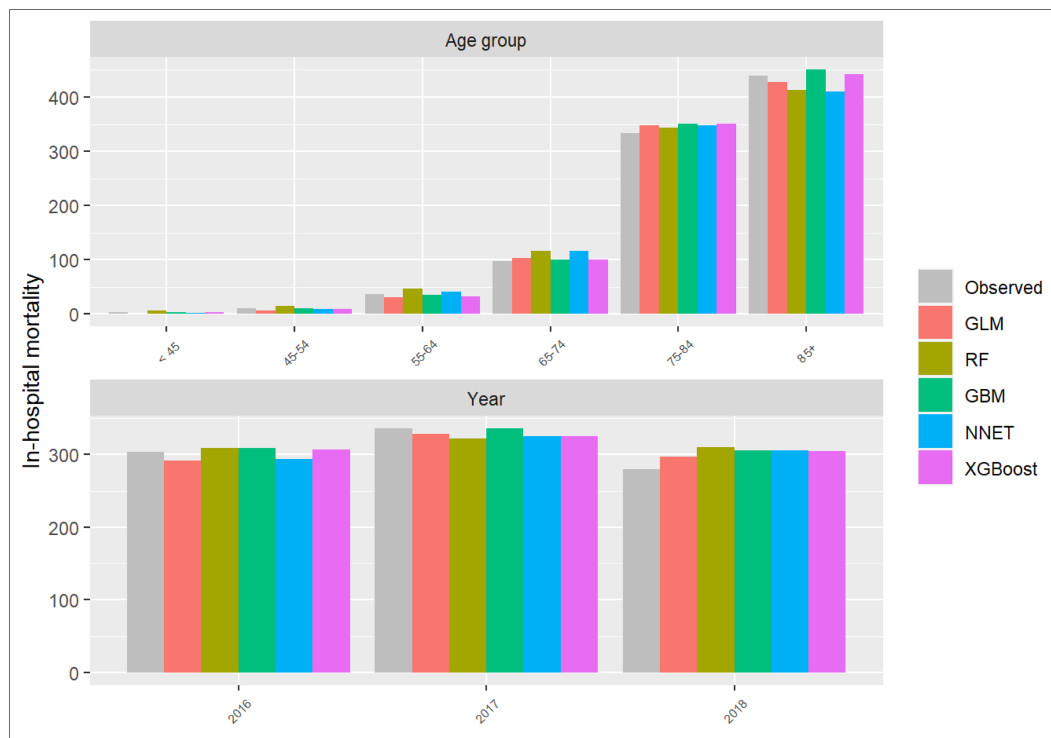


Figure 3 Comparison of predicted and observed in-hospital mortality categorized by age and admission year. GBM, gradient boosting machine; GLM, generalized linear models; NNET, single-layer neural network; RF, random forest; XGBoost, extreme gradient boosting.



overall acceptable AUCs of 0.72 to 0.78 with only one of them predicting in-hospital mortality specifically.^{28,29} Nevertheless, a direct comparison with our models was also not possible because several information included in both referenced analyses were not available from our dataset (medication use, number of prior HF-related hospitalizations, ethnicity, and insurance information). The observed event rate in our cohort was rather higher when compared with the cohort of Lagu *et al.*, which is likely to influence the model quality in the development phase potentially leading to higher AUCs. Differing baseline characteristics including a higher proportion of patients older than 75 years could be one explanation for higher in-hospital mortality rates in our HF cohort. The better performance of both our regression models and ML algorithms when compared with the work of Desai and colleagues may be linked to the comparatively low number of patients of less than 10 000 cases in their database exceeded more than six-fold by our cohort. Interestingly, they also analysed different ML algorithms with GBM showing the best AUC value but without a significant difference to a classic regression model.²⁸ Other studies examining ML methods for mortality prediction reported of comparable or lower AUC values but did so in either different patient cohorts (both inpatients and outpatients) or investigating different endpoints.^{8,21} RF and dynamic radius means were presented as the most reliable models, which differs from our findings. However, the different subsets of implemented variables as well as the already mentioned methodological deviations should be major influencing factors.

When looking for other cardiovascular studies comparing several ML methods for endpoint prediction aside from mortality in HF, there are publications supporting different models also investigated in our manuscript. In a recent analysis of data from GARFIELD-AF and ORBIT-AF, predictive ability of GBM and GLM was comparable whereas neural network models had the lowest discriminatory power.³⁰ In contrast, another study propagated neural networks to be superior to other ML methods under several circumstances with regard to data diversity and density when predicting incident HF.³¹ Our results are in between showing no significant differences between NNET and RF, GBM, or XGBoost but with a better performance compared with GLM. Two other studies propagated RF models to outperform established risk prediction tools including GLM, which is in line with our findings.^{32,33} Nevertheless, both of them did not compare RF with other ML algorithms, which would be of interest because values of AUC and AUPRC were slightly worse for RF than those of GBM and XGBoost in our calculations, even when not meeting statistical significance. Lastly, in an Australian multicentre analysis of almost 40 000 patients, several tested ML models showed higher predictive abilities in comparison with classic GLM in predicting in-hospital mortality following out-of-hospital cardiac arrest, which could be transferred to our

findings.³⁴ Whether an augmentation by data from electronic medical records, especially clinical and lab data, has the potential to further improve our models will be subject of future research.^{28,35} However, one goal of our analysis was the introduction of a reliable model, which can be reproduced from widely available data.

Our models are not intended as mortality prediction tools for individual patients resulting in therapy adaptations. Rather, after future validation, it is conceivable that those models could contribute to further scientific research as well as quality management programs by the implementation of risk-adjusted and therefore standardized calculation of expected mortality rates in large HF cohorts comparing different geographic regions, hospitals, or time periods. In a second step, a broad integration is imaginable at least in the context of quality management programs. Furthermore, the impact of certain external factors such as the ongoing pandemic caused by SARS-CoV-2 on health care of HF patients could be examined by comparing expected and observed in-hospital mortality rates.

Limitations

Retrospectively collected data are widely considered to be of inferior quality compared with datasets with a prospective data assessment. Nevertheless, existing literature suggested that data collection mode per se did not influence the discriminatory power of the derived model.¹¹ The analysis of data derived from the Helios hospital network only could possibly result in a selection bias with respect to local differences regarding socio-economic factors, ethnicity, and other factors. However, due to the size of our cohort and the distribution of Helios hospitals across Germany serving about 7–8% of all German inpatient cases, we presume our database to be representative. Moreover, standardized treatment protocols within our hospital group may influence care pathways and outcome. Nonetheless, this analysis includes data from all hospital classes, including community hospitals as well as tertiary or university centres, which should partly alleviate the last-mentioned bias.

Our study analysed administrative data not being stored for research interests but for remuneration reasons, which could potentially affect the encoded information. Quality of the results depends to a large extent on the correct encoding of hospital discharge diagnoses.^{14,36} However, regarding the main discharge diagnosis and the adequacy of hospitalization as well as encoding, there is a continuous evaluation by reimbursement companies/health insurances, which supports the assumption of overall valid information with respect to the diagnoses and the appropriateness of the patient's hospitalization. Moreover, used ICD codes have been validated internally using electronic medical records in previous works

of this working group.^{12,37} Information regarding patients' specific medical history, time of the onset of symptoms, cardiac imaging, laboratory results, medication, and treatment-related data was not available due to the type and structure of the analysed database. In particular, NYHA classification is influenced by the subjective assessment of the treating physician, and the addition of more objective criteria of disease severity is not possible when only using administrative data. Moreover, information regarding prior HF-related hospitalizations would have been likely to improve the performance of our models. Finally, an external validation of our models is required.

Conclusions

Machine learning algorithms are reliable metrics to calculate expected in-hospital mortality rates in HF patients outperforming regression analyses in a large, multicentre, real-world administrative database. Identifying a model based on widely available data with high predictive power could supplement quality measurement programs and therefore improve future HF patient care.

Acknowledgement

Open access funding enabled and organized by Projekt DEAL.

Conflict of interest

G.H. received grants through the Leipzig Heart Institute from Boston Scientific (Boston Scientific Corporation, Marlborough, Massachusetts, USA) and Abbott/St. Jude Medical (Abbott Laboratories, Chicago, Illinois, USA); no personal payments are declared. All other authors state that there is nothing to declare.

References

1. Conrad N, Judge A, Canoy D, Tran J, Pinho-Gomes AC, Millett ERC, Salimi-Khorshidi G, Cleland JG, McMurray JJV, Rahimi K. Temporal trends and patterns in mortality after incident heart failure: a longitudinal analysis of 86000 individuals. *JAMA Cardiol* 2019; 4: 1102–1111.
2. Taylor CJ, Ordonez-Mena JM, Roalfe AK, Lay-Flurrie S, Jones NR, Marshall T, Hobbs FR. Trends in survival after a diagnosis of heart failure in the United Kingdom 2000–2017: population based cohort study. *BMJ* 2019; 364: 1223.
3. Tromp J, Ferreira JP, Janwanishstaporn S, Shah M, Greenberg B, Zannad F, Lam CSP. Heart failure around the world. *Eur J Heart Fail* 2019; 21: 1187–1196.
4. Joynt Maddox K, Bleser WK, Crook HL, Nelson AJ, Hamilton Lopez M, Saunders RS, McClellan M, Brown N, American Heart Association Value-Based Models Learning Collaborative. Advancing value-based models for heart failure: a call to action from the value in healthcare initiative's value-based models learning collaborative. *Circ Cardiovasc Qual Outcomes* 2020; 13: e006483.
5. Wadhera RK, Vaduganathan M, Jiang GY, Song Y, Xu J, Shen C, Bhatt DL, Yeh RW, Fonarow GC. Performance in federal value-based programs of hospitals recognized by the American Heart Association and American College of Cardiology for high-quality heart failure

Funding

There has been no funding in connection with this study.

Permissions information

The authors do hereby declare that all illustrations and figures in the manuscript are entirely original and do not require reprint permission.

Data availability statement

The data underlying this article will be shared on reasonable request to the corresponding author. The same applies to the codes used to perform the analyses.

Supporting information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Table S1: ICD-codes used as tracers for the identification of HF and for the definition of NYHA classes.

Table S2: ICD-codes used to calculate Elixhauser Comorbidity Score (weighting according to AHRQ algorithm).

Table S3: Comparison of the total dataset and the cohort after exclusion of cases with missing NYHA classes.

Table S4: Baseline characteristics overall and as a comparison for training and testing dataset.

Table S5: Variable importance values for all tested models.

- and acute myocardial infarction care. *JAMA Cardiol* 2020; **5**: 515–521.
6. Di Tanna GL, Wirtz H, Burrows KL, Globe G. Evaluating risk prediction models for adults with heart failure: a systematic literature review. *PLoS One* 2020; **15**: e0224135.
 7. Gottdiener JS, Fohner AE. Risk prediction in heart failure: new methods, old problems. *JACC Heart Fail* 2020; **8**: 22–24.
 8. Adler ED, Voors AA, Klein L, Macheret F, Braun OO, Urey MA, Zhu W, Sama I, Tadel M, Campagnari C, Greenberg B, Yagil A. Improving risk prediction in heart failure using machine learning. *Eur J Heart Fail* 2020; **22**: 139–147.
 9. Jing L, Ulloa Cerna AE, Good CW, Sauer NM, Schneider G, Hartzel DN, Leader JB, Kirchner HL, Hu Y, Riviello DM, Stough JV, Gazes S, Haggerty A, Raghunath S, Carry BJ, Haggerty CM, Fornwalt BK. A machine learning approach to management of heart failure populations. *JACC Heart Fail* 2020; **8**: 578–587.
 10. Shameer K, Johnson KW, Yahi A, Miotto R, Li LI, Ricks D, Jebakaran J, Kovatch P, Sengupta PP, Gelijns S, Moskovitz A, Darrow B, David DL, Kasarskis A, Tatonetti NP, Pinney S, Dudley JT. Predictive modeling of hospital readmission rates using electronic medical record-wide machine learning: a case-study using Mount Sinai heart failure cohort. *Pac Symp Biocomput* 2017; **22**: 276–287.
 11. Rahimi K, Bennett D, Conrad N, Williams TM, Basu J, Dwight J, Woodward M, Patel A, McMurray J, MacMahon S. Risk prediction in patients with heart failure: a systematic review and analysis. *JACC Heart Fail* 2014; **2**: 440–446.
 12. König S, Hohenstein S, Meier-Hellmann A, Kuhlen R, Hindricks G, Bollmann A, Helios Hospitals. In-hospital care in acute heart failure during the COVID-19 pandemic: insights from the German-wide Helios hospital network. *Eur J Heart Fail* 2020; **22**: 2190–2201.
 13. Moore BJ, White S, Washington R, Coenen N, Elixhauser A. Identifying increased risk of readmission and in-hospital mortality using hospital administrative data: the AHRQ Elixhauser comorbidity index. *Med Care* 2017; **55**: 698–705.
 14. Quan H, Sundararajan V, Halfon P, Fong A, Burnand B, Luthi JC, Saunders LD, Beck CA, Feasby TE, Ghali WA. Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data. *Med Care* 2005; **43**: 1130–1139.
 15. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing; 2019. <https://www.R-project.org> (14 January 2021).
 16. Berthelot S, Lang ES, Quan H, Stelfox HT. Development of a hospital standardized mortality ratio for emergency department care. *Ann Emerg Med* 2016; **67**: 517–524 e26.
 17. Berthelot S, Lang ES, Quan H, Stelfox HT. Canadian in-hospital mortality for patients with emergency-sensitive conditions: a retrospective cohort study. *BMC Emerg Med* 2019; **19**: 57.
 18. Dormann CF, Keil P. Calibration of probability predictions from machine-learning and statistical models. *Glob Ecol Biogeogr* 2020; **29**: 760–765.
 19. Johnston A, Fink D, Reynolds MD, Hochachka WM, Sullivan BL, Bruns NE, Hallstein E, Merrifield MS, Matsumoto S, Kelling S. Abundance models improve spatial and temporal prioritization of conservation resources. *Ecol Appl* 2015; **25**: 1749–1756.
 20. Van Calster B, Nieboer D, Vergouwe Y, De Cock B, Pencina MJ, Steyerberg EW. A calibration hierarchy for risk models was defined: from utopia to empirical data. *J Clin Epidemiol* 2016; **74**: 167–176.
 21. Angraal S, Mortazavi BJ, Gupta A, Khera R, Ahmad T, Desai NR, Jacoby DL, Masoudi FA, Spertus JA, Krumholz HM. Machine learning prediction of mortality and hospitalization in heart failure with preserved ejection fraction. *JACC Heart Fail* 2020; **8**: 12–21.
 22. Canepa M, Fonseca C, Chioncel O, Laroche C, Crespo-Leiro MG, Coats AJS, Mebazaa A, Piepoli MF, Tavazzi L, Maggioni AP, Crespo-Leiro M, Anker S, Mebazaa A, Coats A, Filippatos G, Ferrari R, Maggioni AP, Piepoli MF, Amir O, Chioncel O, Dahlström U, Delgado Jimenez JF, Drodz J, Erglis A, Fazlibegovic E, Fonseca C, Fruhwald F, Gatzov P, Goncalvesova E, Hassanein M, Hradec J, Kavoliuniene A, Lainscak M, Logeart D, Merkely B, Metra M, Otljanska M, Seferovic PM, Srbinovska Kostovska E, Temizhan A, Tousoulis D, Ferreira T, Andarala M, Fiorucci E, Folkesson Lefrancq E, Glénot M, Gracia G, Konte M, Laroche C, McNeill PA, Missiamenou V, Taylor C, Auer J, Ablasser K, Fruhwald F, Dolze T, Brandner K, Gstrein S, Poelzl G, Moertl D, Reiter S, Podcezek-Schweighofer A, Muslibegovic A, Vasilj M, Fazlibegovic E, Cesko M, Zelenika D, Palic B, Pravidic D, Cuk D, Vitlianova K, Katova T, Velikov T, Kurteva T, Gatzov P, Kamenova D, Antova M, Sirakova V, Krejci J, Mikolaskova M, Spinar J, Krupicka J, Malek F, Hegarova M, Lazarova M, Monhart Z, Hassanein M, Sobhy M, el Messiry F, el Shazly AH, Elrakshy Y, Youssef A, Moneim AA, Noamany M, Reda A, Abdel Dayem TK, Farag N, Ibrahim Halawa S, Abdel Hamid M, Said K, Saleh A, Beid H, Hanna R, Aziz R, Louis O, Enen MA, Ibrahim BS, Nasr G, Elbahry A, Sobhy H, Ashmawy M, Gouda M, Aboleineen W, Bernard Y, Luporsi P, Meneveau N, Pillot M, Morel M, Seronde MF, Schiele F, Briand F, Delahaye F, Damy T, Eicher JC, de Groote P, Fertin M, Lamblin N, Isnard R, Lefol C, Thevenin S, Hagege A, Jondeau G, Logeart D, le Marcis V, Ly JF, Coisne D, Lequeux B, le Moal V, Mascle S, Lotton P, Behar N, Donal E, Thebault C, Ridard C, Reynaud A, Basquin A, Bauer F, Codjia R, Galinier M, Tourikis P, Stavroula M, Tousoulis D, Stefanadis C, Chrysohoou C, Kotrogiannis I, Matzaraki V, Dimitroula T, Karavidas A, Tsitsinakis G, Kapelios C, Nanas J, Kampouri H, Nana E, Kaldara E, Eugenidou A, Vardas P, Saloustros I, Patrianakos A, Tsaknakis T, Evangelou S, Nikoloulis N, Tziourganou H, Tsaroucha A, Papadopoulou A, Douras A, Polgar L, Merkely B, Kosztin A, Nyolczas N, Csaba Nagy A, Halmosi R, Elber J, Alony I, Shotan A, Vazan Fuhrmann A, Amir O, Romano S, Marcon S, Penco M, di Mauro M, Lemme E, Carubelli V, Rovetta R, Metra M, Bulgari M, Quinzani F, Lombardi C, Bosi S, Schiavina G, Squeri A, Barbieri A, di Tano G, Pirelli S, Ferrari R, Fucili A, Passero T, Musio S, di Biase M, Correale M, Salvemini G, Brognoli S, Zanelli E, Giordano A, Agostoni P, Italiano G, Salvioni E, Copelli S, Modena MG, Reggiani L, Valenti C, Oлару A, Bandino S, Deidda M, Mercurio G, Cadeddu Dessalvi C, Marino PN, di Ruocco MV, Sartori C, Piccinino C, Parrinello G, Licata G, Torres D, Giambanco S, Busalacchi S, Arrotti S, Novo S, Inciardi RM, Pieri P, Chirco PR, Ausilia Galifi M, Teresi G, Buccheri D, Minacapelli A, Veniani M, Frisinghelli A, Priori SG, Cattaneo S, Opasich C, Gualco A, Pagliaro M, Mancone M, Fedele F, Cinque A, Vellini M, Scarfo I, Romeo F, Ferraiuolo F, Sergi D, Anselmi M, Melandri F, Leci E, Iori E, Bovo V, Pidello S, Frea S, Bergerone S, Botta M, Canavosio FG, Gaita F, Merlo M, Cinquetti M, Sinagra G, Ramani F, Fabris E, Stolfo D, Artico J, Miani D, Fresco C, Daneluzzi C, Proclemer A, Ciccoira M, Zanolli L, Marchese G, Torelli F, Vassanelli C, Voronina N, Erglis A, Tamakauskas V, Smalinskas V, Karaliute R, Petraskiene I, Kazakauskaite E, Rumbinaite E, Kavoliuniene A, Vysniauskas V, Brazyte-Ramaniuskene R, Petraskiene D, Stankala S, Switala P, Juszczak Z, Sinkiewicz W, Gilewski W, Pietrzak J, Orzel T, Kaszelowicz P, Kardaszewicz P, Lazorko-Piega M, Gabryel J, Mosakowska K, Bellwon J, Rynkiewicz A, Raczak G, Lewicka E, Dabrowska-Kugacka A, Bartkowiak R, Sosnowska-Pasiarska B, Wozakowska-Kaplon B, Krzeminski A, Zabojszcz M, Mirek-Bryniarska E, Grzegorzko A, Bury K, Nessler J, Zalewski J, Furman A, Broncel M, Poliwczak A, Bala A, Zycinski P, Rudzinska M, Jankowski L, Kasprzak JD, Michalak L, Wojtczak Soska K,

- Drozd J, Huziuk I, Retwinski A, Flis P, Weglarz J, Bodys A, Grajek S, Kaluzna-Oleksy M, Straburzynska-Migaj E, Dankowski R, Szymanowska K, Grabia J, Szyszka A, Nowicka A, Samcik M, Wolniewicz L, Baczynska K, Komorowska K, Poprawa I, Komorowska E, Sajnaga D, Zolbach A, Dudzik-Plocica A, Abdulkarim AF, Lauko-Rachocka A, Kaminski L, Kostka A, Cichy A, Ruszkowski P, Splawski M, Fitas G, Szymczyk A, Serwicka A, Fiega A, Zysko D, Krysiak W, Szabowski S, Skorek E, Pruszczyk P, Bienias P, Ciurzynski M, Welnicki M, Mamcarz A, Folga A, Zielinski T, Rywik T, Leszek P, Sobieszczanska-Malek M, Piotrowska M, Kozar-Kaminska K, Komuda K, Wisniewska J, Tarnowska A, Balsam P, Marchel M, Opolski G, Kaplon-Cieslicka A, Gil RJ, Mozenska O, Byczkowska K, Gil K, Pawlak A, Michalek A, Krzesinski P, Piotrowicz K, Uzieblo-Zyczkowska B, Stanczyk A, Skrobowski A, Ponikowski P, Jankowska E, Rozentryt P, Polonski L, Gadula-Gacek E, Nowalany-Kozielska E, Kuczaj A, Kalarus Z, Szulik M, Przybylska K, Klys J, Prokop-Lewicka G, Kleinrok A, Tavares Aguiar C, Ventosa A, Pereira S, Faria R, Chin J, de Jesus I, Santos R, Silva P, Moreno N, Queirós C, Lourenço C, Pereira A, Castro A, Andrade A, Oliveira Guimaraes T, Martins S, Placido R, Lima G, Brito D, Francisco AR, Cardiga R, Proenca M, Araujo I, Marques F, Fonseca C, Moura B, Leite S, Campelo M, Silva-Cardoso J, Rodrigues J, Rangel I, Martins E, Sofia Correia A, Peres M, Marta L, Ferreira da Silva G, Severino D, Duraõ D, Leao S, Magalhaes P, Moreira I, Filipa Cordeiro A, Ferreira C, Araujo C, Ferreira A, Baptista A, Radoi M, Bicescu G, Vinereanu D, Sinescu CJ, Macarie C, Popescu R, Daha I, Dan GA, Stanescu C, Dan A, Craiu E, Nechita E, Aursulesei V, Christodorescu R, Otasevic P, Seferovic PM, Simeunovic D, Ristic AD, Celic V, Pavlovic-Kleut M, Suzic Latic J, Stojcevski B, Pencic B, Stevanovic A, Andric A, Iric-Cupic V, Jovic M, Davidovic G, Milanov S, Mitic V, Atanaskovic V, Antic S, Pavlovic M, Stanojevic D, Stoickov V, Ilic S, Deljanin Ilic M, Petrovic D, Stojsic S, Kecojevic S, Dodic S, Cemic Adic N, Cankovic M, Stojiljkovic J, Mihajlovic B, Radin A, Radovanovic S, Krotin M, Klambnik A, Goncalvesova E, Pernicky M, Murin J, Kovar F, Kmec J, Semjanova H, Strasek M, Savnik Iskra M, Ravnikar T, Cernic Suligoj N, Komel J, Fras Z, Jug B, Glavic T, Losic R, Bombek M, Krajnc I, Krunic B, Horvat S, Kovac D, Rajtman D, Cencic V, Letonja M, Winkler R, Valentincic M, Melihen-Bartolic C, Bartolic A, Pusnik Vrckovnik M, Kladnik M, Slemenik Pusnik C, Marolt A, Klen J, Drnovsek B, Leskobar B, Fernandez Anguita MJ, Gallego Page JC, Salmeron Martinez FM, Andres J, Genis AB, Mirabet S, Mendez A, Garcia-Cosio L, Roig E, Leon V, Gonzalez-Costello J, Muntane G, Garay A, Alcade-Martinez V, Lopez Fernandez S, Rivera-Lopez R, Puga-Martinez M, Fernandez-Alvarez M, Serrano-Martinez JL, Crespo-Leiro M, Grille-Cancela Z, Marzoa-Rivas R, Blanco-Canosa P, Paniagua-Martin MJ, Barge-Caballero E, Laynez Cerdena I, Famara Hernandez Baldomero I, Lara Padron A, Ofelia Rosillo S, Dalmau Gonzalez-Gallarza R, Salvador Montanes O, Iniesta Manjavacas AM, Castro Conde A, Araujo A, Soria T, Garcia-Pavia P, Gomez-Bueno M, Cobo-Marcos M, Alonso-Pulpon L, Segovia Cubero J, Sayago I, Gonzalez-Segovia A, Briceno A, Escribano Subias P, Vicente Hernandez M, Ruiz Cano MJ, Gomez Sanchez MA, Delgado Jimenez JF, Barrios Garrido-Lestache E, Garcia Pinilla JM, Garcia de la Villa B, Sahuquillo A, Bravo Marques R, Torres Calvo F, Perez-Martinez MT, Gracia-Rodenas MR, Garrido-Bravo IP, Pastor-Perez F, Pascual-Figal DA, Diaz Molina B, Orus J, Epelde Gonzalo F, Bertomeu V, Valero R, Martinez-Abellan R, Quiles J, Rodriguez-Ortega A, Mateo I, ElAmrani A, Fernandez-Vivancos C, Bierge Valero D, Almenar-Bonet L, Sanchez-Lazaro IJ, Marques-Sule E, Facila-Rubio L, Perez-Silvestre J, Garcia-Gonzalez P, Ridocci-Soriano F, Garcia-Escriba D, Pellicer-Cabo A, de la Fuente Galan L, Lopez Diaz J, Recio Platero A, Arias JC, Blasco-Peiro T, Sanz Julve M, Sanchez-Insua E, Aured-Guallar C, Portoles-Ocampo A, Melin M, Hägglund E, Stenberg A, Lindahl IM, Asserlund B, Olsson L, Dahlström U, Afzelius M, Karlström P, Tengvall L, Wiklund PA, Olsson B, Kalayci S, Temizhan A, Cavusoglu Y, Gencer E, Yilmaz MB, Gunes H. Performance of prognostic risk scores in chronic heart failure patients enrolled in the European Society of Cardiology Heart Failure Long-Term Registry. *JACC Heart Fail* 2018; **6**: 452–462.
23. Hwang IC, Cho GY, Choi HM, Yoon YE, Park JJ, Park JB, Lee SP, Kim HK, Kim YJ, Sohn DW. Derivation and validation of a mortality risk prediction model using global longitudinal strain in patients with acute heart failure. *Eur Heart J Cardiovasc Imaging* 2020; **21**: 1412–1420.
24. Lim NK, Lee SE, Lee HY, Cho HJ, Choe WS, Kim H, Choi JO, Jeon ES, Kim MS, Kim JJ, Hwang KK, Chae SC, Baek SH, Kang SM, Choi DJ, Yoo BS, Kim KH, Cho MC, Oh BH, Park HY. Risk prediction for 30-day heart failure-specific readmission or death after discharge: data from the Korean Acute Heart Failure (KorAHF) registry. *J Cardiol* 2019; **73**: 108–113.
25. Yap J, Lim FY, Chia SY, Allen JC Jr, Jaufferally FR, Macdonald MR, Chai P, Loh SY, Lim P, Zaw MWW, Teo L, Sim D, Lam CSP. Prediction of survival in Asian patients hospitalized with heart failure: validation of the OPTIMIZE-HF risk score. *J Card Fail* 2019; **25**: 571–575.
26. Fonarow GC, Adams KF Jr, Abraham WT, Yancy CW, Boscardin WJ, ADHERE Scientific Advisory Committee SG. Risk stratification for in-hospital mortality in acutely decompensated heart failure: classification and regression tree analysis. *JAMA* 2005; **293**: 572–580.
27. Peterson NM, Rumsfeld JS, Liang L, Albert NN, Hernandez AF, Peterson ED, Fonarow GC, Masoudi FA, American Heart Association Get With the Guidelines-Heart Failure Program. A validated risk score for in-hospital mortality in patients with heart failure from the American Heart Association get with the guidelines program. *Circ Cardiovasc Qual Outcomes* 2010; **3**: 25–32.
28. Desai RJ, Wang SV, Vaduganathan M, Evers T, Schneeweiss S. Comparison of machine learning methods with traditional models for use of administrative claims with electronic medical records to predict heart failure outcomes. *JAMA Netw Open* 2020; **3**: e1918962.
29. Lagu T, Pekow PS, Stefan MS, Shieh MS, Pack QR, Kashef MA, Atreya AR, Valania G, Slawsky MT, Lindenauer PK. Derivation and validation of an in-hospital mortality prediction model suitable for profiling hospital performance in heart failure. *J Am Heart Assoc* 2018; **7**.
30. Loring Z, Mehrotra S, Piccini JP, Camm J, Carlson D, Fonarow GC, Fox KAA, Peterson ED, Pieper K, Kakkav AK. Machine learning does not improve upon traditional regression in predicting outcomes in atrial fibrillation: an analysis of the ORBIT-AF and GARFIELD-AF registries. *Europace* 2020; **22**: 1635–1644.
31. Chen R, Stewart WF, Sun J, Ng K, Yan X. Recurrent neural networks for early detection of heart failure from longitudinal electronic health record data: implications for temporal modeling with respect to time before diagnosis, data density, data quantity, and data type. *Circ Cardiovasc Qual Outcomes* 2019; **12**: e005114.
32. Ambale-Venkatesh B, Yang X, Wu CO, Liu K, Hundley WG, McClelland R, Gomes AS, Folsom AR, Shea S, Guallar E, Bluemke DA, Lima JAC. Cardiovascular event prediction by machine learning: the Multi-Ethnic Study of Atherosclerosis. *Circ Res* 2017; **121**: 1092–1101.
33. Zack CJ, Senecal C, Kinar Y, Metzger Y, Bar-Sinai Y, Widmer RJ, Lennon R, Singh M, Bell MR, Lerman A, Gulati R. Leveraging machine learning techniques to forecast patient prognosis after percutaneous coronary intervention. *JACC Cardiovasc Interv* 2019; **12**: 1304–1311.

34. Nanayakkara S, Fogarty S, Tremeer M, Ross K, Richards B, Bergmeir C, Xu S, Stub D, Smith K, Tacey M, Liew D, Pilcher D, Kaye DM. Characterising risk of in-hospital mortality following cardiac arrest using machine learning: a retrospective international registry study. *PLoS Med* 2018; **15**: e1002709.
35. Essay P, Balkan B, Subbian V. Decom-pensation in critical care: early prediction of acute heart failure onset. *JMIR Med Inform* 2020; **8**: e19892.
36. Rangachari P. Coding for quality measurement: the relationship between hospital structural characteristics and coding accuracy from the perspective of quality measurement. *Perspect Health Inf Manag* 2007; **4**: 3.
37. König S, Ueberham L, Schuler E, Wiedemann M, Reithmann C, Seyfarth M, Sause A, Tebbenjohanns J, Schade A, Shin DI, Staudt A. In-hospital mortality of patients with atrial arrhythmias: insights from the German-wide Helios hospital network of 161 502 patients and 34 025 arrhythmia-related procedures. *Eur Heart J* 2018; **39**: 3947–3957.