ESC European Society of Cardiology

European Heart Journal - Digital Health (2023) **4**, 175–187
https://doi.org/10.1093/ehjdh/ztac054

ORIGINAL ARTICLE

What drives performance in machine learning models for predicting heart failure outcome?

Rom Gutman[1], Doron Aronson[2,3], Oren Caspi [2,3,*†], and Uri Shalit[1,†]

[1]William Davidson Faculty of Industrial Engineering and Management, Technion, Haifa, Israel; [2]Department of Cardiology, Rambam Health Care Campus; and [3]the Bruce Rappaport Faculty of Medicine, Technion, Haifa, Israel

Received 27 May 2022; revised 19 August 2022; online publish-ahead-of-print 30 September 2022

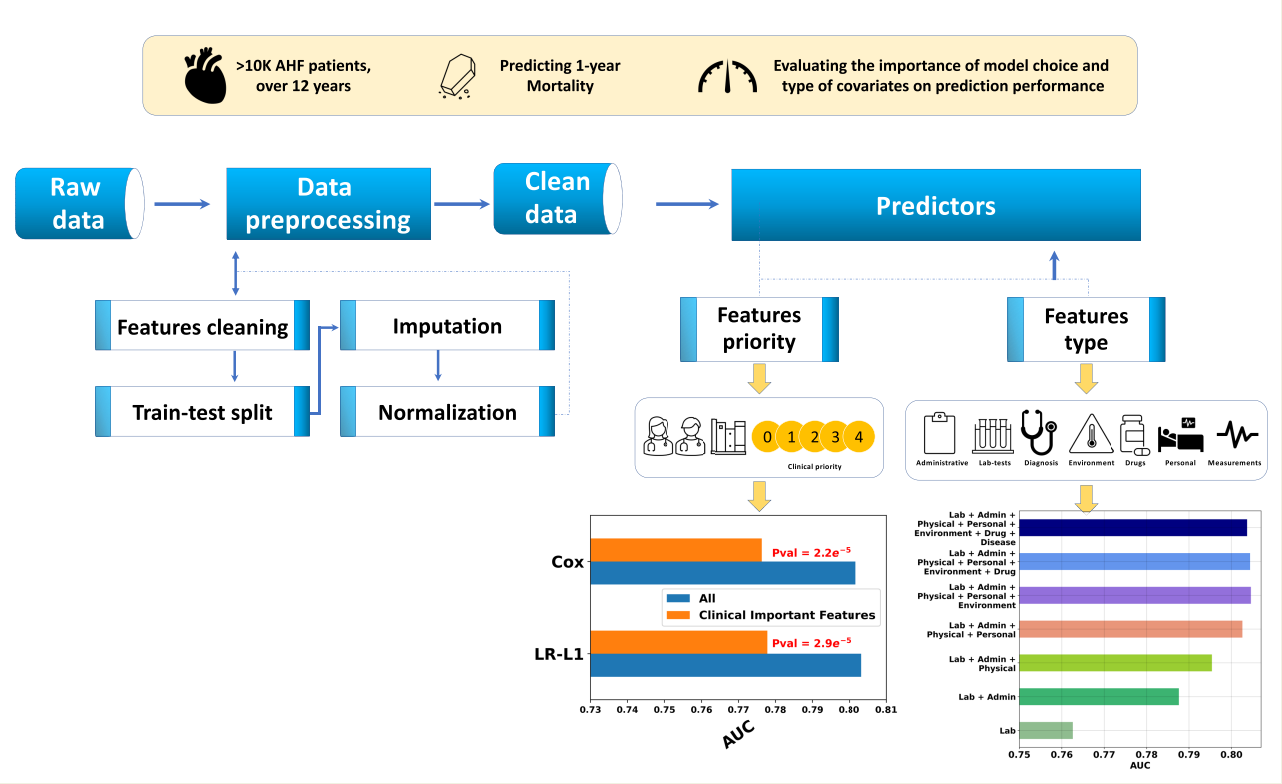
Aims The development of acute heart failure (AHF) is a critical decision point in the natural history of the disease and carries a dismal prognosis. The lack of appropriate risk-stratification tools at hospital discharge of AHF patients significantly limits clinical ability to precisely tailor patient-specific therapeutic regimen at this pivotal juncture. Machine learning-based strategies may improve risk stratification by incorporating analysis of high-dimensional patient data with multiple covariates and novel prediction methodologies. In the current study, we aimed at evaluating the drivers for success in prediction models and establishing an institute-tailored artificial Intelligence-based prediction model for real-time decision support.

Methods and results We used a cohort of all 10 868 patients AHF patients admitted to a tertiary hospital during a 12 years period. A total of 372 covariates were collected from admission to the end of the hospitalization. We assessed model performance across two axes: (i) type of prediction method and (ii) type and number of covariates. The primary outcome was 1-year survival from hospital discharge. For the model-type axis, we experimented with seven different methods: logistic regression (LR) with either L_1 or L_2 regularization, random forest (RF), Cox proportional hazards model (Cox), extreme gradient boosting (XGBoost), a deep neural-net (NeuralNet) and an ensemble classifier of all the above methods. We were able to achieve an area under receiver operator curve (AUROC) prediction accuracy of more than 80% with most prediction models including L1/L2-LR (80.4%/80.3%), Cox (80.2%), XGBoost (80.5%), NeuralNet (80.4%). RF was inferior to other methods (78.8%), and the ensemble model was slightly superior (81.2%). The number of covariates was a significant modifier ($P < 0.001$) of prediction success, the use of multiplex-covariates preformed significantly better (AUROC 80.4% for L1-LR) compared with a set of known clinical covariates (AUROC 77.8%). Demographics followed by lab-tests and administrative data resulted in the largest gain in model performance.

Conclusions The choice of the predictive modelling method is secondary to the multiplicity and type of covariates for predicting AHF prognosis. The application of a structured data pre-processing combined with the use of multiple-covariates results in an accurate, institute-tailored risk prediction in AHF

* Corresponding author. Tel: 972-4-777-3781, Fax: 972-4-777-2176, Email: o_caspi@rmc.gov.il
†The last two authors contributed equally to the study.

Graphical Abstract



Introduction

Despite the immense progress in cardiology in the last decades, heart failure is a growing pandemic worldwide. It is a leading cause of morbidity and mortality and a colossal economic burden on healthcare systems.[1–3] Acute heart failure (AHF) is the most common cause for hospitalization for adults (>65 years) in western societies.[4] The prognosis of patients admitted with AHF is dismal, with 30% re-admission rate and 5 to 15% mortality rate at 60 to 90 days post-discharge.[5,6] Nevertheless, AHF is a heterogenous syndrome with a variable prognosis and identifying patients for early events may have a broad impact on health care delivery.[7] While in recent years, the armamentarium for mitigating high risk heart failure has been expanded, the lack of appropriate risk-stratification tools for AHF patients limits physician ability for tailoring the appropriate therapeutic and follow-up regimen. This limitation may contribute to the unacceptably high re-hospitalization rate and short-term mortality associated with AHF.[6,8]

Advances in statistical machine learning (ML) method have led to an explosion of personalized risk scores based on a large number of patient-specific covariates. Many of these models were developed using electronic medical records (EMRs) of large patient populations. When assessing the value of these risk scores, several factors are often conflated: the use of large high-dimensional patient data derived from EMR, and the use of novel statistical methods for prediction. Such methods are often identified with the field of ML and include random-forest,[9] gradient-boosted decision trees,[10] sparse-regression methods such as least absolute shrinkage and selection operator[11] and deep-neural nets.[12]

Prediction based on hospitalization data is of major importance for supporting decisions related to optimizing patient discharge and examining the appropriate therapeutic strategy (e.g. eligibility for advanced heart failure device treatment and heart transplantation) and the surveillance schedule. In the current study, we used a single-centre cohort of AHF patients aiming at developing a clinically useful model for the prediction of mortality and HF re-hospitalization after hospital discharge. Toward this aim we evaluated the relative contribution of two axes of prediction drivers: (i) the type and number of covariates; (ii) The type of the prediction model including classical statistical models and ML-based methods.

Methods

Study population

We used a database of all patients admitted to the Rambam Medical Center, Haifa, Israel, with the primary diagnosis of AHF between September 2004 and December 2015. Eligible patients were those

hospitalized with new-onset or worsening of preexisting heart failure as the primary cause of admission,[13–15] using the European Society of Cardiology criteria.[16] The study was performed in accordance with the Declaration of Helsinki and approved by the institutional review committee on human research. Like most of the observational data in the healthcare domain, the patients' data that used in our study were assembled from EMR. The basic characteristics of our cohort are detailed in *Table 1*.

Study outcomes

Study outcomes were predefined prior to data analysis. The primary outcome was 1 year mortality from hospital discharge. The secondary outcome was a composite of one year mortality or rehospitalization for AHF at 1 year. Analysis for the primary outcome is included in the main manuscript. Analysis for the secondary outcome is included in the Supplementary material online, supplementary data.

Data collection and preparation

Using observational data (rather than prospective clinical research) is cheap, relatively easy to obtain and often provides a good representation of the real-world population of interest. However, it poses two major challenges that need to be addressed prior to the prediction efforts: wrong (noisy) data, and missing data. Below we describe how we handle these challenges. Overall, we used 372 covariates collected from admission to the end of the hospitalization including demographics, administrative data, lab tests, medical therapies, and echocardiographic data. A full list is available in the Supplementary material online, *Table S8*. A schematic representation of our data preparation pipeline is given in *Figure 1*.

Data imputation and optimization

We searched for errors in the data by analyzing the distribution of each covariate in order to map outliers (e.g. mixture of [cm] and [m] in height data, swap between systolic blood pressure and diastolic blood pressure values), identify ill-calculated features (e.g. body mass index), and spot data points that are out of domain (e.g. text values in lab result feature). Then, each outlier was manually checked and fixed by the appropriate action, to ensure the data integrity. Covariates that were found to be too noisy or irrelevant were removed.

We addressed missingness by a two-stage process: indication and imputation. For each covariate with at least one missing value, we created a dummy variable to indicate which data-point was missing in the original data. Then, we imputed missing data[17] using median imputation (see Supplementary material online, *Table S2*). We also experimented with multivariate imputation by chained equations (MICEs) imputation,[18–20] and found that the former achieved slightly better results (see Supplementary material online, *Table S1*). Therefore, we chose to use median imputation for all our analyses. Finally, the data were normalized using z-scoring: $\hat{x} = \frac{x - mean(X)}{std(X)}$. Some covariates (e.g. lab dates) were redefined to represent the number of days from admission, to ensure common domain.

Statistical analysis

We define the outcome y as a binary indicator of whether the patient died during a span of 1-year from hospital discharge. Given a set of patient covariates measured before and during hospitalization, denoted x, we estimated $p(y|x)$, the probability of the outcome for a patient with the given set of covariates. The evaluation (test) set is temporally distinct from the set used to fit the model: 2004–2013 as train (model fitting) set, 2014–2015 as evaluation set.

Our aim was to assess model performance across two axes—(i) type of prediction method and (ii) type and number of covariates. For the model-type axis, we experimented with seven different methods for estimating $p(y|x)$: logistic regression (LR) with either L_1 or L_2 regularization, random forest (RF), Cox proportional hazards model (Cox), extreme gradient boosting (XgBoost),[10] a deep neural-net (NeuralNet), and an ensemble classifier (average of the LR- L_1, Cox proportional hazards model, XgBoost, and NeuralNet predictions). The LR, RF, and majority vote classifiers were implemented using the Sci-kit learn python package.[21] The Cox model required regularization because of the large number of covariates. Since Cox model predicts time-to-event, and the other methods simply predict a probability of the event happening, we needed to slightly modify the way we use the Cox predictor for a direct comparison with the other models. See details below.

Transforming time-to-event prediction to binary prediction

We consider the difference between time-to-event analysis as performed by Cox regression and the rest of the models we use. Most predictors provide a probability that a sample is in a particular class (e.g. in the binary case, probability for mortality for a patient in the 1− year span from discharge), survival analysis does this and more. A time-to-event model provides the cumulative probability of a sample to have an event before (or at) a time t, where in our case t is the time from the baseline, to the event, measured in days.

Therefore, for an apples-to-apples comparison, we use the cumulative probability $F(t) = P(T \leq t) = 1 - S(t)$, where $S(t)$ is the survival function, as given by the Cox analysis, and t is set to 365 days. This is exactly the probability for an event occurring during one year from the time of hospital discharge.

Neural network

We designed the neural network using a fully connected architecture. Each layer built from linear weights, a non-linear activation function (ReLU), and applying dropout during model fitting.[22] We also added an L_1 regularization term on the weights of the first layer.

Covariate type and priority

In order to assess the relative contribution of choice of covariates, we conducted two analyses: using the *type* of the features and the *clinical-priority* of the features. From now on, we will refer to them as 'Type' models and 'Priority' models, respectively.

For the 'Type' analysis, we partitioned the covariates into seven types: Laboratory results ('*lab tests*'), demographic ('*personal*'), data related to administrative processes in the hospital such as relative time from admission to first blood test ('*admin*'), patient comorbidities ('*disease*'), drug related data, both prior to hospitalization, during hospitalization, and at discharge ('*drug*'), physical measurements such as patient's blood pressure ('*physical*'), and environmental related features such as the season of admission ('*environment*'). We ran each prediction method on every combination of the seven feature types outlined above, for a total of $2^7 - 1 = 127$ different sets of covariates per prediction method.

For the 'Priority' analysis, we used physicians' expertise (heart failure cardiologists) to select for the most validated and established covariates both in terms of prediction ability and clinical prior knowledge.[23–26] Each covariate received a clinical-relevance score from 0 to 4, where 0 is 'not relevant' and 4 is 'highly relevant'. Examples for priority 4 covariates are the last read of patient diastolic and systolic blood pressure, and examples for priority 0 covariates are whether the patient weight was

Table 1 Backgroud clinical characteristics of patients at development and validation cohorts

Variable	Value	Development	Validation	Missing
No. of patients		8520	2348	
Demographics				
Age, (years)		74.4 ± 12.1	74.3 ± 12.4	0
Gender (female)		49.4% (4209)	48.1% (1129)	
BMI, Kg/m^2		29.9 ± 6.4	29.8 ± 6.4	8832
Comorbidities				
Chronic renal failure (*n*)		33.4% (2843)	28.2% (662)	0
Valvular heart disease (*n*)		13.8% (1176)	12.9% (303)	0
AKI, % (*n*)		18.39% (1567)	21.97% (516)	
Diabetes				
COPD		52.6% (4483)	54.6% (1282)	0
Ischaemic heart disease		13.5% (1148)	15.1% (354)	0
Atrial fibrillation		22.3% (1899)	18.9% (443)	0
		41.7% (3556)	41.8% (981)	0
Smoking	not smoking	74.2% (6323)	70.4% (1654)	
	past smoker	14.8% (1257)	14.5% (341)	
	active smoker	11.0% (940)	15.0% (353)	
Clinical parameters				
Diastolic blood pressure, mmHg	Mean	68.6 ± 9.6	69.1 ± 9.6	2281
Systolic blood pressure, mmHg	Mean	131.2 ± 22.0	131.3 ± 19.5	2281
MAP, mmHg	Mean	89.5 ± 11.6	89.9 ± 11.3	2281
body temperature, celsius	First	37.1 ± 10.3	36.7 ± 0.4	2309
Estimated LVEF	Mean	44.3 ± 47.3	44.2 ± 19.5	6000
Lab tests				
Albumin, g/dL		3.3 ± 0.5	3.2 ± 0.5	3672
BNP, ng/mL	First	1313.8 ± 1124.3	1112.3 ± 1011.4	5941
	Last	1306.1 ± 1118.6	1112.3 ± 1012.6	5941
BUN, mg/dL	Mean	36.0 ± 20.2	36.5 ± 21.0	2472
Creatinine, mg/dL	Mean	1.7 ± 1.1	1.6 ± 1.1	2473
RDW (%)	First	15.7 ± 1.9	16.1 ± 2.3	3496
Sodium, mmol/L	Mean	138.3 ± 3.5	138.1 ± 3.7	2469
Haemoglobin, g/dL	First	11.6 ± 1.9	11.5 ± 2.0	2504
	Last	11.5 ± 1.8	11.4 ± 1.9	2498
Troponin, ng/mL	First	0.2 (1.5)	0.2 (1.1)	4088
HCT (%)	First	35.3 (5.7)	35.1 (5.9)	2502
Medical therapy				
β-blockers (*n*)	Background	64.8% (5518)	78.7% (1849)	0
	In hospital	61.7% (5255)	88.4% (2075)	0
	Discharge	65.2% (5554)	83.3% (1955)	0
ACE inhibitors (*n*)	Background	60% (5114)	66.3% (1557)	0
Angiotensin receptor blocker, % (*n*)	In hospital	41.2% (3509)	49.6% (1165)	0
In hospital	15.5% (1322)	22.8% (536)	0	
In hospital	17.9% (1522)	27.3% (642)	0	
				MRA, % (*n*)
Furosemide, % (*n*)	Background	59.19% (5043)	68.44% (1607)	0
Target				
1-year mortality, % (*n*)		30% (2552)	31.3% (736)	0

AKI, acute kidney injury; BMI, body mass index; BNP, brain natriuretic peptide; HCT-Hematocrit, MAP, mean arterial pressure, MRA, mineralocorticoid receptor antagonist; RDW, red cell distribution width.

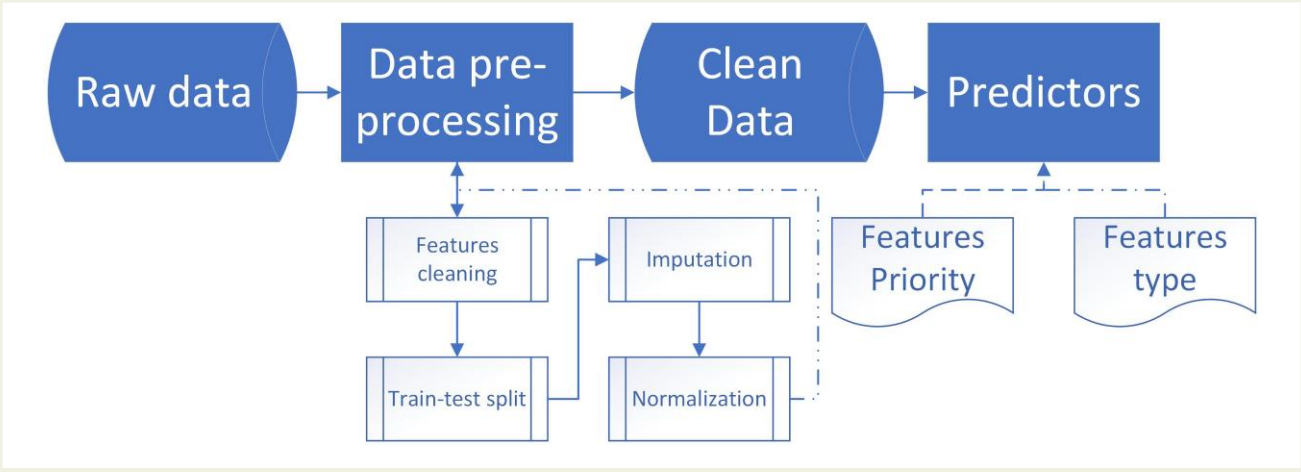


Figure 1 The suggested pipeline of the predicting process.

recorded in admission. The full list of variables their subtype and the clinical priority is detailed in the Supplementary material online, *Table S8*. We use this clinical prior knowledge input to train our predictors as follows: The first model used only the most important clinical features (4), the next one used these together with the second most important ones, so we denote it 3+, and so on, where the last model used all the features, which we denote 0+. Overall, for the Priority analysis, we evaluated five sets of covariates per prediction method.

Thus, in total we considered 127 + 5 = 132 sets of covariates, and seven prediction methods, giving a total of 132 × 7–7 = 917 models (we subtract 7 because the models fit on 0+ priority are identical to those fit using all Types of features). Each model's hyper-parameters (e.g. level of regularization) was optimized for Area under receiver operator curve (AUROC) using 10-fold stratified cross-validation on the training set. The AUROC of different models was compared using the DeLong method for comparing areas under the ROC curve.[27,28] We emphasize that our goal is not to find the single most accurate model out of the 917—rather we wish to understand which factors account for better or worse model accuracy across the settings.

In addition to AUROC, we also give a decision curve analysis (net-benefit)[29] and report the following metrics, as suggest by Shameer *et al.*[30]: sensitivity, specificity, negative predictive value (NPV), positive predictive value (PPV), accuracy, and Brier Score. See Supplementary material online, *Table S9* for the full details.

Code availability

Full code is available in https://github.com/RomGutman/ADHF.

Results

A cohort of patients with a primary diagnosis of AHF admitted to a tertiary hospital during a period of 12 years was used in order to assess study outcomes. The cohort included 10 868 cases from which a total of 372 covariates were collected from admission to discharge. The training cohort consisted of 8520 cases and the validation cohort included 2348 cases.

The performance of the seven prediction models was evaluated across seven predetermined sets of covariates (demographic, physical examination, patient's comorbidities, laboratory tests, drug therapy, environmental, and administrative) classified according to their clinical level of priority (indicated as low priority-0 to high priority-4). Heart failure clinicians identified 52 covariates as high priority level. The results on prediction model performances for predicting the primary outcome (1-year mortality) are presented over a total of 917 settings assessed on a temporally held-out dataset.

The models applied predicted one year mortality with a relatively high level of accuracy. Overall, the performance of the different methods was found to be comparable (*Figure 2*), with the exception of the RF model that significantly underperformed all other models (*Table 2*). Specifically, L1- and L2-regularized LR (80.4% and 80.3% AUROC, respectively), XgBoost (80.5%) and NeuralNet (80.4%) predicted the occurrence of one year mortality with essentially the same AUROC when using all covariates, whereas the RF (78.8%) performed slightly worse than other models and the ensemble model (81.2%) slightly better. The probable reason for RF to preform less than other models here may be related to overfitting of the training data. Importantly, Cox based model, which is not based on a ML paradigm, preformed similarly (80.2%).

To further characterize the difference in the predictive capacity of the models a comparative analysis between pairs of models was conducted using the DeLong method for comparing areas under the ROC (AUROC) curve[27,28] (*Table 2*). A *P*-value for a one-sided test, i.e. whether model A has higher AUROC than model B, of all the pairs of methods when using the full set of covariates. The analysis demonstrate similar performance between the models despite the fact that the models we compare make vastly different assumptions about the outcome (linear vs. non-linear, sparse or not), and have a widely varying number of parameters (neural networks have 1.4 million parameters vs. 372 parameters for the LR models). To evaluate the contribution of the number of covariates and their clinical importance we evaluated the differential performance of each model according to the assigned priority of the clinical variables (*Figure 2*). A comparative analysis demonstrated that inclusion of low priority variables improved the predictive performance of the

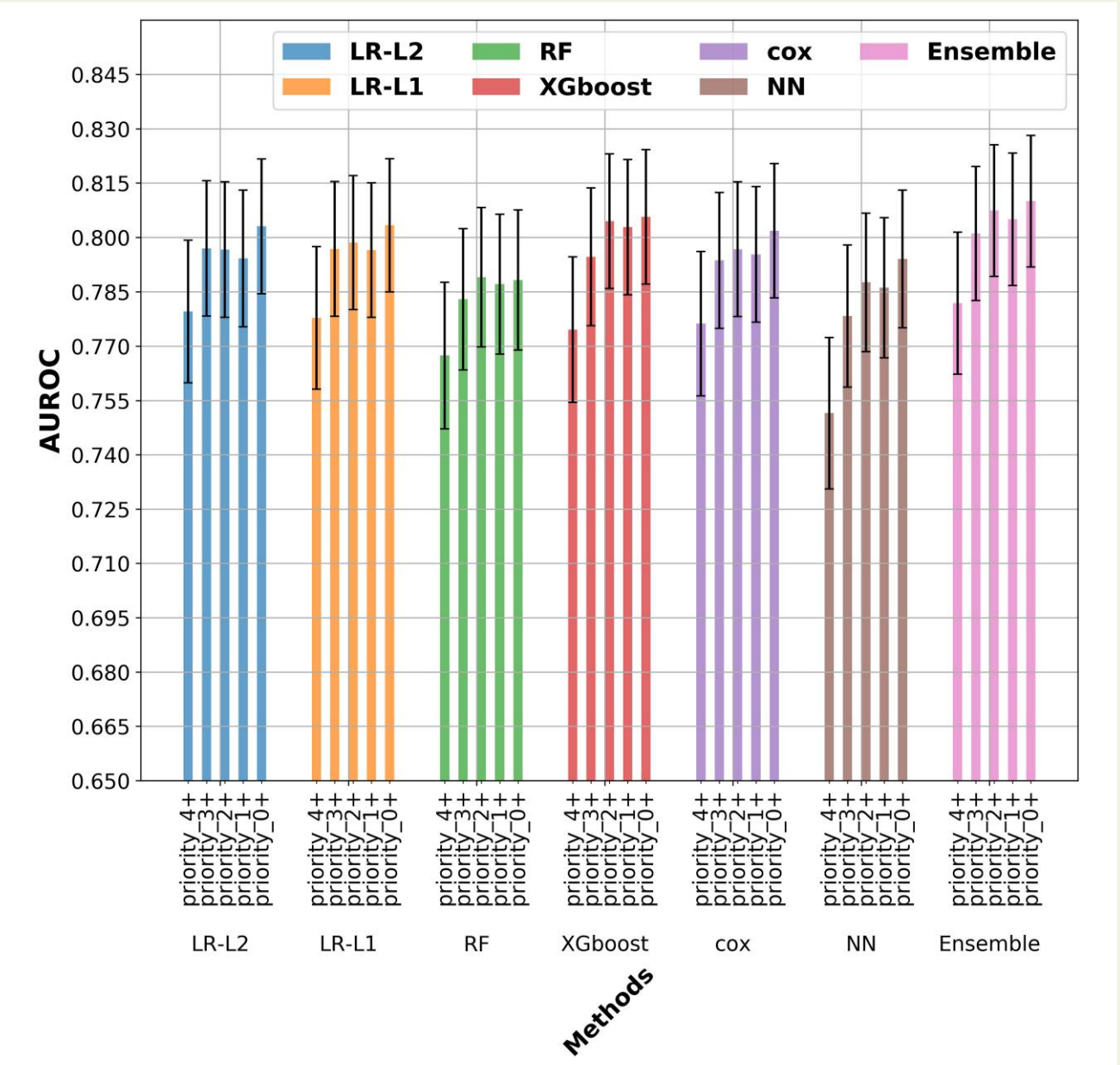


Figure 2 Prediction of the models based on the set of clinical covariates utilized. Area under receiver operator curve of the models used splitted by clinical priority of the variables. The confidence interval in this graph are the prediction interval for the area under receiver operator curve prediction. LR, logistic regression; NN, neuron network; RF, random forest; XgBoost, extreme gradient boosting trees; Cox, Cox regression; Majority Ensemble, ensemble model using LR-L1, Cox, XgBoost and NN.

models even when comparing analyses based on priority 0+ compared to priority 1+ (*Table 3*).

Calibration plots found a good calibration for most models with slight underestimation of the predicted risk for low risk patients using RF (*Figure 3*). To assess the clinical implication of the model in predicting mortality outcome for AHF patients, we carried out a net-benefit analysis (*Figure 4*).[31] A reasonable threshold for 1 year mortality risk in heart failure patients that entails consideration of advanced therapies (e.g. Heart transplantation or mechanical circulatory support) is 20%.[32–34] All models give a significant net benefit and perform similarly across a wide range of decision thresholds. We further see that most models are indistinguishable in their net benefit, with RF again the only model with diminished performance compared to the other models. In addition, the sensitivity, specificity, NPV, PPV, and Brier score, for all models using the full set of covariates is detailed in *Table 4*. There are differences between the models in their sensitivity and PPV, but these are mostly the result of slightly different decision thresholds of the models, which disappear when using analyses that consider all possible thresholds such as AUROC and net benefit.

To estimate the relative cumulative contribution of different types of covariate, we analyzed their relative contribution to model

Table 2 Difference in predictive capacity for predicting the primary outcome between pairs of methods

	LR-L1	LR-L2	RF	NN	XGboost	Cox	Ensemble
LR-L1	**80.4%**	0.624	0.991	0.432	0.396	0.72	0.0002
LR-L2	0.376	**80.3%**	0.988	0.346	0.33	0.66	6.82E-05
RF	0.0085	0.012	**78.8%**	0.004	0.0003	0.007	9.6E-07
NeuralNet	0.568	0.654	0.996	**80.4%**	0.467	0.75	0.005
XgBoost	0.604	0.673	1.0	0.533	**80.5%**	0.75	0.006
Cox	0.28	0.34	0.993	0.249	0.253	**80.2%**	1.11E-05
Ensemble	1.0	1.0	1.0	0.995	0.994	1.0	**81.2%**

The table diagonal contains cells with the AUROC value for each prediction model (bold). The off-diagonal cells represent the *P*-value of a one-sided paired test, testing whether the ROC curve of the row model is smaller than the model in the column. Cells highlighted in red indicate that the result is significant for $P < 0.05$. All models were generated the full set of covariates (i.e. 300 + covariates).
LR, logistic regression; NeuralNet, neural network; RF, random forest; XgBboost, extreme gradient boosting trees; Cox, Cox regression; Ensemble, ensemble model using LR-L1, Cox, XgBoost and NeuralNet.

performance. For this purpose, we conducted the analysis for the L1 LR model (*Figure 5*). Specifically, we ordered the types by the biggest gain each type gave over the previously added types, with the provision that age and sex are always included. Using only sex and age gives AUROC of no more than 61.6% (ranging from 56.3 to 61.6% across the other models). The addition of lab-test gives by far the biggest gain in performance, followed by adding administrative data (examples of administrative data is whether the patient had a previous AHF hospitalization in the previous years as indicated in the database, or whether a weight of the patient was recorded). We further see diminishing returns when adding two other types of information: personal (demographic data) and physical examination. On top of these, adding diseases (comorbidities), environmental and drug information gives essentially only modest gain in model performance. In *Table 5*, we summarize the most statistically significant covariates ($P < 0.05$) obtained when using L1 LR as the predictive model, applied to the full set of covariates.

Discussion

The present study is aimed at understanding what drives success in mortality prediction based on a contemporary AHF hospitalization cohort utilizing a variety of ML methodologies. The study's main finding is that the choice of the predictive modelling method is secondary to the choice and variety of covariates: most methods, including both classic methods such as Cox regression and newer methods such as gradient boosted trees, perform similarly when given the same set of covariates. The study also demonstrates that adding a large number of covariates, some of which are deemed less clinically relevant by clinical experts, results in a significant gain in accuracy, largely irrespective of the statistical method chosen. Finally, an additional finding of the study is that integrating non-clinical, administrative variables to demographic and laboratory variables significantly improves the predictive capabilities of the model.

In the current study, building prediction models with all available information resulted in a reliable prediction of 1-year mortality and rehospitalization using data gained solely from hospital EMR. Specifically, standard *regularized* linear methods such as LR and Cox regression were shown to be non-inferior to ML based methodologies. An important finding of the current study is that the use of multiple covariates is crucial for augmenting the prediction accuracy. On the priority front, we found that using 113 covariates (out of 372) that were set as high priority by clinical experts were inferior to using a multiple-covariates strategy (including variables deemed as non-significant, priority 0). These results were gained on future split of the data, which shows that data from the past can be used for accurate prediction.

Prediction of heart failure rehospitalization and mortality has been a long-standing challenge. Given the severe consequences associated with the disease many efforts were and are currently invested at improving the prediction the AHF prognosis, with limited success. Currently, there are two main prediction tasks that have been studied. First, for AHF patients, short-term prediction of rehospitalization, and in-hospital or 30 day mortality, mainly aiming to assess the quality of heart failure treatment during the hospitalization.[35] Second, for ambulatory chronic heart failure patient predicting 1-year and 3-year outcomes.[23,35–39] The clinical setting chosen for the present study, at patient discharge from acute hospitalization, is critical for choosing the appropriate therapeutic strategy (e.g. advanced therapies or Uptitration of heart failure medications) and for tailoring the surveillance programme and the medical therapy. While several models have been devised for risk stratifying heart failure patients at admission to hospital or ambulatory patients[24,40–42] none thus far have been developed for AHF patients at this critical decision point. The current study elucidates the relevant factors affecting AHF prognosis and the relative contribution of the factor-type to the risk stratification using an institution tailored methodology.

The main predictive variables in the study (listed in *Table 5*) include demographic, laboratory, drug therapy, and administrative factors. Age, albumin levels, and RDW have been described as predictive factors for mortality in general and heart failure mortality specifically.[43–48] Our study is in agreement with previous studies showing that discharge (or last) BNP is a stronger predictor of AHF prognosis than baseline BNP and provide a substantial tool for risk stratification.[49] In recent years, there is growing evidence that admission hypochloraemia is a marker for mortality and rehospitalization in chronic heart failure patients,[50] and the present study reiterates this finding and

Table 3 Difference in prediction capacity according to the model used and the assigned clinical priority of the variables used

	LR-L1 priority 4+	LR-L1 priority 3+	LR-L1 priority 2+	LR-L1 priority 1+	LR-L1 priority 0+	Cox priority 4+	Cox priority 3+	Cox priority 2+	Cox priority 1+	Cox priority 0+
LR-L1 priority 4+	**77.8%**	0.001	0.001	0.002	2.9E-05	0.690	0.004	0.001	0.003	9E-05
LR-L1 priority 3+	0.999	**79.7%**	0.242	0.547	0.012	0.999	0.856	0.505	0.656	0.096
LR-L1 priority 2+	0.999	0.758	**79.9%**	0.990	0.003	1.000	0.906	0.711	0.838	0.177
LR-L1 priority 1+	0.998	0.453	0.010	**79.7%**	1E-05	0.999	0.771	0.469	0.642	0.065
LR-L1 priority 0+	1.000	0.988	0.997	1.000	**80.3%**	1.000	0.993	0.965	0.988	0.675
Cox priority 4+	0.310	0.001	0.000	0.001	2.4E-05	**77.6%**	0.001	0.0003	0.001	2.2E-05
Cox priority 3+	0.996	0.144	0.094	0.229	0.007	0.999	**79.4%**	0.102	0.261	0.003
Cox priority 2+	0.999	0.495	0.289	0.531	0.035	1.000	0.898	**79.7%**	0.939	0.003
Cox priority 1+	0.997	0.344	0.162	0.358	0.012	0.999	0.739	0.061	**79.5%**	5.3E-05
Cox priority 0+	1.000	0.904	0.823	0.935	0.325	1.000	0.997	0.997	1.000	**80.2%**

The table diagonal contains cells with the AUROC value for each prediction model and variables priority set (bold). The off-diagonal cells represent the *P*-value of a one-sided paired test testing whether the ROC curve of the row model is smaller than the model in the column. Cells highlighted in red indicate that the result is significant for $P < 0.05$. The 'priority x+' postfix for each model indicate that it uses the covariates that were assigned to have priority $x \leq$ (e.g. cox priority 3 + indicates that the cox model was trained with covariates that have priority 4 and 3).
LR, logistic regression; NeuralNet, neural network; RF, random forest; XgBboost, extreme gradient boosting trees; Cox, Cox regression; Ensemble, ensemble model using LR-L1, Cox, XgBoost and NeuralNet.

demonstrates that low plasma chloride levels at admission (more than low sodium levels) serve as a significant prognostic factor in AHF for predicting 1 year mortality. A key finding of the study is that administrative variables significantly boost the predictive performance of the evaluated models. Among those variables, while a history of recent AHF hospitalization is widely described as a prognostic factor,[51,52] factors such as length of stay and missing weight measurement are less established as risk markers and may be used to alert clinicians for the associated risks, serving as proxies for underlying patient characteristics.

The application of ML methodologies for predicting adverse events and outcomes based on electronic health records (EHRs) is a promising approach, highly relevant to acute exacerbation of chronic diseases such as heart failure.[42,53–55] The benefit of ML in predicting prognosis in the setting of heart failure is not yet established and its added advantage over traditional models is still controversial. In the study of Frizzell *et al.*,[56] use of a number of ML algorithms did not improve prediction of 30-day all-cause readmissions 30 days after discharge from a heart failure hospitalization compared with traditional LR models. Desai *et al.*[57] recently demonstrated that multiple ML methodologies have minimal benefit over traditional LR in the setting of chronic heart failure. However, the authors noted that ML approaches generally fared better when the models included numerous continuous variables such as laboratory tests.[57]

A recent systematic review compared the results of LR and other ML methods, across different risk-scores using clinical data, found no evidence of superior performance of ML techniques in terms of AUROC.[58] More recent reviews have found and discussed similar finding regarding general benefit for using Deep-learning techniques.[59–61] Our findings generally agree, as we were unable to demonstrate any systematic benefit of ML methods in terms of AUROC, calibration or net benefit. Importantly, it remains to be determined whether ML can improve model calibration, which is considered as the 'Achilles heel' of predictive analytics, and specifically in heart failure (underestimating the risk for high-risk patients and overestimating the risk for low-risk patients).[62–64]

The present study adds an important observation: using a very large and diverse set of covariates, many more than are used in classic risk scores, carries a significant benefit in terms of model discrimination regardless of the method used, without adversely affecting model calibration. This lesson is likely to be applicable not only for AHF outcome prediction but also for other chronic diseases undergoing acute decompensations with complex EHR data, such as inflammatory bowel disease and chronic obstructive pulmonary disease.

Unlike much of the above work, we propose not a single model but a framework for understanding what drives performance when building site-specific models. When building predictive models, there is usually a tradeoff between robustness and accuracy[65,66] In this work, we focus on the accuracy end of this tradeoff, where we are willing to use many hospital-specific and population-specific variables in order to boost accuracy *at the site*, with the possible price of building models which will not generalize well to other sites or populations. Thus, the limitation of using a large number of covariates is the adaptation and stability of the prediction model when applied to other EHRs in alternate institutions. For example, the

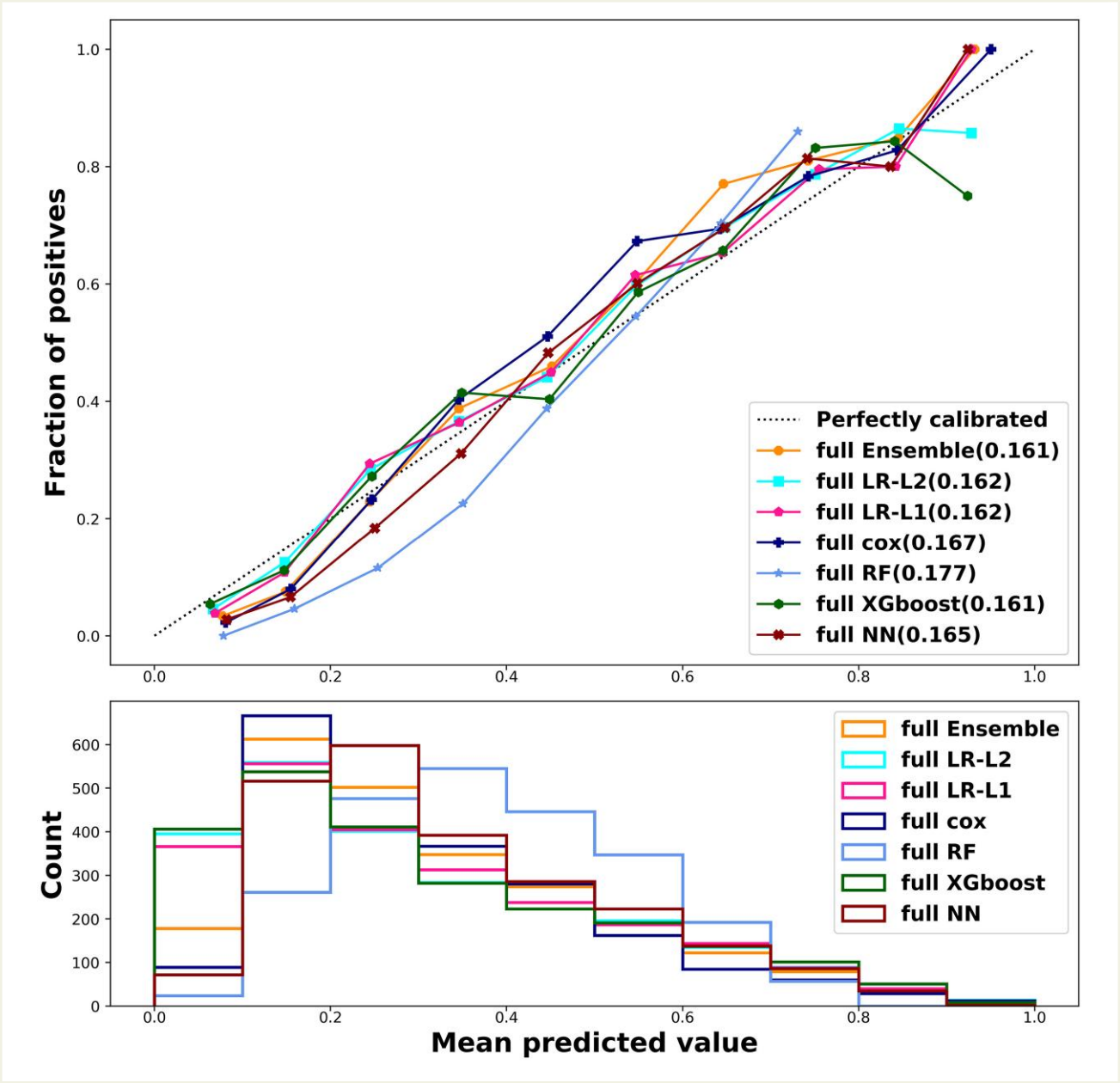


Figure 3 Calibration plot. All models applied were reasonably calibrated (upper panel). Risk distribution of the patients evaluated using the prediction models delineates that most patients had a medium risk for the unified outcome—higher than that assessed for the 1 year mortality outcome (lower panel). The number in parentheses after each model indicates the brier score of each model. The 'full' prefix for each model is to indicate that it uses all the 300 + covariates. LR, logistic regression; NN, neuron network; RF, random forest; XGboost, extreme gradient boosting trees; Cox, Cox regression; Majority Ensemble, ensemble model using LR-L1, Cox, XGboost and NN.

typical time until a certain lab test is administered will very likely vary between hospitals. This means the models our framework yielded when applied to a specific population cannot and should not be used 'as is' in different environments. Rather, the structured framework for constructing models such as ours, each tailored to the patient population and practices of a specific hospital and a specific clinical scenario, should be used. Towards that end, we are sharing the code base used to construct and evaluate our models (https://github.com/RomGutman/ADHF). We also note that such predictors should be re-calibrated at some-future-point, as we cannot expect a model to be accurate for indefinitely long periods of time.

In conclusion, our study demonstrates that by applying ML strategies for optimal use of data, combined with the use of multiple covariates, we were able to reliably predict the outcome of a chronic disease such as heart failure based solely on hospitalization data during acute decompensation. However, we were unable to demonstrate a clear superiority of ML methods over the traditional Cox regression model (when used with a regularization factor). We believe that a similar multiple covariate

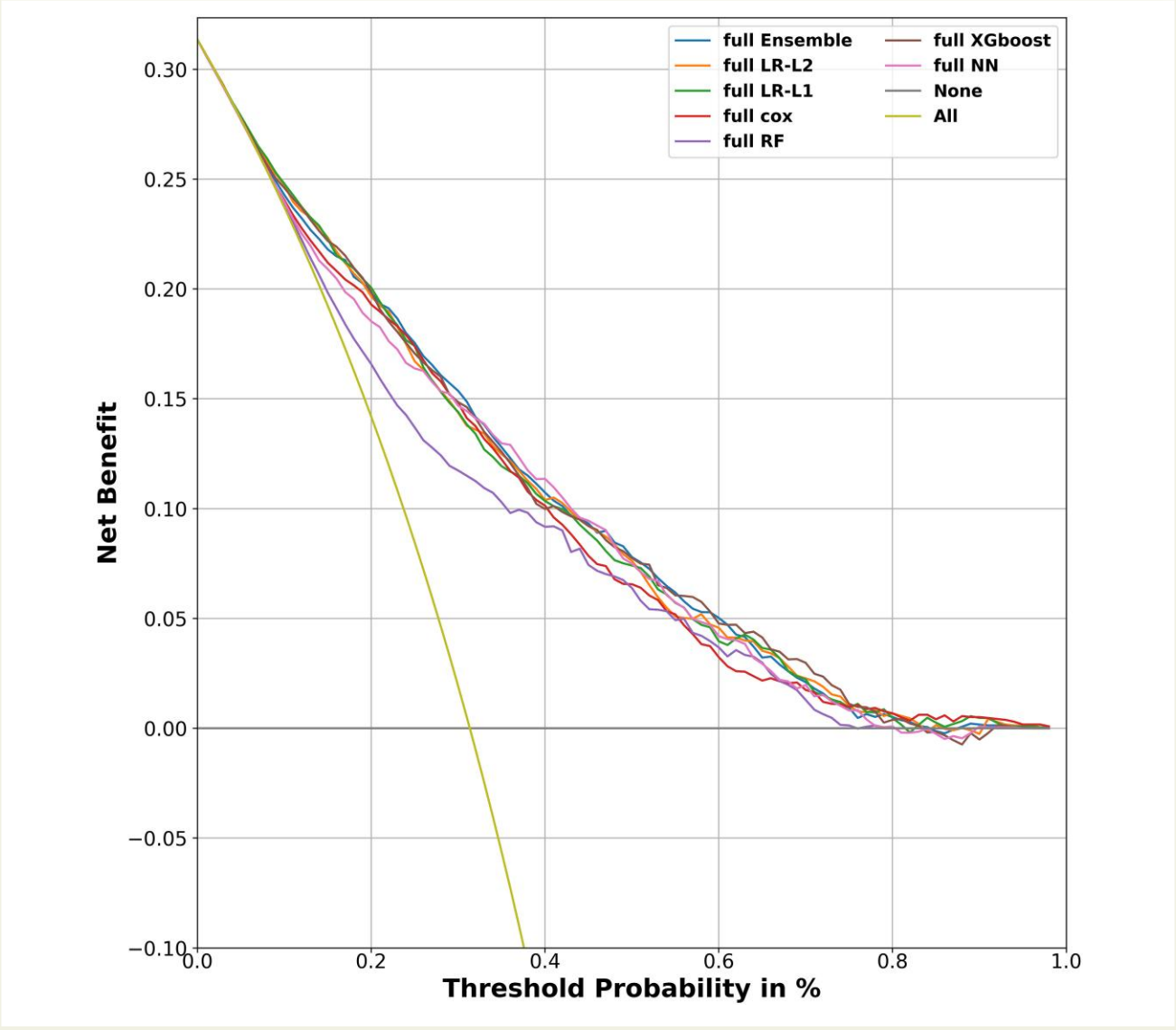


Figure 4 Net benefit of the prediction models. Net benefit allows to evaluate the clinical implication of the model in predicting mortality outcome. A threshold of 20% risk for mortality is considered significant for altering decisions regarding treatment strategies and patient's categorization (Stage C vs. Stage D heart failure). All models give a significant net benefit with random forest being the least beneficial. The straight yellow line denotes the benefit of a model that predicts mortality for all patients. LR, logistic regression; NN, neuron network; RF, random forest; XGboost, extreme gradient boosting trees; Cox, Cox regression; Ensemble, ensemble model using LR-L1, Cox, xgboost and NN.

Table 4 The metrics of the prediction models using all the features (variables) for predicting the primary outcome

	AUROC	Sensitivity	Specificity	PPV	NPV	Accuracy	Brier
Cox	80.18%	0.34	0.94	0.72	0.76	0.75	0.167
LR-L1	80.37%	0.44	0.91	0.69	0.78	0.76	0.162
LR-L2	80.29%	0.44	0.91	0.69	0.78	0.76	0.162
Ensemble	81.23%	0.42	0.92	0.71	0.78	0.76	0.161
NeuralNet	80.45%	0.45	0.90	0.68	0.78	0.76	0.165
RF	78.84%	0.51	0.86	0.63	0.79	0.75	0.177
XgBoost	80.49%	0.46	0.91	0.69	0.78	0.76	0.161

A threshold of 0.5 is used.
AUROC, area under receiver operator curve; NPV, negative predictive value; PPV, positive predictive value.

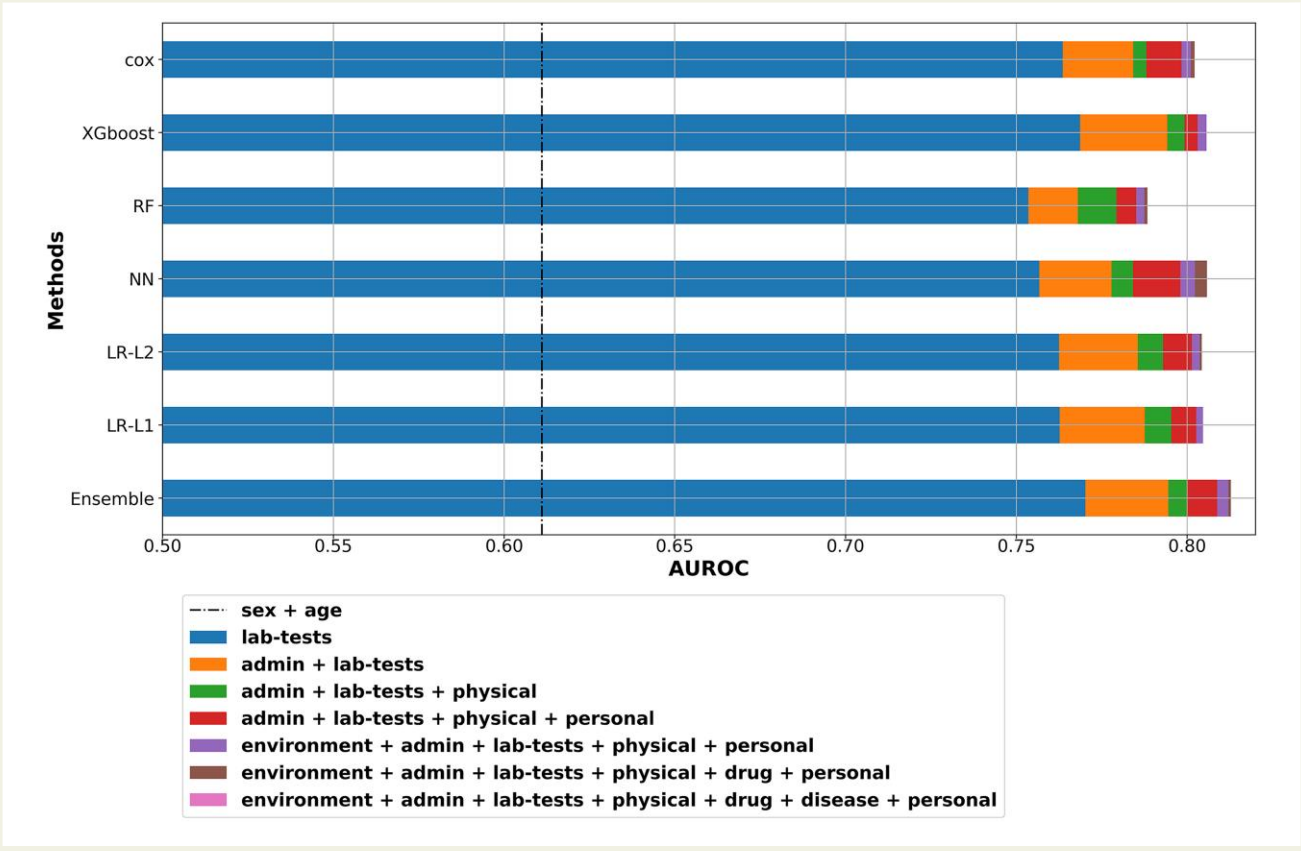


Figure 5 Area under receiver operator curve plot by types. The relative effect of the various variable types on the performance of the prediction model is delineated. Variable types that are not shown in the graph had a minor (or negative) contribution to the model's performance. Sex and age contribution are represented by the dashed vertical line. Admin, administrative variables; lab-tests, laboratory tests; physical, physical examination; personal, demographic variables; drug, drug treatment; disease, comorbidities; LR, logistic regression; NN, neuron network; RF, random forest; XGboost, extreme gradient boosting trees; Cox, Cox regression; Majority Ensemble, ensemble model using LR-L1, cox, XGboost and NN.

Table 5 Most significant variables predicting one year survival

Variable	Odds ratio	Std.Err	Z-value	P>|z|	[0.025	0.975]
Age	1.52	0.03	13.23	6.13E-40	1.43	1.61
Patient readmission	1.31	0.028	9.46	2.99E-21	1.24	1.38
Albumin	0.77	0.03	−8.76	1.92E-18	0.73	0.82
RDW(%)[a]	1.12	0.03	6.24	4.42E-10	1.13	1.26
Length of stay	1.24	0.04	5.49	3.95E-08	1.15	1.34
Last BNP	1.14	0.03	4.84	1.31E-06	1.08	1.2
Hyperlipidaemia	0.78	0.06	−4.33	1.49E-05	0.70	0.88
Statins at discharge	0.76	0.06	−4.14	3.42E-05	0.67	0.87
Missing weight measurement	1.32	0.07	4.05	5.05E-05	1.15	1.51
Sex (female)	0.80	0.06	−3.81	0.00014	0.72	0.9
Chloride[a]	0.90	0.03	−3.05	0.002	0.84	0.96
Furosemide (background)	1.21	0.07	2.85	0.004	1.06	1.38

BNP, brain natriuretic peptide; RDW, red blood cell distribution width.
[a]First exams taken during hospitalization.

strategy may be used as a disease and institute specific methodology for constructing decision support tools for risk stratification at discharge for patients with chronic diseases such as heart failure.

Supplementary material

Supplementary material is available at *European Heart Journal – Digital Health* online.

Funding

Funding support for this research was given by Yad Hanadiv Foundation, Israel Science Foundation, and the Council of Higher Education (Israel) for student funding.

Conflict of interest: None declared.

Data availability

All data are included in the submission/manuscript file.

References

1. Lund LH, Rich MW, Hauptman PJ. Complexities of the global heart failure epidemic. *J Card Fail* 2018;**24**:813–814.
2. Maggioni AP. Epidemiology of heart failure in Europe. *Heart Fail Clin* 2015;**11**:625–635.
3. Thavendiranathan P, Nolan MT. An emerging epidemic: cancer and heart failure. *Clin Sci (Lond)* 2017;**131**:113–121.
4. Ambrosy AP, Fonarow GC, Butler J, Chioncel O, Greene SJ, Vaduganathan M, Nodari S, Lam CSP, Sato N, Shah AN, Gheorghiade M. The global health and economic burden of hospitalizations for heart failure: lessons learned from hospitalized heart failure registries. *J Am Coll Cardiol* 2014;**63**:1123–1133.
5. Gheorghiade M, Vaduganathan M, Fonarow GC, Bonow RO. Rehospitalization for heart failure: problems and perspectives. *J Am Coll Cardiol* 2013;**61**:391–403.
6. Blecker S, Herrin J, Li L, Yu H, Grady JN, Horwitz LI. Trends in hospital readmission of medicare-covered patients with heart failure. *J Am Coll Cardiol* 2019;**73**:1004–1012.
7. Crespo-Leiro MG, Metra M, Lund LH, Milicic D, Costanzo MR, Filippatos G, Gustafsson F, Tsui S, Barge-Caballero E, De Jonge N, Frigerio M, Hamdan R, Hasin T, Hulsmann M, Nalbantgil S, Potena L, Bauersachs J, Gkouziouta A, Ruhparwar A, Ristic AD, Straburzynska-Migaj E, McDonagh T, Seferovic P, Ruschitzka F. Advanced heart failure: a position statement of the heart failure association of the European society of cardiology. *Eur J Heart Fail* 2018;**20**:1505–1535.
8. O'Connor CM. High heart failure readmission rates: is it the health system's fault? *JACC Heart Fail* 2017;**5**:393.
9. Breiman L. Random forests. *Mach Learn* 2001;**45**:5–32.
10. Chen T, Guestrin C. XGBoost. p.785–794.
11. Tibshirani R. Regression shrinkage and selection via the LASSO. *J R Stat Soc Ser B (Methodological)* 1996;**58**:267–288.
12. Lecun Y, Bengio Y, editors. *Hinton G. Deep learning*. In: Nature Publishing Group; 2015. p436–444.
13. Caspi O, Naami R, Halfin E, Aronson D. Adverse dose-dependent effects of morphine therapy in acute heart failure. *Int J Cardiol* 2019.
14. Aronson D, Darawsha W, Atamna A, Kaplan M, Makhoul BF, Mutlak D, Lessick J, Carasso S, Reisner S, Agmon Y, Dragu R, Azzam ZS. Pulmonary hypertension, right ventricular function, and clinical outcome in acute decompensated heart failure. *J Card Fail* 2013;**19**:665–671.
15. Mutlak D, Lessick J, Khalil S, Yalonetsky S, Agmon Y, Aronson D. Tricuspid regurgitation in acute heart failure: is there any incremental risk? *Eur Heart J Cardiovasc Imaging* 2018;**19**:993–1001.
16. Ponikowski P, Voors AA, Anker SD, Bueno H, Cleland JGF, Coats AJS, Falk V, González-Juanatey JR, Harjola V-P, Jankowska EA, Jessup M, Linde C, Nihoyannopoulos P, Parissis JT, Pieske B, Riley JP, Rosano GMC, Ruilope LM, Ruschitzka F, Rutten FH, van der Meer P. 2016 ESC guidelines for the diagnosis and treatment of acute and chronic heart failure. *Eur J Heart Fail* 2016;**18**:891–975.
17. Steyerberg EW, van Veen M. Imputation is beneficial for handling missing data in predictive models. *J Clin Epidemiol* 2007;**60**:979.
18. Azur MJ, Stuart EA, Frangakis C, Leaf PJ. Multiple imputation by chained equations: what is it and how does it work? *Int J Methods Psychiatr Res* 2011;**20**:40–49.
19. Raghunathan TE, Lepkowski JM, Van Hoewyk J, Solenberger P, van Hoewyk J. A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models Key Words: Item nonresponse; Missing at random; Multiple imputation; Nonignorable missing mechanism; Regression; Sampling properties and simulations. In; 2001.
20. Rubin DB. Multiple imputation for nonresponse in surveys; 2004.
21. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E. Scikit-learn: machine learning in {P}ython. *J Mach Learn Res* 2011;**12**:2825–2830.
22. Srivastava N, Hinton G, Krizhevsky A, Salakhutdinov R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. In; 2014, 1929–1958.
23. Pocock SJ, Ariti CA, McMurray JJV, Maggioni A, Køber L, Squire IB, Swedberg K, Dobson J, Poppe KK, Whalley GA, Doughty RN. Predicting survival in heart failure: a risk score based on 39 372 patients from 30 studies. *Eur Heart J* 2013;**34**:1404–1413.
24. Wussler D, Kozhuharov N, Sabti Z, Walter J, Strebel I, Scholl L, Miro O, Rossello X, Martin-Sanchez FJ, Pocock SJ, Nowak A, Badertscher P, Twerenbold R, Wildi K, Puelacher C, du Fay de Lavallaz J, Shrestha S, Strauch O, Flores D, Nestelberger T, Boeddinghaus J, Schumacher C, Goudev A, Pfister O, Breidthardt T, Mueller C. External validation of the MEESSI acute heart failure risk score: a cohort study. *Ann Intern Med* 2019.
25. Fonarow GC. Risk stratification for in-hospital mortality in acutely decompensated heart failure< SUBTITLE> classification and regression tree analysis</SUBTITLE&gt. *JAMA* 2005;**293**:572.
26. Felker GM, Leimberger JD, Califf RM, Cuffe MS, Massie BM, Adams KF, Jr, Gheorghiade M, O'Connor CM. Risk stratification after hospitalization for decompensated heart failure. *J Card Fail* 2004;**10**:460–466.
27. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 1988;**44**:837–837.
28. Sun X, Xu W. Fast implementation of DeLong's algorithm for comparing the areas under correlated receiver operating characteristic curves. *IEEE Signal Process Lett* 2014;**21**:1389–1393.
29. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Making* 2006;**26**:565–574.
30. Shameer K, Johnson KW, Glicksberg BS, Dudley JT, Sengupta PP. Machine learning in cardiovascular medicine: are we there yet? *Heart (British Cardiac Society)* 2018;**104**:1156–1164.
31. Vickers AJ, Van Calster B, Steyerberg EW. Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. *BMJ* 2016:i6.
32. Samman-Tahhan A, Hedley JS, McCue AA, Bjork JB, Georgiopoulou VV, Morris AA, Butler J, Kalogeropoulos AP. INTERMACS profiles and outcomes among non-inotrope-dependent outpatients with heart failure and reduced ejection fraction. *JACC Heart Fail* 2018;**6**:743–753.
33. Starling RC, Estep JD, Horstmanshof DA, Milano CA, Stehlik J, Shah KB, Bruckner BA, Lee S, Long JW, Selzman CH, Kasirajan V, Haas DC, Boyle AJ, Chuang J, Farrar DJ, Rogers JG. Risk assessment and comparative effectiveness of left ventricular assist device and medical management in ambulatory heart failure patients: the ROADMAP study 2-year results. *JACC Heart Fail* 2017;**5**:518–527.
34. Estep JD, Starling RC, Horstmanshof DA, Milano CA, Selzman CH, Shah KB, Loebe M, Moazami N, Long JW, Stehlik J, Kasirajan V, Haas DC, O'Connell JB, Boyle AJ, Farrar DJ, Rogers JG. Risk assessment and comparative effectiveness of left ventricular assist device and medical management in ambulatory heart failure patients: results from the ROADMAP study. *J Am Coll Cardiol* 2015;**66**:1747–1761.
35. Lee DS, Stitt A, Austin PC, Stukel TA, Schull MJ, Chong A, Newton GE, Lee JS, Tu JV. Prediction of heart failure mortality in emergent care: a cohort study. *Ann Intern Med* 2012; **156**:767–775, w-261, w-262.
36. Lee DS, Austin PC, Rouleau JL, Liu PP, Naimark D, Tu JV. Predicting mortality among patients hospitalized for heart failure. *JAMA* 2003;**290**:2581–2581.
37. Levy WC, Mozaffarian D, Linker DT, Sutradhar SC, Anker SD, Cropp AB, Anand I, Maggioni A, Burton P, Sullivan MD, Pitt B, Poole-Wilson PA, Mann DL, Packer M. The Seattle heart failure model: prediction of survival in heart failure. *Circulation* 2006;**113**:1424–1433.
38. Nakada Y, Kawakami R, Matsushima S, Ide T, Kanaoka K, Ueda T, Ishihara S, Nishida T, Onoue K, Soeda T, Okayama S, Watanabe M, Okura H, Tsuchihashi-Makaya M, Tsutsui H, Saito Y. Simple risk score to predict survival in acute decompensated heart failure – A2B score –. *Circ J* 2019.
39. Scrutinio D, Ammirati E, Guida P, Passantino A, Raimondo R, Guida V, Braga SS, Pedretti RFE, Lagioia R, Frigerio M, Catanzaro R, Oliva F. Clinical utility of N-terminal pro-B-type natriuretic peptide for risk stratification of patients with acute decompensated heart failure. Derivation and validation of the ADHF/NT-proBNP risk score. *Int J Cardiol* 2013;**168**:2120–2126.
40. Passantino A, Monitillo F, Iacoviello M, Scrutinio D. Predicting mortality in patients with acute heart failure: role of risk scores. *World J Cardiol* 2015;**7**:902–911.

41. Miro O, Rossello X, Gil V, Martin-Sanchez FJ, Llorens P, Herrero-Puente P, Jacob J, Bueno H, Pocock SJ. Predicting 30-day mortality for patients with acute heart failure in the emergency department: a cohort study. *Ann Intern Med* 2017;**167**:698–705.
42. 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.
43. Cheng Y-L, Sung S-H, Cheng H-M, Hsu P-F, Guo C-Y, Yu W-C, Chen C-H. Prognostic nutritional Index and the risk of mortality in patients with acute heart failure. *J Am Heart Assoc* 2017;**6**:e004876.
44. Gatta A, Verardo A, Bolognesi M. Hypoalbuminemia. *Intern Emerg Med* 2012;**7**: S193–S199.
45. Peng W, Zhang C, Wang Z, Yang W. Prediction of all-cause mortality with hypoalbuminemia in patients with heart failure: a meta-analysis. *Biomarkers* 2019;**24**: 631–637.
46. Turcato G, Zorzi E, Prati D, Ricci G, Bonora A, Zannoni M, Maccagnani A, Salvagno GL, Sanchis-Gomar F, Cervellin G, Lippi G. Early in-hospital variation of red blood cell distribution width predicts mortality in patients with acute heart failure. *Int J Cardiol* 2017;**243**:306–310.
47. Dabbah S, Hammerman H, Markiewicz W, Aronson D. Relation between red cell distribution width and clinical outcomes after acute myocardial infarction. *Am J Cardiol* 2010;**105**:312–317.
48. Makhoul BF, Khourieh A, Kaplan M, Bahouth F, Aronson D, Azzam ZS. Relation between changes in red cell distribution width and clinical outcomes in acute decompensated heart failure. *Int J Cardiol* 2013;**167**:1412–1416.
49. Omar HR, Guglin M. Discharge BNP is a stronger predictor of 6-month mortality in acute heart failure compared with baseline BNP and admission-to-discharge percentage BNP reduction. *Int J Cardiol* 2016;**221**:1116–1122.
50. Grodin JL, Simon J, Hachamovitch R, Wu Y, Jackson G, Halkar M, Starling RC, Testani JM, Tang WHW. Prognostic role of serum chloride levels in acute decompensated heart failure. *J Am Coll Cardiol* 2015;**66**:659–666.
51. Bowen GS, Diop MS, Jiang L, Wu W-C, Rudolph JL. A multivariable prediction model for mortality in individuals admitted for heart failure. *J Am Geriatr Soc* 2018;**66**: 902–908.
52. Voors AA, Ouwerkerk W, Zannad F, van Veldhuisen DJ, Samani NJ, Ponikowski P, Ng LL, Metra M, Ter Maaten JM, Lang CC, Hillege HL, van der Harst P, Filippatos G, Dickstein K, Cleland JG, Anker SD, Zwinderman AH. Development and validation of multivariable models to predict mortality and hospitalization in patients with heart failure. *Eur J Heart Fail* 2017;**19**:627–634.
53. Tomašev N, Glorot X, Rae JW, Zielinski M, Askham H, Saraiva A, Mottram A, Meyer C, Ravuri S, Protsyuk I, Connell A, Hughes CO, Karthikesalingam A, Cornebise J, Montgomery H, Rees G, Laing C, Baker CR, Peterson K, Reeves R, Hassabis D, King D, Suleyman M, Back T, Nielson C, Ledsam JR, Mohamed S. A clinically applicable approach to continuous prediction of future acute kidney injury. *Nature* 2019; **572**:116–119.
54. Goldstein BA, Navar AM, Carter RE. Moving beyond regression techniques in cardiovascular risk prediction: applying machine learning to address analytic challenges. *Eur Heart J* 2017;**38**:1805–1814.
55. Goto T, Camargo CA Jr, Faridi MK, Yun BJ, Hasegawa K. Machine learning approaches for predicting disposition of asthma and COPD exacerbations in the ED. *Am J Emerg Med* 2018;**36**:1650–1654.
56. Frizzell JD, Liang L, Schulte PJ, Yancy CW, Heidenreich PA, Hernandez AF, Bhatt DL, Fonarow GC, Laskey WK. Prediction of 30-day all-cause readmissions in patients hospitalized for heart failure: comparison of machine learning and other statistical approaches. *JAMA Cardiol* 2017;**2**:204–209.
57. 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.
58. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *J Clin Epidemiol* 2019;**110**:12–22.
59. Rudin C. Interpretable machine learning: fundamental principles and 10 grand challenges. *Stat Surv* 2021;**16**:1–85.
60. Grinsztajn LE, Oyallon E, Gaël V. Tabular data: deep learning is not all you need. *arXiv* 2022.
61. Shwartz-Ziv R, Armon A. Tabular data: deep learning is not all you need. *Information Fusion* 2022;**81**:84–90.
62. Alba AC, Agoritsas T, Jankowski M, Courvoisier D, Walter SD, Guyatt GH, Ross HJ. Risk prediction models for mortality in ambulatory patients with heart failure. *Circ Heart Fail* 2013;**6**:881–889.
63. Shah ND, Steyerberg EW, Kent DM. Big data and predictive analytics: recalibrating expectations. *JAMA* 2018;**320**:27–28.
64. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW. Topic Group 'Evaluating diagnostic t, prediction models' of the Si. Calibration: the achilles heel of predictive analytics. *BMC Med* 2019;**17**:230.
65. Subbaswamy A, Chen B, Saria S. A unifying causal framework for analyzing dataset shift-stable learning algorithms. *J Causal Inference* 2022;**10**:64–89.
66. Hongyang Z, Yaodong Y, Jiantao J. Theoretically principled trade-off between robustness and accuracy. *PMLR* 2019.