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

Rom Gutman MSc, Doron Aronson MD, Oren Caspi MD PhD, Uri Shalit PhD.

From the William Davidson Faculty of Industrial Engineering and Management, Technion, Haifa, Israel, and From the Department of Cardiology, Rambam Health Care Campus, and the Bruce Rappaport Faculty of Medicine, Technion, Haifa, Israel

Short title: Models for Predicting Heart Failure Outcome.

Equally contributing authors

*Corresponding author:

Oren Caspi, MD PhD

Heart Failure Unit, Department of Cardiology
Rambam Health Care Campus and
the B. Rappaport Faculty of Medicine, Technion, Haifa, Israel
POB 9602, Haifa 31096
Israel
Phone: 972-4-777-3781
Fax: 972-4-777-2176
E-mail: o_caspi@rmc.gov.il
Abstract

Introduction: 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 AI based prediction model for real-time decision support.

Methods: 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: (1) type of prediction method and (2) type and number of covariates. The primary outcome was one-year survival from hospital discharge. For the model-type axis we experimented with seven different methods: Logistic Regression (LR) with either L1 or L2 regularization, Random Forest (RF), Cox proportional hazards model (Cox), XGBoost, a deep neural-net (NeuralNet) and an ensemble classifier of all the above methods.

Results: We were able to achieve a 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%). Random Forest 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 modeling 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.
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. Acute heart failure (AHF) is the most common cause for hospitalization for adults (>65 years) in western societies. The prognosis of patients admitted with AHF is dismal, with 30% readmission rate and 5% to 15% mortality rate at 60 to 90 days post-discharge. 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. 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.

Advances in statistical machine learning 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 Machine Learning (ML), and include random-forest, gradient-boosted decision trees, sparse-regression methods such as LASSO and deep-neural nets.

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 center 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: (1) the type and number of covariates; (2) 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 Dec 2015. Eligible patients were those hospitalized with new-onset or worsening of preexisting heart failure as the primary cause of admission, using the European Society of Cardiology criteria. 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 was 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 one year mortality from hospital discharge. The secondary outcome was a composite of one year mortality or rehospitalization for acute heart failure at one year. Analysis for the primary outcome is included in the main manuscript. Analysis for the secondary outcome is included in the supplementary data.

**Data Collection and preparation**

Using observational data 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 supplemental data Table 8. A schematic representation of our data preparation pipeline is given in Figure 5.

**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 SBP and DBP values), identify ill-calculated features (e.g. BMI) 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 using median imputation (Supplementary data, Table 2). We also experimented with multivariate imputation by chained equations (MICE) imputation, and found that the former achieved slightly better results (Supplementary data, Table 1). Therefore, we chose to use median imputation for all our analyses. Finally, the data was normalized using z-scoring: \( \hat{x} = \frac{x - \text{mean}(X)}{\text{std}(X)} \). 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 one-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 – (1) type of prediction method and (2) type and number of covariates. For the model-type axis we experimented with 7 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), 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 logistic regression, Random Forest and majority vote classifiers were implemented using the Sci-kit learn python package$^{23}$. 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$^{21}$. 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 7 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 7 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. 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 recorded in admission. The full list of variables their subtype and the clinical priority is detailed in the supplementary data in Table 8. 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 5 sets of covariates per prediction method.

Thus, in total we considered $127+5=132$ sets of covariates, and 7 prediction methods, giving a total of $132 \times 7 - 7 = 917$ 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 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. 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), and report the following metrics, as suggest by Shameer et al.: Sensitivity, Specificity, Negative Predictive Value (NPV), Positive Predictive Value (PPV), accuracy and Brier Score. See Table 9 in the Supplementary data, 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 7 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 (one-year mortality) are presented over a total of 917 settings assessed on a temporally held-out data-set.

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 1), with the exception of the Random Forest model that significantly underperformed all other models (Table 2). Specifically, L1- and L2-regularized logistic regression (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 Random Forest (78.8%) performed slightly worse than other models and the ensemble model (81.2%) slightly better. The probable reason for Random Forest to perform less than other models here may be related to overfitting of the training data. Importantly, Cox based model, which is not based on a machine learning 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 (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 logistic regression 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 1). A comparative analysis demonstrated that inclusion of low priority variables improved the predictive performance of the models even when comparing analyses based on priority 0+ compared to priority 1+ (Table 5).
Calibration plots found a good calibration for most models with slight underestimation of the predicted risk for low risk patients using RF (Figure 2). To assess the clinical implication of the model in predicting mortality outcome for AHF patients we carried out a net-benefit analysis (Figure 3). A reasonable threshold for one year mortality risk in heart failure patients that entails consideration of advanced therapies (e.g. Heart transplantation or mechanical circulatory support) is 20%. 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 Random Forest 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 3. 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 performance. For this purpose we conducted the analysis for the L1 logistic regression model (Figure 4). 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 4 we summarize the most statistically significant covariates (p < 0.05) obtained when using L1 logistic regression 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 machine learning methodologies. The study’s main finding is that the choice of the predictive modeling 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 logistic regression and Cox regression were shown to be non-inferior to machine learning 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. The first, for acute heart failure 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. The second, for ambulatory chronic heart failure patient predicting 1-year and 3-year outcomes. 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 program and the medical therapy. While several models have been devised for risk stratifying heart failure patients at admission to hospital or ambulatory patients, 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 4) 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. 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. In recent years, there is growing evidence that admission hypochloremia is a marker for mortality and rehospitalization in chronic heart
failure patients\textsuperscript{51}, and the present study reiterates this finding and 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\textsuperscript{52,53}, 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 machine learning methodologies for predicting adverse events and outcomes based on electronic health records (EHR) is a promising approach, highly relevant to acute exacerbation of chronic diseases such as heart failure\textsuperscript{43,54-56}. 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., 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.\textsuperscript{57} Desai et al. recently demonstrated that multiple ML methodologies have minimal benefit over traditional logistic regression in the setting of chronic heart failure\textsuperscript{58}. However, the authors noted that machine learning approaches generally fared better when the models included numerous continuous variables such as laboratory tests.\textsuperscript{58}

A recent systematic review compared the results of logistic regression and other machine learning methods, across different risk-scores using clinical data, found no evidence of superior performance of ML techniques in terms of AUROC\textsuperscript{59}. More recent reviews have found and discussed similar finding regarding general benefit for using Deep-learning techniques\textsuperscript{60-62}. 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).\textsuperscript{31,63,64}

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\textsuperscript{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 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 machine learning 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 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.
Figures and Tables

<table>
<thead>
<tr>
<th>variable</th>
<th>value</th>
<th>Development</th>
<th>Validation</th>
<th>missing</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of patients</td>
<td>8520</td>
<td>2348</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Demographics</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>age, (years)</td>
<td>74.4 ± 12.1</td>
<td>74.3 ± 12.4</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Gender (Female)</td>
<td>49.4% (4209)</td>
<td>48.1% (1129)</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>BMI, Kg/m^2</td>
<td>29.9 ± 6.4</td>
<td>29.8 ± 6.4</td>
<td>8832</td>
<td></td>
</tr>
<tr>
<td>Comorbidities</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chronic Renal failure (n)</td>
<td>33.4% (2843)</td>
<td>28.2% (662)</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Valvular Heart Disease (n)</td>
<td>13.8% (1176)</td>
<td>12.9% (303)</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>AKI, % (n)</td>
<td>18.39% (1567)</td>
<td>21.97% (516)</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Diabetes</td>
<td>52.6% (4483)</td>
<td>54.6% (1282)</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>COPD</td>
<td>13.5% (1148)</td>
<td>15.1% (354)</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Ischemic heart disease</td>
<td>22.3% (1899)</td>
<td>18.9% (443)</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Atrial fibrillation</td>
<td>41.7% (3556)</td>
<td>41.8% (981)</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Smoking</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>not smoking</td>
<td>74.2% (6323)</td>
<td>70.4% (1654)</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>past smoker</td>
<td>14.8% (1257)</td>
<td>14.5% (341)</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>active smoker</td>
<td>11.0% (940)</td>
<td>15.0% (353)</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Clinical Parameters</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>diastolic blood pressure, mmHg</td>
<td>mean 68.6 ± 9.6</td>
<td>69.1 ± 9.6</td>
<td>2281</td>
<td></td>
</tr>
<tr>
<td>systolic blood pressure, mmHg</td>
<td>mean 131.2 ± 22.0</td>
<td>131.3 ± 19.5</td>
<td>2281</td>
<td></td>
</tr>
<tr>
<td>MAP, mmHg</td>
<td>mean 89.5 ± 11.6</td>
<td>89.9 ± 11.3</td>
<td>2281</td>
<td></td>
</tr>
<tr>
<td>body temperature, Celsius</td>
<td>first 37.1 ± 10.3</td>
<td>36.7 ± 0.4</td>
<td>2309</td>
<td></td>
</tr>
<tr>
<td>Estimated LVEF</td>
<td>mean 44.3 ± 47.3</td>
<td>44.2 ± 19.5</td>
<td>6000</td>
<td></td>
</tr>
<tr>
<td>Lab tests</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Albumin, g/dL</td>
<td>3.3 ± 0.5</td>
<td>3.2 ± 0.5</td>
<td>3672</td>
<td></td>
</tr>
<tr>
<td>BNP, ng/mL</td>
<td>first 1313.8 ± 1124.3</td>
<td>1112.3 ± 1011.4</td>
<td>5941</td>
<td></td>
</tr>
<tr>
<td></td>
<td>last 1306.1 ± 1118.6</td>
<td>1112.3 ± 1012.6</td>
<td>5941</td>
<td></td>
</tr>
<tr>
<td>BUN, mg/dL</td>
<td>mean 36.0 ± 20.2</td>
<td>36.5 ± 21.0</td>
<td>2472</td>
<td></td>
</tr>
<tr>
<td>Creatinine, mg/dL</td>
<td>mean 1.7 ± 1.1</td>
<td>1.6 ± 1.1</td>
<td>2473</td>
<td></td>
</tr>
<tr>
<td>RDW (%)</td>
<td>first 15.7 ± 1.9</td>
<td>16.1 ± 2.3</td>
<td>3496</td>
<td></td>
</tr>
<tr>
<td>Sodium, mmol/L</td>
<td>mean 138.3 ± 3.5</td>
<td>138.1 ± 3.7</td>
<td>2469</td>
<td></td>
</tr>
<tr>
<td>Hemoglobin, g/dL</td>
<td>first 11.6 ± 1.9</td>
<td>11.5 ± 2.0</td>
<td>2504</td>
<td></td>
</tr>
<tr>
<td></td>
<td>last 11.5 ± 1.8</td>
<td>11.4 ± 1.9</td>
<td>2498</td>
<td></td>
</tr>
<tr>
<td>Troponin, ng/mL</td>
<td>first 0.2 (1.5)</td>
<td>0.2 (1.1)</td>
<td>4088</td>
<td></td>
</tr>
<tr>
<td>HCT (%)</td>
<td>first 35.3 (5.7)</td>
<td>35.1 (5.9)</td>
<td>2502</td>
<td></td>
</tr>
<tr>
<td>Medical Therapy</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>β-Blockers (n)</td>
<td>background 64.8% (5518)</td>
<td>78.7% (1849)</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>In hospital 61.7% (5255)</td>
<td>88.4% (2075)</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>
Table 1 - Background clinical characteristics of patients at development and validation cohorts. BMI - Body mass index, AKI - Acute kidney injury, MAP - Mean arterial pressure, BNP - Brain Natriuretic peptide, RDW - Red cell distribution width, HCT - Hematocrit, MRA – Mineralocorticoid receptor antagonist.

<table>
<thead>
<tr>
<th></th>
<th>discharge</th>
<th>background</th>
<th>In hospital</th>
<th>In hospital</th>
<th>In hospital</th>
<th>In hospital</th>
<th>Target</th>
<th>1 Year mortality, % (n)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ace Inhibitors (n)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Angiotensin receptor blocker, % (n)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MRA, % (n)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Furosemide, % (n)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2 - Difference in predictive capacity for predicting the primary outcome between pairs of methods. 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-L1: Logistic Regression, NeuralNet: Neural Network, RF – Random Forest, XgBoost – Extreme Gradient Boosting Trees, Cox – Cox Regression, Ensemble – ensemble model using LR-L1, Cox, XgBoost and NeuralNet.
Table 3: The metrics of the prediction models using all the features (variables) for predicting the primary outcome. The results are using threshold of 0.5. AUROC=Area under Receiver Operator Curve; PPV-Positive Predictive Value; NPV-Negative Predictive Value.

| 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(%)*                         | 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   |
| Hyperlipidemia                  | 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*                       | 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  |

Table 4: Most significant variables predicting one year survival. BNP- Brain natriuretic peptide; RDW- Red blood cell distribution width; *First exams taken during hospitalization.
<table>
<thead>
<tr>
<th>Cox priority 0+</th>
<th>1.000</th>
<th>0.904</th>
<th>0.823</th>
<th>0.935</th>
<th>0.325</th>
<th>1.000</th>
<th>0.997</th>
<th>0.997</th>
<th>1.000</th>
<th>80.2%</th>
</tr>
</thead>
</table>

Table 5 - Difference in prediction capacity according to the model used and the assigned clinical priority of the variables used. 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 ≤ (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, XgBoost – Extreme Gradient Boosting Trees, Cox – Cox Regression, Ensemble – ensemble model using LR-L1 ,Cox, XgBoost and NeuralNet.
Funding Statement:
Funding support for this research was given by Yad Hanadiv Foundation, Israeli Science Foundation, and the Israeli Council for Higher Education for student funding.

Conflict of Interest statement:
No conflict of interests.

Data statement:
All data is included in the submission/manuscript file

References:
Results From the ROADMAP Study of Left Ventricular Assist Device and Medical Management in Ambulatory Heart Failure Patients:

- **Fail Non** prediction models, molecular markers, and diagnostic tests.
- **Non** ignorable missing mechanism; Nonignorable missing mechanism; Sampling properties and simulations. In: 2001.
- **Fonarow GC.** Risk Stratification for In-Hospital Mortality in Acutely Decompensated Heart Failure. &lt;SUBTITLE&gt;Classification and Regression Tree Analysis&lt;/SUBTITLE&gt; JAMA 2005;293:572. doi: 10.1001/jama.293.5.572
- **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. doi: 10.1136/heartjnl-2017-311198
- **Vickers AJ, Van Calster B, Steyerberg EW.** Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. *BMJ* 2016;i6. doi: 10.1136/bmj.i6
- **Samman-Tahhan A, Hedley JS, McCue AA, et al.** INTERMACS Profiles and Outcomes Among Non-Inotrope-Dependent Outpatients With Heart Failure and Reduced Ejection Fraction. *JACC Heart Fail* 2018;6:743-753. doi: 10.1016/j.jchf.2018.03.018


Figure Legends

Figure 1: Prediction of the models based on the set of clinical covariates utilized. AUROC of the models used split by clinical priority of the variables. The confidence interval (CI) in this graph are the prediction interval for the AUROC 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).

Figure 2: 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).

Figure 3: 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 RF being the least beneficial. The 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).

Figure 4: AUROC 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.

Figure 5 - The suggested pipeline of the predicting process.
Figure 1
165x165 mm (4.1 x DPI)
Figure 2
165x165 mm (4.1 x DPI)
Figure 3
165x165 mm (4.1 x DPI)
Figure 4

165x110 mm (4.1 x DPI)
Figure 5
165x57 mm (4.1 x DPI)
Conclusions

- The choice of the predictive modeling 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.