ESC European Society of Cardiology

European Journal of Heart Failure (2020) **22**, 139–147
doi:10.1002/ejhf.1628

RESEARCH ARTICLE

Improving risk prediction in heart failure using machine learning

Eric D. Adler[1], Adriaan A. Voors[2], Liviu Klein[3], Fima Macheret[4], Oscar O. Braun[5], Marcus A. Urey[1], Wenhong Zhu[4], Iziah Sama[2], Matevz Tadel[6], Claudio Campagnari[7†], Barry Greenberg[1*†], and Avi Yagil[1,6]

[1]Division of Cardiology, Department of Medicine, UC San Diego, La Jolla, CA, USA; [2]University of Groningen, University Medical Center Groningen, Groningen, The Netherlands; [3]Division of Cardiology, Department of Medicine, UC San Francisco, San Francisco, CA, USA; [4]Altman Clinical and Translational Research Institute (ACTRI), UC San Diego, La Jolla, CA, USA; [5]Cardiology, Department of Clinical Sciences, Lund University and Skåne University Hospital, Lund, Sweden; [6]Physics Department, UC San Diego, La Jolla, CA, USA; and [7]Physics Department, UC Santa Barbara, Santa Barbara, CA, USA

Received 10 July 2019; revised 24 August 2019; accepted 25 August 2019; online publish-ahead-of-print 12 November 2019

Background Predicting mortality is important in patients with heart failure (HF). However, current strategies for predicting risk are only modestly successful, likely because they are derived from statistical analysis methods that fail to capture prognostic information in large data sets containing multi-dimensional interactions.

Methods and results We used a machine learning algorithm to capture correlations between patient characteristics and mortality. A model was built by training a boosted decision tree algorithm to relate a subset of the patient data with a very high or very low mortality risk in a cohort of 5822 hospitalized and ambulatory patients with HF. From this model we derived a risk score that accurately discriminated between low and high-risk of death by identifying eight variables (diastolic blood pressure, creatinine, blood urea nitrogen, haemoglobin, white blood cell count, platelets, albumin, and red blood cell distribution width). This risk score had an area under the curve (AUC) of 0.88 and was predictive across the full spectrum of risk. External validation in two separate HF populations gave AUCs of 0.84 and 0.81, which were superior to those obtained with two available risk scores in these same populations.

Conclusions Using machine learning and readily available variables, we generated and validated a mortality risk score in patients with HF that was more accurate than other risk scores to which it was compared. These results support the use of this machine learning approach for the evaluation of patients with HF and in other settings where predicting risk has been challenging.

Keywords Heart failure • Outcomes • Machine learning

Introduction

Predicting mortality in heart failure (HF) is critically important to patients, their providers, healthcare systems, and third-party payers alike. The ability to accurately assess outcomes in patients with HF, however, has proven to be a difficult task. Although a number of tools, including biomarkers,[1,2] risk scores[3–10] and their combination[11–13] have been developed for this purpose, most have achieved only modest success, particularly when they are employed in HF populations other than those from which the score was derived.[14–19] For instance, the Meta-Analysis Global Group in Chronic Heart Failure (MAGGIC) risk score[8] achieved a C-statistic for the area under the curve (AUC) of 0.74 for predicting mortality risk in a large cohort of patients followed in the Swedish Heart Failure Registry, but performed less well in patients on a transplant

*Corresponding author. Division of Cardiology, Department of Medicine, UC San Diego, 9452 Medical Center Drive, La Jolla, CA 92037, USA. Tel: +1 858 246 2987, Email: bgreenberg@ucsd.edu
†Co-senior authors on this manuscript.

waiting list and in a large population of ambulatory HF patients in the US where the C-statistics were 0.69 and 0.70, respectively.[16,18] The results with previous HF risk scores are likely due to several causes including dependence on variables that are not universally available and temporal dispersion in the collection of variables so that the state of the patient at a discrete point in the course of their disease is not captured.[14,15] Most importantly, though, is that previous HF prediction tools were largely derived using statistical analysis methods that fail to capture multi-dimensional correlations that contain prognostic information. In contrast, machine learning, which has long been used by other fields, including high-energy physics[20] to discriminate between signal and background, uses non-parametric analysis methods to incorporate these interactions. As this approach offers theoretical advantages over ones used in the past, we hypothesized that it could be used to generate a model that more accurately predicts mortality risk among patients with HF than previously published scores.

Methods

The study conforms with the principles outlined in the Declaration of Helsinki[21] and was approved by the University of California, San Diego (UCSD) Health Institutional Review Board.

Identification of three separate cohorts of patients with heart failure

This study utilizes data derived from three distinct patient populations. The algorithm was developed from retrospective analysis of a cohort of patients followed at UCSD who were identified at the time HF was first noted in their medical records. To locate these patients, we queried the institutional electronic medical record (EMR) (EPIC 2015, Verona, WI, USA) for patients with the earliest recorded occurrence (HF index event) of the International Classification of Disease-10 (ICD-10) codes listed in the online supplementary *Table S1*. This cohort includes patients admitted for HF, had emergency room visits, or were seen in an outpatient setting at the time when HF was first coded, regardless of ejection fraction between 2006 and 2017. For every patient in this cohort, we extracted results from the first instance of the complete blood count, comprehensive metabolic panel, vital sign measurement, electrocardiogram, and echocardiogram that occurred within 7 days of the patient's HF index event. This relatively narrow time window allows for a precise capture of the state of the patient and preserves the correlation between variables that could be diluted over time.

The second set of patients came from the University of California, San Francisco Medical Center (UCSF) and was identified following an identical procedure (ICD-10 codes and database extraction script) as the one used for the data extraction at UCSD. This was possible as both medical centres use the same underlying EMR system (EPIC 2015). For this cohort, the study was approved by the UCSF Health Institutional Review Board.

The third set of patients were those from the European A systems BIOlogy Study to TAilored Treatment in Chronic Heart Failure (BIOSTAT-CHF) project, which enrolled from 69 centres in 11 European countries between 2010 and 2012 to determine profiles of patients with HF that do not respond to recommended therapies, despite anticipated up-titration. The design and first results of the study and patients have been described elsewhere.[22]

Derivation of the Machine learning Assessment of RisK and EaRly mortality in Heart Failure (MARKER-HF) risk model

We developed a model using data from 14 589 patients identified at the time of the first occurrence of HF in the UCSD EMRs. After exclusion of patients older than 80 years at the time of the clinical encounter ($n = 2444$), with cardiac implantable electronic device (CIED) including implantable cardioverter-defibrillator or pacemaker ($n = 612$), with findings consistent with sepsis ($n = 40$), who had died within 7 days of initial encounter ($n = 254$), who had medical records with obvious database errors such as recorded date of death prior to a recorded clinical encounter or spurious lab results ($n = 651$) and those who had missing data ($n = 4766$), the cohort used in the derivation and initial validation consisted of 5822 patients, all of whom were under 80 years of age. The rationale for excluding patients older than 80 years in our derivation cohort is that age is the best predictor of death in older patients and that inclusion of this variable would have diluted the ability of the model to 'learn' the salient clinical features relevant to all ages. For patients with a recorded date of death in the EMR, the time from index event to mortality was calculated. The 407 patients who died within 90 days of index event were classified as high-risk. For patients without a recorded date of death, we calculated the time interval from index event to the last known follow-up defined as the latest date of an actual physical encounter recorded in the EMR. The 966 patients with a last known follow-up 800 (or more) days after index event and with no recorded date of death were defined as low-risk.

MARKER-HF was then designed by building a model to discriminate between high and low-risk patients. The choice to not include the population with 'intermediate' outcomes (mortality after 90 days or no recorded follow-up after 800 days) in the derivation of the model was dictated by two considerations. First, the algorithm that was chosen to build the model is based on automated training with two well-defined populations, which in this case are the high and low-risk cohorts. Second, by excluding the patients with intermediate outcomes, the two populations used in the training are clearly separated in outcomes, and this helps the algorithm focus on the most important distinguishing characteristics and correlations.

The model was built by training a boosted decision tree (BDT) algorithm[23] to relate a subset of the patient data, as detailed below, to the two extreme outcomes. The BDT was based on the *AdaBoost* algorithm,[24] as implemented in the TMVA toolkit in version 6.13/01 of the CERN ROOT software package.[25,26] The number of variables selected was limited in order to increase inclusiveness by avoiding loss of patients with missing data and to minimize overfitting that can result in over-training and loss of robustness when the score is applied to other populations. The variables selected are inexpensive to obtain and commonly available. We used this algorithm to iteratively select the smallest, most common and discriminating subset of variables out of those available in the UCSD cohort. This approach identified a composite of eight variables [diastolic blood pressure, creatinine, blood urea nitrogen, haemoglobin, white blood cell count, platelets, albumin, and red blood cell distribution width (RDW)] that provided excellent discrimination between high and low-risk patients. In developing the model, the high and low-risk cohorts were randomly divided into equal-sized *training (derivation)* and *test (validation)* samples. The high and low-risk training samples were used to train the BDT; the validation samples were used to test the performance in a statistically independent way. Training was performed assuming equal a-priori probabilities

for any individual patient to belong to the low or high-risk samples. Modifying these relative probabilities to 20–80% or 80–20% did not result in significant changes in performance. The output of the BDT algorithm for a given patient is the MARKER-HF score, where a higher score indicates a higher likelihood of death. Although the MARKER-HF score was derived from the high and low-risk patients, it can be calculated for all patients, not just those in the low and high-risk cohorts, and subsequently testing was performed in patients across the full spectrum of mortality risk in all three populations studied.

MARKER-HF performance on the University of California, San Diego cohort

The performance of MARKER-HF was tested by calculating the C-statistic in the separation of the high and low-risk cohorts, and by examining the relationship between MARKER-HF and life expectancy in the full cohort, i.e. including patients with intermediate outcomes, as defined previously. In addition, MARKER-HF performance is studied in subsets of patients divided by gender, ethnicity, inpatients vs. outpatients, and by the ICD code used to identify HF at the identifying occurrence.

External validation of the MARKER-HF risk model

In order to determine if the score accurately predicts risk of death in other distinct patient cohorts, we compared C-statistics and correlation between MARKER-HF and life expectancy in patients from the UCSD, UCSF and BIOSTAT-CHF populations. For this test RDW information was imputed in both the BIOSTAT-CHF and UCSF patients, since it was not available for these patients.

Results

Description of the three cohorts and the covariates used in MARKER-HF

Age, gender, and mean values for the eight variables used to construct the MARKER-HF score of the patients studied in the three populations are summarized in *Table 1*.

Table 1 Age, gender and variables used in MARKER-HF in the three cohorts

	UCSD ($n = 5822$)	UCSF ($n = 1516$)	BIOSTAT-CHF ($n = 888$)
Demographics			
Age, years	59	57	65
Female sex, %	41	41	28
MARKER-HF covariates, mean (SD)			
Diastolic blood pressure, mmHg	70 (14)	68 (13)	77 (15)
BUN, mg/dL	23 (14)	25 (16)	28 (17)
Creatinine, mg/dL	1.2 (0.7)	1.4 (0.9)	1.2 (0.4)
Haemoglobin, g/dL	11.3 (2.6)	11.6 (2.4)	13.3 (2.0)
White blood cell count, $\times 10^3$/L	9.6 (6.0)	9.3 (7.7)	8.3 (3.0)
Platelet count, $\times 10^3$/L	211 (108)	206 (107)	231 (84)
Albumin, g/dL	3.6 (0.7)	3.1 (0.7)	3.2 (0.9)
RDW, %	15.4 (2.5)	NA	NA

BUN, blood urea nitrogen; NA, not available; RDW, red blood cell distribution width; SD, standard deviation; UCSD, University of California, San Diego; UCSF, University of California, San Francisco.

Performance of MARKER-HF in the University of California, San Diego cohort

Using a BDT algorithm, a score between −1 and +1 was generated for each patient. As shown in *Figure 1A*, the MARKER-HF score was highly effective in separating the high and low-risk UCSD patients. The un-binned Kolmogorov–Smirnov probabilities for the compatibility between distributions of MARKER-HF scores for the training and validation cohorts are 84% (high-risk cohorts) and 31% (low-risk cohorts), demonstrating no overtraining by the underlying algorithm. To further visualize how MARKER-HF discriminates between high and low-risk patients in the validation sample, we constructed a receiver operator characteristic (ROC) curve (*Figure 1B*) and calculated 95% confidence intervals for the area under the ROC curve (AUC or C-statistic). The AUC in the validation sample is 0.88 (95% confidence interval 0.85–0.90). It is important to note that none of the eight input covariates is a powerful discriminator in and of itself, since their individual AUCs are in the range 0.54–0.78.

We calculated life expectancy in exclusive ranges of MARKER-HF further distinguishing the training and validation UCSD cohorts and splitting the remaining population randomly between the two. To allow adequate follow-up time, we only included here patients with index event prior to 31 December 2015 ($n = 1986$). *Figure 2A* shows 1-year survival rates in ranges of MARKER-HF score for these patients. The red (blue) curves show the results for the training (validation) cohorts, respectively. Note that while MARKER-HF was developed excluding intermediate risk patients, the distribution shown in *Figure 2A* demonstrates the applicability of the MARKER-HF algorithm across the full range of risk.

MARKER-HF performance on different cohorts

The performance of MARKER-HF was subsequently further validated in two additional patient populations. For this purpose, we used a cohort of patients from UCSF identified at the first recorded occurrence of HF in the EMR and a cohort of patients from a European registry, BIOSTAT-CHF, which included HF patients who had not responded to recommended therapies. To this end, we first defined high and low-risk cohorts for UCSF and BIOSTAT-CHF in the same way as was done for the UCSD cohort. We then assigned a MARKER-HF score to each patient using the model developed on the UCSD cohort, and finally calculated the AUCs for separating high and low-risk patients. The results, including a comparison with those on the UCSD cohort, are summarized in *Table 2* and *Figure 2B*. As mentioned previously, RDW information was not available in the UCSF and BIOSTAT-CHF cohorts. For these patients we imputed the RDW value as the mean RDW

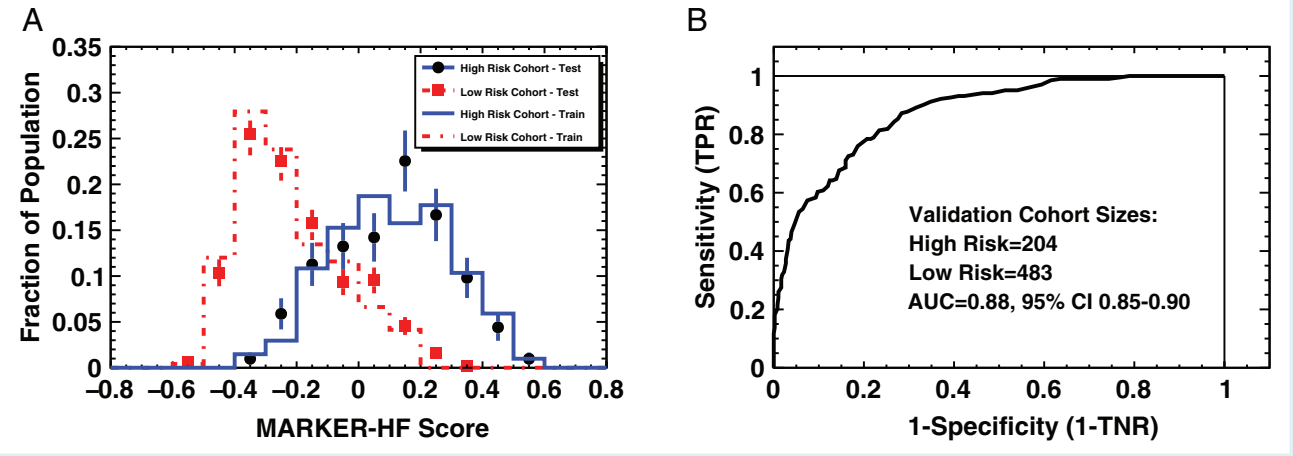

Figure 1 Comparison of high-risk and low-risk populations. (*A*) Distributions of MARKER-HF scores in the high-risk (blue) and low-risk (red) populations. Both training (histograms) and testing/validation (data points) samples are shown. The vertical error bars on the data points represent the $\sqrt{n}$ statistical uncertainties in each data point. The corresponding uncertainties in the histogram are of the same order, and they are not shown for simplicity. (*B*) Receiver operating characteristic curve for discrimination of high-risk from low-risk patients in the validation cohort. AUC, area under the curve; CI, confidence interval; TNR, true negative rate; TPR, true positive rate.

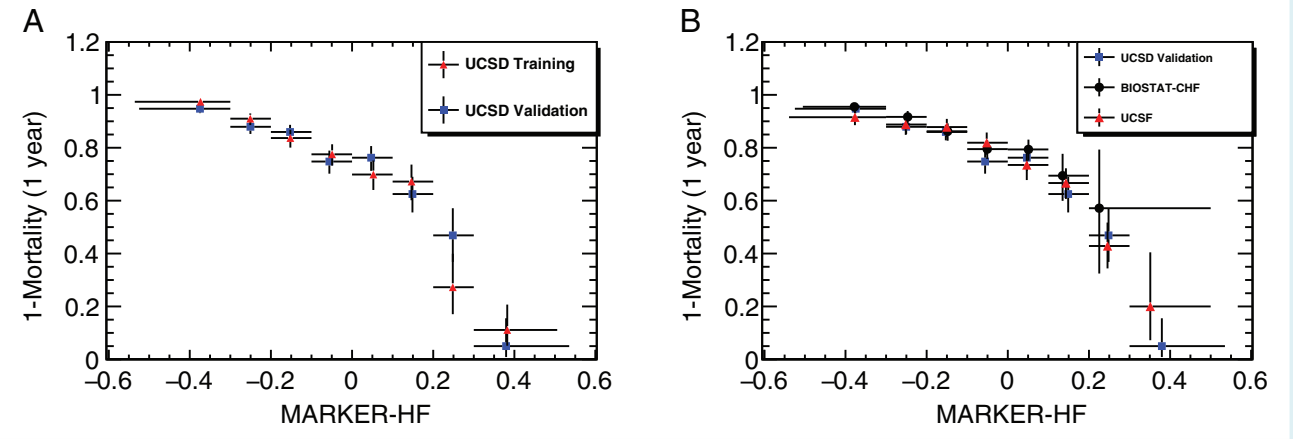

Figure 2 MARKER-HF performance as a risk score: 1-year survival rate in ranges of MARKER-HF comparing (*A*) University of California, San Diego (UCSD) training and validation cohorts, and (*B*) the UCSD (n = 1986, index event before 31 December 2015), University of California, San Francisco (UCSF; n = 1516, index event before 31 December 2016), and BIOSTAT-CHF (n = 888, index event required to be at least 1 year prior to end-of-study date) cohorts. The horizontal bar on each point represents the range in MARKER-HF of the corresponding bin. The data point is placed at the weighted-mean position in the bin range. The vertical error bar on each point represents the one sigma Clopper–Pearson interval. Note the reproducibility between cohorts over wide range of risks.

from the UCSD cohort. As shown in *Table 2*, when we tested this strategy on the UCSD cohort there was only a small deterioration in performance (row 1 vs. row 2). Note that for rows 2, 3, and 4 of *Table 2* MARKER-HF was computed using the same RDW value in every cohort, which effectively removes it as a distinguishing feature between high and low-risk cohorts. Within the somewhat limited statistical power of the comparison, the AUCs on the three samples are consistent with each other (χ^2/n.d.o.f = 4.1/2, P = 0.13) indicating the ability of MARKER-HF to predict outcomes across different populations. To test the performance in a broader set of patients, we compared the 1-year mortality prediction in ranges of MARKER-HF (*Figure 2B*) in the UCSF and BIOSTAT-CHF populations following the same procedure described previously for *Figure 2A*. We observed consistent performance, within statistics, over the full risk spectrum in all three populations.

Comparing MARKER-HF to NT-proBNP in determining risk

We also compared the performance of MARKER-HF with N-terminal pro-B-type natriuretic peptide (NT-proBNP), a well-validated biomarker associated with HF for which elevation

Table 2 MARKER-HF area under the curve (AUC) in different cohorts. Comparisons of AUCs to separate high- and low-risk cohorts in the University of California, San Diego validation cohort, the University of California, San Francisco cohort, and the BIOSTAT-CHF cohort

Cohort	High-risk, *n*	Low-risk, *n*	AUC	95% CI
UCSD (all variables)	204	483	0.88	0.85–0.90
UCSD (RDW imputed)	204	483	0.87	0.84–0.89
UCSF[a]	135	330	0.81	0.77–0.86
BIOSTAT-CHF[a]	35	228	0.84	0.78–0.90

CI, confidence interval; RDW, red blood cell distribution width; UCSD, University of California, San Diego; UCSF, University of California, San Francisco.
[a]Note that RDW was imputed in both UCSF and BIOSTAT-CHF cohorts.

is strongly and independently associated with mortality.[1,2] When we examined NT-proBNP values stratified by the MARKER-HF score in UCSD patients for whom it was available (*Figure 3A*), we found that increasing MARKER-HF scores are clearly linked with increasing NT-proBNP values. However, in contrast to the more powerful AUC of 0.88 for MARKER-HF, the mortality-risk predictive power of NT-proBNP has an AUC of only 0.69 to separate high and low-risk patients (*Figure 3B*). We chose to exclude NT-proBNP from the set of variables in the training of the model because of its low availability (~50%) in the UCSD cohort. Furthermore, as shown in the online supplementary *Table S2*, we found that adding NT-proBNP to the other eight variables in MARKER-HF did not result in a measurable improvement for the predictive power of the model.

MARKER-HF testing in subgroups of the University of California, San Diego population

Our results show that MARKER-HF performs consistently across the full range of risk in subgroups of the UCSD population based on sex, race, and site of entry into the study (inpatient vs. outpatient) (*Figure 4A–4C*). We also compared patients with an ICD code for HF (I50) to the ICD code for pulmonary oedema (J81) and found no difference in the predictive power of MARKER-HF (*Figure 4D*). Finally, as shown in the online supplementary *Table S3* and *Figure S1*, when we compared MARKER-HF in UCSD patients <55 years, 55–66 years and >66 years, no significant differences were seen in the correlation coefficients relating MARKER-HF score and time to death between the three age groups.

Comparison of MARKER-HF to other scores

We compared the performance of MARKER-HF in the UCSD population to other risk scores. The C-statistic for the AUC for MARKER-HF (0.88) was significantly greater than that for either the Intermountain Risk Score (IMRS),[7] the Get With the Guidelines-HF (GWTG-HF) risk score[6] and the Acute Decompensated Heart Failure Registry (ADHERE) score[3] where the C-statistics were 0.78, 0.74 and 0.63, respectively ($P < 0.001$ for each compared to MARKER-HF). The Pearson correlation coefficient between MARKER-HF and time to death ($r = -0.41$) was also superior to those of IMRS (−0.31), GWTG-HF (−0.25) and ADHERE (−0.14). Similar patterns were found using different measures of correlation (Kendall and Spearman). In addition, as shown in the online supplementary *Table S4*, MARKER-HF was superior to both the GWTG-HF and the ADHERE risk

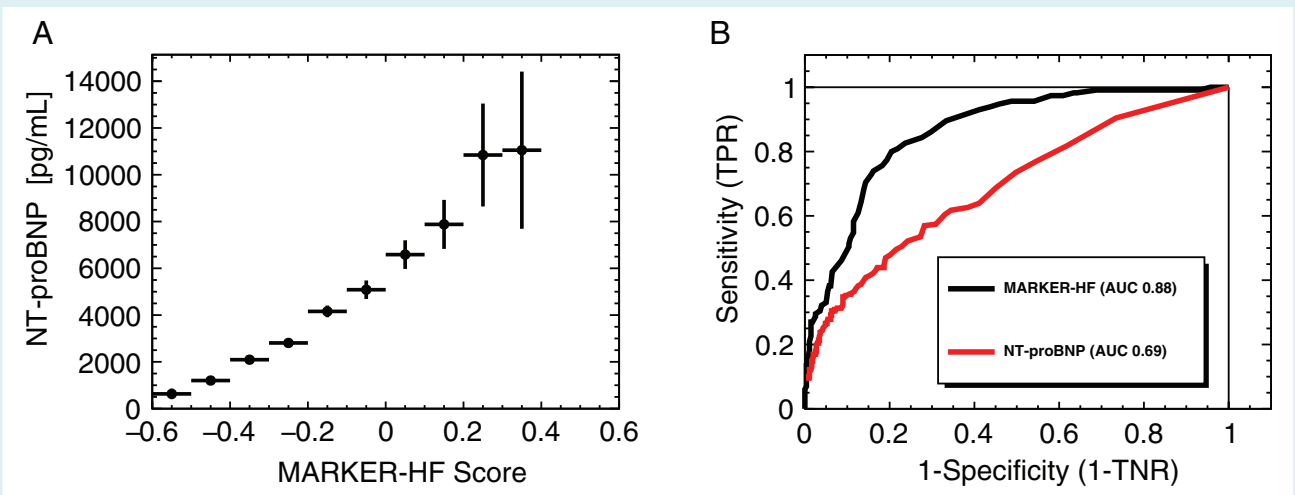

Figure 3 N-terminal pro-B-type natriuretic peptide (NT-proBNP) in MARKER-HF cohort and comparison as risk predictors: (*A*) mean NT-proBNP vs. MARKER-HF in the subset of patients that had NT-proBNP tested. Error bars represent the one sigma statistical uncertainty on the mean. (*B*) Receiver operating characteristic curves for NT-proBNP (red) and MARKER-HF (black). MARKER-HF has significantly higher C-statistic. AUC, area under the curve; TNR, true negative rate; TPR, true positive rate.

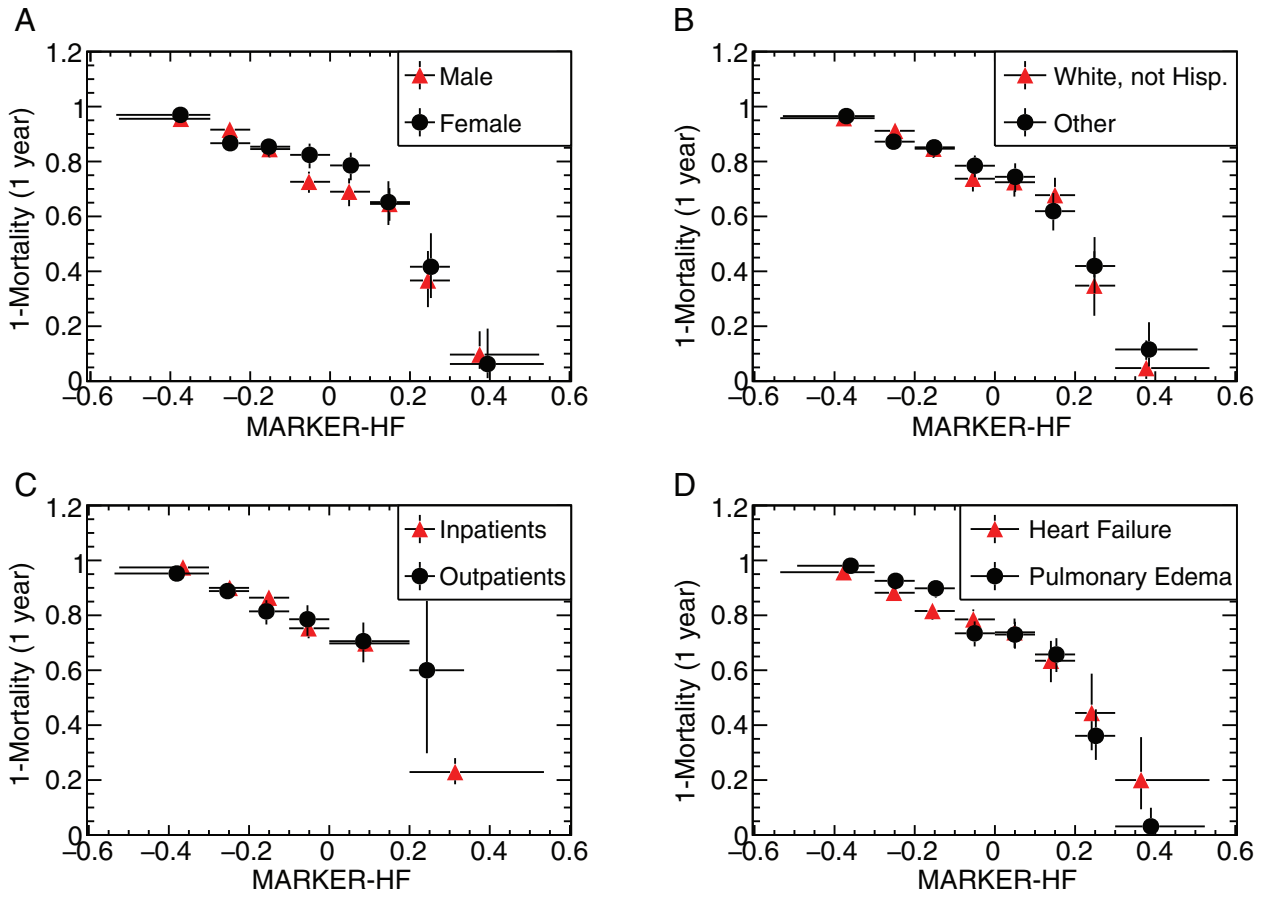

Figure 4 Comparisons of 1-year mortality in subsets of the University of California, San Diego cohorts: (*A*) male vs. female; (*B*) white non-Hispanics vs. other races; (*C*) inpatients vs. outpatients; (*D*) patients with main index diagnosis being heart failure vs. patients with main index diagnosis being pulmonary oedema.

scores when these were applied to the UCSD and BIOSTAT-CHF populations.

Discussion

We applied novel machine learning analytical techniques to create a mortality risk model in patients with HF. The resulting MARKER-HF score demonstrated excellent discriminatory power in assessing mortality risk with an AUC of 0.88. MARKER-HF performed well across a wide range of risk, and in subgroups based on gender, race and whether patients were identified in the clinic or during hospitalization. External validation of this risk score model in two independent study cohorts yielded similarly high AUCs, indicating the applicability of this machine learning approach to other populations of patients with HF. In addition, MARKER-HF performed better than other risk scores in the three populations in which it was tested.

Numerous risk scores have been developed to predict mortality in patients with HF.[3–13] These tools generally do not have the discriminatory power of MARKER-HF. The moderate to poor performance of previous risk prediction models has often been ascribed to the heterogeneity of patients with HF. However, their modest performance is more likely due to the failure of the statistical methods used in their design to capture prognostically important multi-dimensional correlations between the state of the patient and their outcome. In contrast, MARKER-HF was specifically designed to capture these correlations. We also ensured that covariates were taken within a limited time frame to ensure that the correlations between them defined the patient at a specific well-defined point in time. This approach resulted in the generation of a risk score that greatly improves the accuracy of determining risk of mortality in HF patients.

The eight variables used in MARKER-HF (diastolic blood pressure, creatinine, blood urea nitrogen, haemoglobin, white blood cell count, platelets, albumin, and RDW) were selected in an iterative optimization process based on learning done in the training cohort. Starting from the superset of all available variables, a set of variables was selected that maximized discriminating power (as indicated by the C-statistic). The number of variables selected was limited in order to increase inclusiveness by avoiding loss of patients with missing data and also to minimize over-fitting that can result

in over-training and loss of robustness when the score is applied to other populations.

The variables selected are inexpensive to obtain and commonly available. They are generally ordered as part of the initial evaluation of most patients with HF. Of note, the addition of other clinical variables did not significantly improve the performance of MARKER-HF, including several variables more directly physiologically associated with HF, such as heart rate, left ventricular ejection fraction, QRS duration. In order to keep the number of variables low and to avoid overfitting, we excluded this additional information from the construction of MARKER-HF. When creating MARKER-HF we chose not to use age as a variable, even though it was available in all of the patients in the cohorts and is strongly associated with death. By including age as a variable in the training, the 'machine' would essentially learn that when you are old, you are more likely to die. Thus, the addition of age would dilute our ability to identify salient clinical variables and associations that were predictive. To assess the impact of age on the model, we performed additional analysis in which age was added to the set of variables on which the algorithm was trained. The AUC of the retrained BDT was identical to what we had found initially (i.e. without adding age) but that there was slightly worse overtraining. Furthermore, when this nine variable model of MARKER-HF was assessed in to an expanded UCSD population that included patients between 80–89 years, the AUC was not improved confirming our hypothesis that age dilutes the ability of the 'machine' to detect salient clinical features that are predictive of outcome. Other variables traditionally associated with outcomes in HF, including gender, and ethnicity,[27,28] did not significantly improve the predictive accuracy of MARKER-HF. Moreover, MARKER-HF distributions showed no dependence on clinical setting in which the patient was identified (outpatient vs. hospitalization) or whether the primary diagnosis was listed as pulmonary oedema or another HF diagnosis.

Previous studies have shown a relationship between NT-proBNP levels and mortality in patients with HF. The demonstration that NT-proBNP and MARKER-HF closely approximated each other in their ability to predict mortality serves as additional and independent validation of the relevance and efficacy of MARKER-HF. We decided, however, not to include NT-proBNP in the derivation of MARKER-HF as its inclusion would have greatly reduced the number of patients that were available. In support of this decision was the finding that the addition of NT-proBNP as a ninth variable in MARKER-HF did not improve the predictive accuracy of the model. Finally, when we compared the predictive power of NT-proBNP to MARKER-HF in the subset of our cohort with available NT-proBNP levels, we observed that MARKER-HF was superior to the natriuretic peptide.

The applicability of MARKER-HF was further demonstrated by observing a very similar performance in cohorts of patients from UCSF and the BIOSTAT-CHF study as was seen in the UCSD population. The ability of MARKER-HF to largely retain its predictive ability in two additional, distinct cohorts despite having to remove one of the test variables (RDW) speaks to the validity of our findings across a spectrum of HF patients. It is also worth noting that the performance of MARKER-HF was superior to that of either the IMRS, GWTG-HF or AHDERE risk scores in the UCSD population and also to the GWTG-HF or AHDERE risk scores in two additional independent populations in which it was tested.

Our study has potentially important clinical ramifications. HF remains a leading cause of death in the United States as well as in most other countries in the world.[27–30] A central challenge in HF management is the identification of mortality risk, as the clinical course is often unpredictable. Prompt identification of high-risk patients using MARKER-HF could help allow for the deployment of additional resources, including more intensive assessment by physicians and health care extenders, in appropriate cases. It could be used to alert patients and their families of the severity of the patient's illness and encourage discussions regarding advanced care or end of life directives. MARKER-HF might also be useful in the evaluation and prioritization of patients for interventions such as CIEDs, mechanical circulatory support and heart transplantation, although this possibility would need to be confirmed in future studies.

Limitations

MARKER-HF was initially derived from a single centre (with two hospitals) in San Diego, California, and hence may be subject to a particular demographic selection bias. However, it performed quite well in defining risk of mortality in two separate groups of patients, including the European-based BIOSTAT-CHF cohort. In order to capture higher order correlations between the input variables, we decided not to impute missing data and this may have introduced additional bias. Although this decision resulted in an approximate 40% reduction in the study cohort size, we found that the distributions of the other individual variables for the rejected patients did not differ significantly from those in the patients who were retained in the analysis, suggesting that this decision did not influence the characteristics of the patients used to generate MARKER-HF.

For the derivation stage of the score, we excluded certain groups of patients in order to enable the algorithm to capture the salient features distinguishing high from low-risk cases. This included patients over 80 years old as well as those with CIED and those with left ventricular assist devices. We explicitly excluded elderly patients when deriving the score as they are more likely to die from causes not necessarily related to HF and their presence in the cohort used to define the model will weaken it. However, since MARKER-HF was derived and validated in HF populations that are relatively young, its performance should be assessed in older populations to more fully determine its generalizability. Similarly, we excluded patients with a left ventricular assist device or a CIED (at index event) when deriving the score as our goal was to assess prognosis at a time as close as possible to the onset of the disease and its earliest recorded diagnosis. Sensitivity analysis in which patients with CIED were included (data not shown) did not significantly affect either the derivation or validation of the score. Future study will be required to assess the utility of MARKER-HF in these patients.

Future directions

Although MARKER-HF performed well in the three populations studied, further validation in other populations including those on transplant waiting lists and in countries not represented in this report would be helpful in determining its usefulness in clinical practice. In addition, we are currently in the process of prospectively utilizing MARKER-HF in the UCSD HF population to determine its impact on decision making in clinical practice.

Conclusions

Using a machine learning approach, we generated and validated a risk score that was highly predictive of mortality in patients with HF. MARKER-HF, which uses readily available variables, provided consistent results in three different cohorts, within various subgroups (e.g. men vs. women, inpatients vs. outpatients) and across a full range of risk. These findings suggest that MARKER-HF can be a clinically useful tool in the management of patients with HF. They also raise the possibility that the approach taken to develop MARKER-HF may be useful for determining risk of events in other disease states. However, this possibility will need to be confirmed by additional studies in the future.

Supplementary Information

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

Table S1. ICD-10 codes used to identify patients in the UCSD cohort.

Table S2. Predictive value of MARKER-HF in the UCSD population using the original eight variables and with the addition of age and NT-proBNP.

Table S3. Correlation coefficients between MARKER-HF and time to death for the three age groups.

Table S4. Performance of MARKER-HF, GWTG-HF risk score and the ADHERE risk score in the UCSD, UCSF and BIOSTAT-CHF populations.

Figure S1. Distribution of MARKER-HF scores in different age groups of the UCSD population.

Acknowledgements

We thank the entire Sulpizio Cardiovascular Center's clinical staff and Prof. Oren Caspi (Technion) for reviewing our preliminary results and his insights. We thank the Altman Clinical and Translational Research Institute (ACTRI) staff for their support in extracting the data.

Funding

BIOSTAT-CHF was supported by the European Commission (FP7-242209-BIOSTAT4 CHF; EudraCT 2010–020808-29).

Conflict of interest: none declared.

References

1. Januzzi JL Jr, Sakhuja R, O'Donoghue M, Baggish AL, Anwaruddin S, Chae CU, Cameron R, Krauser DG, Tung R, Camargo CA Jr, Lloyd-Jones DM. Utility of amino-terminal pro-brain natriuretic peptide testing for prediction of 1-year mortality in patients with dyspnea treated in the emergency department. *Arch Intern Med* 2006;**166**:315–320.
2. McKie PM, Cataliotti A, Lahr BD, Martin FL, Redfield MM, Bailey KR, Rodeheffer RJ, Burnett JC Jr. The prognostic value of N-terminal pro-B-type natriuretic peptide for death and cardiovascular events in healthy normal and stage A/B heart failure subjects. *J Am Coll Cardiol* 2010;**55**:2140–2147.
3. Fonarow GC, Adams KF Jr, Abraham WT, Yancy CW, Boscardin WJ; ADHERE Scientific Advisory Committee, Study Group, and Investigators. Risk stratification for in-hospital mortality in acutely decompensated heart failure: classification and regression tree analysis. *JAMA* 2005;**293**:572–580.
4. 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.
5. Alba AC, Rao V, Ivanov J, Ross HJ, Delgado DH. Usefulness of the INTERMACS scale to predict outcomes after mechanical assist device implantation. *J Heart Lung Transplant* 2009;**28**:827–833.
6. Peterson PN, Rumsfeld JS, Liang L, Albert NM, Hernandez AF, Peterson ED, Fonarow GC, Masoudi FA. A validated risk score for in-hospital mortality in patients with heart failure from the American Heart Association Get With the Guidelines Program. *Circ Cardiovasc Qual Outcomes* 2010;**3**:25–32.
7. Horne BD, May HT, Kfoury AG, Renlund DG, Muhlestein JB, Lappé DL, Rasmusson KD, Bunch TJ, Carlquist JF, Bair TL, Jensen KR, Ronnow BS, Anderson JL. The Intermountain Risk Score (including the red cell distribution width) predicts heart failure and other morbidity endpoints. *Eur J Heart Fail* 2010;**12**:1203–1213.
8. Sartipy U, Dahlstrom U, Edner M, Lund LH. Predicting survival in heart failure: validation of the MAGGIC heart failure risk score in 51,043 patients from the Swedish Heart Failure Registry. *Eur J Heart Fail* 2014;**16**:173–179.
9. Nilsson J, Ohlsson M, Hoglund P, Ekmehag B, Koul B, Andersson B. The International Heart Transplant Survival Algorithm (IHTSA): a new model to improve organ sharing and survival. *PLoS One* 2015;**10**:e0118644.
10. Pocock SJ, Ariti CA, McMurray JJ, Maggioni A, Kober L, Squire IB, Swedberg K, Dobson J, Poppe KK, Whalley GA, Doughty RN; Meta-Analysis Global Group in Chronic Heart Failure. Predicting survival in heart failure: a risk score based on 39 372 patients from 30 studies. *Eur Heart J* 2013;**34**:1404–1413.
11. May HT, Horne BD, Levy WC, Kfoury AG, Rasmusson KD, Linker DT, Mozaffarian D, Anderson JL, Renlund DG. Validation of the Seattle Heart Failure Model in a community-based heart failure population and enhancement by adding B-type natriuretic peptide. *Am J Cardiol* 2007;**100**:697–700.
12. Sawano M, Shiraishi Y, Kohsaka S, Nagal T, Goda A, Mizuno A, Sujino Y, Nagatomo Y, Kohno T, Anzai T, Fukuda K, Yoshikawa T. Performance of the MAGGIC heart failure risk score and its modification with the addition of discharge natriuretic peptides. *ESC Heart Fail* 2018;**5**:610–619.
13. Doumouras BS, Lee DS, Levy WC, Alba AC. An appraisal of biomarker-based risk-scoring models in chronic heart failure: which one is best? *Curr Heart Fail Rep* 2018;**15**:24–36.
14. Ouwerkerk W, Voors AA, Zwinderman AH. Factors influencing the predictive power of models for predicting mortality and/or heart failure hospitalization in patients with heart failure. *JACC Heart Fail* 2014;**2**:429–436.
15. Lanfear DE, Levy WC, Stehlik J, Estep JD, Rogers JG, Shah KB, Boyle AJ, Chuang J, Farrar DJ, Starling RC. Accuracy of Seattle Heart Failure Model and HeartMate II risk score in non-inotrope-dependent advanced heart failure patients: insights from the ROADMAP study (Risk Assessment and Comparative Effectiveness of Left Ventricular Assist Device and Medical Management in Ambulatory Heart Failure Patients). *Circ Heart Fail* 2017;**10**:e003745.
16. Allen LA, Matlock DD, Shetterly SM, Xu S, Levy WC, Portalupi LB, McIlvennan CK, Gurwitz JH, Johnson ES, Smith DH, Magid DJ. Use of risk models to predict death in the next year among individual ambulatory patients with heart failure. *JAMA Cardiol* 2017;**2**:435–441.
17. Freitas P, Aguiar C, Ferreira A, Tralhao A, Ventosa A, Mendes M. Comparative analysis of four scores to stratify patients with heart failure and reduced ejection fraction. *Am J Cardiol* 2017;**120**:443–449.
18. Nguyen LS, Coutance G, Ouldamar S, Zahr N, Brechot N, Galeone A, Bougle A, Lebreton G, Leprince P, Varnous S. Performance of existing risk scores around heart transplantation: validation study in a 4-year cohort. *Transpl Int* 2018;**31**:520–530.
19. Canepa M, Fonseca C, Chioncel O, Laroche C, Crespo-Leiro MG, Coats AJ, Mebazaa A, Piepoli MF, Tavazzi L, Maggioni AP; ESC HF Long Term Registry Investigators. Performance of prognostic risk scores in chronic heart failure

patients enrolled in the European Society of Cardiology Heart Failure Long-Term Registry. *JACC Heart Fail* 2018;**6**:452–462.
20. CMS Collaboration. Search for top-squark pair production in the sinflelepton final state in pp collision at sqrt(s)=8 TeV. *Eur Phys J C* 2013;**73**: 2677.
21. Rickham PP. Human experimentation. Code of Ethics of the World Medical Association. Declaration of Helsinki. *Br Med J* 1964;**2**:177.
22. Voors AA, Anker SD, Cleland JG, Dickstein K, Filippatos G, van der Harst P, Hillege HL, Lang CC, ter Maaten JM, Ng L, Ponikowski P, Samani NJ, van Veldhuisen DJ, Zannad F, Zwinderman AH, Metra M. A systems BIOlogy Study to TAilored Treatment in Chronic Heart Failure: rationale, design, and baseline characteristics of BIOSTAT-CHF. *Eur J Heart Fail* 2016;**18**: 716–726.
23. Roe BP, Yang HJ, Zhu J, Liu Y, Stancu I, McGregor G. Boosted decision trees as an alternative to artificial neural networks for particle identification. *Nucl Instrum Meth A* 2005;**543**:577–584.
24. Freund Y, Schapire RE. A Decision-theoretic generalization of on-line learning and an application to boosting. *J Comput Syst Sci* 1997;**55**:119–139.
25. Höcker A, Speckmayer P, Stelzer J, Tegenfeldt F, Voss H, Voss K, Christov A, Henrot-Versille S, Jachowski M, Krasznahorkay A Jr, Mahalalel Y, Ospanov R, Prudent X, Wolter M, Zemla A. TMVA - Toolkit for Multivariate Data Analysis. CERN-OPEN-2007-007. arXiv:physics/0703039v5.
26. Antcheva I, Ballintijn M, Bellenot B, Biskup M, Brun R, Buncic N, Canal P, Casadei D, Couet O, Fine V, Franco L, Ganis G, Gheata A, Maline DG, Goto M, Iwaszkiewicz J, Kreshuk A, Segura DM, Maunder R, Moneta L, Naumann A, Offermann E, Onuchin V, Panacek S, Rademakers F, Russo P, Tadel M. ROOT — A C++ framework for petabyte data storage, statistical analysis and visualization. *Computer Physics Communications* 2011;**182**:1384–1385.
27. Roger VL, Weston SA, Redfield MM, Hellermann-Homan JP, Killian J, Yawn BP, Jacobsen SJ. Trends in heart failure incidence and survival in a community-based population. *JAMA* 2004;**292**:344–350.
28. Chen J, Normand SL, Wang Y, Krumholz HM. National and regional trends in heart failure hospitalization and mortality rates for Medicare beneficiaries, 1998–2008. *JAMA* 2011;**306**:1669–1678.
29. Ni H, Xu J. Recent trends in heart failure-related mortality: United States, 2000–2014. *NCHS Data Brief* 2015;**231**:1–8.
30. Dokainish H, Teo K, Zhu J, Roy A, AlHabib KF, ElSayed A, Palileo-Villaneuva L, Lopez-Jaramillo P, Karaye K, Yusoff K, Orlandini A, Sliwa K, Mondo C, Lanas F, Prabhakaran D, Badr A, Elmaghawry M, Damasceno A, Tibazarwa K, Belley-Cote E, Balasubramanian K, Islam S, Yacoub MH, Huffman MD, Harkness K, Grinvalds A, McKelvie R, Bangdiwala SI, Yusuf S; INTER-CHF Investigators. Global mortality variations in patients with heart failure: results from the International Congestive Heart Failure (INTER-CHF) prospective cohort study. *Lancet Glob Health* 2017;**5**:e665–e672.