Machine Learning Applications in Heart Failure Disease Management: Hype or Hope?

Barry Greenberg, MD^{1,*}

Alison Brann, MD¹

Claudio Campagnari, PhD²

Eric Adler, MD¹

Avi Yagil, PhD^{1,3}

Address

^{*1}Cardiology Department, Sulpizio Cardiovascular Center, University of California, San Diego Medical Center, 9454 Medical Center Drive, La Jolla, CA, 92037-7411, USA

Email: bgreenberg@health.ucsd.edu

²Department of Physics, University of California, Santa Barbara, CA, 92037-7411, USA

³Department of Physics, University of California, San Diego, CA, 92037-7411, USA

Published online: 5 May 2021

© The Author(s) 2021

This article is part of the Topical Collection on *Heart Failure*

Keywords Heart failure · Machine learning · Risk prediction

Abstract

Purpose of the review Machine learning (ML) approaches have emerged as powerful tools in medicine. This review focuses on the use ML to assess risk of events in patients with heart failure (HF). It provides an overview of the ML process, challenges in developing risk scores, and strategies to mitigate problems.

Recent findings Risk scores developed using standard statistical methods have limited accuracy, particularly when they are applied to populations other than the one in which they were developed. Computerized ML algorithms which identify correlations between descriptive variables in complex, non-linear, multi-dimensional systems provide an alternative approach to predicting risk of events. The MARKER-HF mortality risk score was developed using data from the electronic health record of HF patients followed at a large academic medical center. The risk score, which uses eight commonly available variables, proved to be highly accurate in predicting mortality across the spectrum of risk. It retained accuracy in independent populations and was superior to other risk scores.

Summary Machine learning approaches can be used to develop risk scores that are superior to ones based on standard statistical methods. Careful attention to detail in curating data, selecting covariates, and trouble-shooting the process is required to optimize results.

Introduction: what is machine learning?

Machine learning (ML) is a subset of artificial intelligence that is defined as the study of computer algorithms that improve automatically through experience. These algorithms are built on sample training data with the goal of making predictions or decisions without being explicitly programmed for this purpose. Machine learning enables computerized analysis of large amounts of data over

relatively short periods of time and can be used to address questions in a variety of fields including medicine. There are several different approaches to ML. These include unsupervised learning which uses only input data to find grouping between the data points and supervised learning that builds models from training data sets containing both inputs and desired outputs.

Application of machine learning to heart failure treatment

Over the past several years, ML approaches have been used in a number of different areas of medicine including cardiology and is frequently used to identify correlations in large data sets, as in studies of the genome [1], proteome and metabolome. Applications currently being explored in HF include analysis of images from tests such as the electrocardiogram, echocardiogram, cardiac magnetic resonance imaging, and other scans that are commonly obtained as part of the evaluation of patients with HF and identification of subgroups of patients with HF who have common characteristics (pheno-grouping) that might provide novel mechanistic insights or predict response to specific therapies. The ability of ML to identify correlations between descriptive variables in complex, non-linear, multi-dimensional systems is also being investigated as a means of predicting risk of the future development of HF in a population and in identifying risk of events in patients with established disease. In this review, we focus on the use of ML as a means of developing tools for predicting risk in patients with HF.

Use of machine learning in detecting risk

Modern ML algorithms for classification are powerful tools that can be used to build clinical risk scores. These algorithms are “trained” to distinguish (“classify”) two or more cohorts of patients with or without known outcomes, such as death or HF hospitalization based on a set of clinical covariates. The ML process helps identify unsuspected associations among the covariates fed into the computer which cumulatively relate to the patients risk of a future event. Once the algorithm is trained, it can be used to assess the risk of an event occurring in an individual patient. The strength of ML is that it can handle complex non-linear correlations between the covariates that are entered into the computer. Figure 1 [2] provides a

simple example. Although the choice of covariates should be guided by clinical and other considerations, a deep understanding of their correlations is not necessary, since many ML algorithms handle them automatically.

Challenges in developing a machine learning algorithm

In practice risk scores are most often designed based on retrospective studies. The data used to generate these scores can originate from a clinical trial, patient registry, or an administrative data base. More recently, use of data from hospitals and healthcare systems has been made possible by the adoption of electronic health record (EHR) systems by institutions around the world. The use of EHR data bases offers several advantages including availability of large amounts of raw data, better representation of the spectrum of HF than in the other settings, and the ability to update scores as the study population grows or when new tests are integrated into medical practice. However, in addition to the obvious issues of de-identification and privacy inherent in using data from EHRs, several other challenges arise in extracting useful data from such systems. These include time variability in the data format/content due to evolutions in the EHR system, missing, incomplete, or poorly formatted data, confusion with diagnostic and procedure codes, duplications, mistakes, etc. While much information in a patients EHR already exists in numerical form that lends itself to computer evaluation, other patient descriptors and characteristics exist only in narrative form. While non-numeric variables including images can be translated into numeric form, this process adds complexity to the process in that it involves integration of steps that allow the capture of information in an appropriate format into the algorithm [3]. Regardless of how it is derived, validation of the data extraction procedure is essential. This can be achieved by proactively searching the dataset for nonsensical features, e.g., lab data supposedly taken months after the patient's death and extremely abnormal clinical values, as well as by randomly selecting a subset of patients and comparing "by hand" the extracted data with the contents of their EHR (chart review).

Fig. 1. Simple visualization how a machine learning algorithm is able to learn how to distinguish between two populations. Two computer-generated populations (a) red and (b) blue, characterized by a pair of covariates (X,Y). A Neural Network (NN) is trained to distinguish between the two. For each (X,Y) pair, the NN provides a score that can be used in this or a different population to give a red vs. blue probability. In panel (c) the output of the NN is visualized (in a third dimension - color) as a function of (X,Y) from deep red (very likely red) to deep blue (very likely blue). As additional covariates are added. Separation of the populations increases.

Approaches to missing data

One of the major limitations in constructing risk scores from the EHR is partially missing patient data. Missing data can be handled by “imputation”; e.g., if diastolic blood pressure (BPDias) is one of the covariates and BPDias data is missing for a given patient, the simplest form of imputation is to assign the patient the mean or the median of BPDias in the cohort. There are more sophisticated techniques for imputation but the obvious disadvantage of all of these procedures is that information is lost and correlations are broken. Alternatively, patients with missing data could be excluded from the study; however, this approach raises other concerns that must be addressed. For example, if missing data are not randomly distributed across the cohort, this procedure can introduce a selection bias. In such cases, it is important to investigate possible biases generated by exclusion of patients with missing data and attempt to mitigate them. Special attention needs to be paid to in-patients vs out-patients, if both are included in the cohort, as out-patients are more likely to have more issues with time-consistency of the covariates used (see section [“Characterizing the State of the Patient at a Prediction Time”](#)).

Exclusion of patients with missing data is most important in the training phase of a ML algorithm, where capturing the correlations between the covariates and the outcomes is “learned” by the computer. It is less so, however, when the model that has been generated is applied to calculating risk for a given patient. At this stage, the loss of a single covariate out of a larger number may slightly reduce but does not destroy the predictive power of the risk score. It is possible to provide specific calibrations for scores calculated with an imputed variable.

Characterizing the state of the patient at a prediction time

As the clinical state of a patient evolves over time, it is important that the covariates used in the creation of the score be measured in tight temporal proximity to each other so that the correlations used to collectively capture the state of the patient at prediction time are preserved. If these correlations are lost or “smeared” by mixing covariates collected at various points over an extended time frame, the ability to define a patient’s state at a precise moment is weakened, reducing the ability to relate that state to the outcomes. Clearly, not all covariates are equally time-sensitive to the same extent and clinical judgment of the necessary time-window is important. For instance, use of left ventricular ejection fraction collected within several weeks or even months from the point in time that risk is being calculated may be acceptable in a patient who has had a stable clinical course, whereas variables such as serum creatinine level or arterial blood gas saturation may be needed to be collected in much closer temporal proximity to the prediction time point.

Overtraining

In most cases, the size of the cohorts on which risk scores can be developed is limited to the hundreds or the thousands, rather than the millions. This usually limits the complexity of the underlying algorithm that can be safely deployed; e.g., “deep” neural networks typically need large sample sizes to avoid overfitting (aka over-training) and hence risk irreproducibility. In general, if the number of covariates is too large with respect to the training data set size, the algorithm is at risk of being “over-trained.” An over-trained algorithm is one that has too much freedom to focus on the detailed properties of the particular

cohort on which it is trained, so that its performance does not generalize. This leads to inferior performance of the score when it is applied to populations other than the one from which it was derived. There are several techniques to recognize and avoid or minimize over-training [4]. The simplest one is to randomly split the cohort into two statistically independent sets: a “training” cohort, and a “testing” cohort. The ML algorithm is first trained, i.e., developed, on the training cohort only. Then its performance is evaluated on and compared between the training and testing cohorts. Statistically significant differences in performance are a sign of over-training.

Machine learning algorithms

There are many different ML algorithms [5–7] that can be used, e.g., neural networks, decision trees, Bayesian methods, and support vector machines, each with several configurable parameters. Picking among different choices is not obvious, and some trial and error is necessary. Fortunately, software codes to implement different possibilities are easily available.

Selection of covariates

The choice of the set of covariates on which to base a risk score is guided by clinical judgment, practical considerations, e.g., what covariates are generally available, and avoidance of over-training. Unlike risk scores that are developed using standard approaches in which variables are selected according to their individual predictive power, covariates that are selected by the computer in ML approaches may not have predictive value by themselves. Testing performance as covariates are added or removed from the training set is also a powerful tool. An important reason to limit the number of covariates is that if a score is to be useful in clinical settings, it needs to be based on widely available and easily accessible data and it should be easy to calculate. Although the addition of a particular variable may improve predictive accuracy of a ML generated risk score, this may represent a pyrrhic victory if that variable is costly or not readily accessible to providers. As discussed in the following section, we have recently developed a score for predicting mortality risk in patients with HF that is based on just eight covariates (diastolic blood pressure and seven covariates from CBC (complete blood count) and CMP (comprehensive metabolic panel)), all of which are readily available in patients with HF.

External validation and bias-testing

Finally, biases due the properties of the cohort on which the risk score is built are probably inevitable. Validations in external cohorts, e.g., from a different medical center or from a different country, are also crucial. A limited choice of widely available covariates also helps in making external validation possible. In addition, examining the performance of the score, stratified by other properties such as age gender, ethnicity, in-patients vs out-patients, is very important, particularly when these data are not included as covariates.

Application of machine learning in predicting risk in heart failure

Importance of knowing risk for future events

The ability to accurately predict risk of clinical events in patients with HF provides important information to providers and patients alike. Determination

of the risk of an event helps identify patients who are most (or least) likely to derive benefit from targeted interventions for preventing future events. Assessment of a patient's risk for these events is particularly valuable when the proposed treatment option is costly or associated with serious side effects. Knowledge of a patient's risk for future events when integrated along with clinical, social, and other information provides a framework for therapeutic recommendations by providers and is helpful to patients and their families in planning for the future.

Approaches to predicting risk

Risk of future events can be determined in a variety of ways including assessment of patient characteristics, clinical status, test results, or biomarkers. Risk scores which combine information from a variety of the above domains using standard statistical analysis methods have also been developed to predict risk in patients with HF [8–13]. These risk scores, however, demonstrate only modest predictive power, especially when applied to populations outside the one used for their derivation [14–17]. A recent analysis of the predictive accuracy of several commonly used risk scores including the CHARM (Candesartan in Heart Failure-Assessment of Reduction in Mortality), GISSI-HF (Gruppo Italiano per lo Studio della Streptochinasi nell'Infarto Miocardico-Heart Failure), MAGGIC (Meta-analysis Global Group in Chronic Heart Failure), and SHFM (Seattle Heart Failure Model) scores in a large European registry of patients with HF indicated only modest accuracy of all of these scores in predicting risk of death at 1 and 2 years [18]. The authors of this analysis concluded that the limited accuracy of these scores made physicians reluctant to use them in clinical practice and that more precise predictive tools are needed.

Machine learning for predicting incident HF risk

Machine learning techniques have been used to predict cardiovascular (CV) events, including the risk of incident HF. In a study of patients in the Multi-Ethnic Study of Atherosclerosis (MESA) cohort, Ambale-Venkatesh et al. used random survival forest (RSF) to predict six CV outcomes including new onset HF [19]. They identified the top-20 variables which were predictive of each outcome from a total 735 variables derived from imaging studies, health questionnaires, and biomarker analysis. While the model predicted incident HF with an AUC of 0.84, this was only a modest improvement compared to the MESA-HF score (AUC of 0.8). A limitation of this risk score is that the top predictive variables included several laboratory values which are not widely available, e.g., tissue necrosis factor alpha soluble receptor and interleukin-2 soluble receptor. This model has not yet been validated in external populations and its generalizability is questionable. Segar et al. used a similar RSF technique to identify predictors of incident HF among patients with type 2 diabetes mellitus (DM) in the Action to Control Cardiovascular Risk in Diabetes (ACCORD) trial [20]. The RSF models demonstrated somewhat better discrimination than the Cox-based method (C-index 0.77 [95% CI 0.75–0.80] vs. 0.73 [0.70–0.76], respectively). From the identified predictors, an integer-based risk score for 5-year HF incidence was created, the WATCH-DM (Weight [BMI], Age, hyperTension, Creatinine, HDL-C, Diabetes control [fasting plasma glucose], QRS Duration, MI, and CABG) risk score. Each 1 unit increment in the risk score

is associated with a 24% higher relative risk of HF within 5 years. The WATCH-DM risk score can easily be applied in clinical practice. Both the risk score and the RSF-based risk prediction model were externally validated in a cohort of individuals with DM using the Antihypertensive and Lipid-Lowering Treatment to Prevent Heart Attack Trial (ALLHAT) [21].

Machine Learning for Predicting Future Events in Patients with Heart Failure

Researchers have also applied ML to predict morbidity and mortality in patients with known HF (Table 1). Although numerous models have been created to predict HF readmissions, most have demonstrated only limited discriminative properties. Frizzell et al. used several ML algorithms to predict all-cause readmissions 30 days after discharge from a HF hospitalization in patients included in the GWTG-HF registry [22]. All of the models developed in this study showed modest discriminatory power, with C statistics consistently around 0.62. Awan et al. demonstrated a similar AUC of 0.62 using a multi-layer perceptron-based approach to predict risk of 30 day HF readmission or death in a population of patients above age 65 years admitted with HF [25]. Golas et al. were able to demonstrate modest improvement in risk prediction using several deep learning algorithms in a population of HF patients admitted within a large healthcare system [24]. The model developed using deep unified networks from >3500 variables from the electronic health record (EHR) demonstrated the best performance with an AUC of 0.705 for prediction of 30-day readmission. To our knowledge, these models have not been externally validated nor directly compared to traditional risk assessment tools, so it is not known if they can be applied to broad HF populations.

Machine learning models have overall been more successful at predicting mortality in patients with HF. In 2018, Ahmad et al. applied random forest modeling to identify predictors of 1-year survival in patients enrolled in the Swedish HF Registry [28]. Their model demonstrated excellent discrimination for survival with a C-statistic of 0.83. Additionally, they used cluster analysis to successfully identify 4 clinically relevant subgroups of HF with marked differences in 1-year survival and response to therapeutics. Samad et al. used ML learning techniques to incorporate a panel of 57 echocardiographic measurements with clinical variables to predict 5-year mortality in all patients undergoing echocardiography, including 15,492 patients with HF. For the patients with HF, the random forest model that was developed resulted in an AUC of 0.8, a significant improvement compared to the Seattle HF Model (AUC 0.63) [9, 30]. Using a deep learning algorithm, Kwon et al. have also demonstrated superior predictive abilities compared to traditional risk scores for predicting in-hospital, 1-year and 3-year mortality for patients with acute HF [29•]. For in-hospital mortality, their ML algorithm demonstrated an excellent AUC of 0.88, compared to the AUC of 0.728 from the GWTG-HF score [10]. They were also able to successfully predict long-term mortality, with an AUC of 0.782 for 1-year mortality and AUC 0.813 for 3-year mortality. This algorithm outperformed the MAGGIC score which demonstrated AUCs of 0.718 and 0.729 for 1- and 3-year mortality, respectively [13]. While these models have demonstrated success at risk prediction, they have not been externally validated and have not yet been widely incorporated into clinical care.

Jing et al. were recently able to demonstrate that in addition to risk stratify patients, ML can be used to identify individuals most likely to benefit from

Table 1. Characteristics and outcomes in selected studies using machine learning to predict mortality and hospitalization in patients with heart failure

Study	Year	Population Size	Endpoints	Methods	Externally Validated	AUC	Comparison with other predictive tools
Frizzell et al. [22]	2017	56,477 from GWTG-HF registry	30-day Readmissions	Several ML algorithms	No	0.62 (TAN), 0.63 (LR), 0.62 (LASSO)	Superior discrimination compared to model from Eapen et al. [23], AUC 0.63 (LR) vs 0.59
Golas [24]	2018	11,510 patients admitted w/ HF	30-day Readmissions	DL algorithm	No	0.664 (LR), 0.65 (GBM), 0.695 (MN), 0.705 (DUNs)	No
Awan et al. [25]	2019	10,757 patients admitted w/ HF	30-day HF readmission or death	MLP-based approach	No	0.62 (MLP) 0.58 (LR)	No
Angraal et al. [26•]	2020	1767 patients with HFpEF	Mortality and HF hospitalization	RF	No	0.72 for mortality, 0.76 for HF hospitalization	No
Adler et al. [27••]	2020	1986 HF patients followed in a hospital system	1-year mortality	BDT algorithm	Yes	0.88	Outperformed the IMRS (AUC 0.78) [11], GWTG-HF (AUC 0.74) [10], and ADHERE scores (AUC 0.63) [8]
Ahmad et al. [28]	2018	44,886 HF patients from Swedish HF Registry	1-year mortality	RF	No	0.83	No
Kwon et al. [29•]	2019	2165 patients with acute HF	In hospital, 1-year and 3-year mortality	DL algorithm	No	0.88 (in-hospital) 0.78 (1-year) 0.81 (3-year)	Outperformed the GWTG-HF [10] and MAGGIC [12] scores for all outcomes
Samad [30]	2018	171,510 patients undergoing echocardiography	5-year mortality	Several ML algorithms	No	0.80 (RF)	Outperformed the SHFM [12], AUC 0.80 vs 0.63

Abbreviations: *GWTG-HF*, Get With the Guidelines Heart Failure; *ML*, machine learning; *TAN*, tree augmented naïve Bayesian network; *LR*, logistic regression; *LASSO*, least absolute shrinkage and selection operator; *DL*, deep learning; *GB*, gradient boosting; *MN*, maxout network; *DUN*, deep unified network; *MLP*, multi-layer perceptron; *MAGGIC*, Meta-Analysis Global Group in Chronic Heart Failure; *SHFM*, Seattle Heart Failure Model; *BDT*, boosted decision tree; *ADHERE*, Acute Decompensated Heart Failure Registry

interventions with evidence-based therapies [31]. Using data from the EHR, they trained ML models to predict 1-year all-cause mortality in 26,971 patients with HF. Their model included clinical variables, diagnostic codes, electrocardiogram and echocardiographic measurements, and 8 evidence-based “care gaps”: flu vaccine, blood pressure of <130/80 mmHg, A1c of <8%, cardiac resynchronization therapy, and active medications (angiotensin-converting enzyme inhibitor/angiotensin II receptor blocker/angiotensin receptor-neprilysin inhibitor, aldosterone receptor antagonist, hydralazine, and beta-blocker). Their best performing model demonstrated an AUC of 0.77 for 1-year all-cause mortality which was superior to the performance of the Seattle HF Model (AUC 0.57) [9]. Of the 13,238 living patients, 2844 (21.5%) patients were predicted to die within 1 year based on the estimated mortality rate. Simulating closure of the 8 care gaps that existed in these patients resulted in a 1.7% improvement in the absolute mortality rate with 231 additional patients predicted to survive beyond 1 year. These findings require prospective evaluation and external validation, but they are promising and highlight the potential utility of ML methods to guide clinical action and to identify patients most likely to benefit from optimization of evidence-based therapies.

Derivation and validation of MARKER-HF

The Machine learning Assessment of Risk and Early mortality in Heart Failure (MARKER-HF) risk score for predicting mortality in patients with HF was developed using non-parametric analysis methods to incorporate interactions between variables that have prognostic value [32]. It was derived and internally validated in patients who were identified at the time of the first mention of a HF diagnosis in either the in-patient or out-patient setting in their University of California, San Diego Healthcare System EHR. Patients who met entry criteria were divided into training and validation cohorts. MARKER-HF was constructed in the training cohort using a boosted decision tree model to discriminate between patients at the extrema of risk of death with patients who died within 90 days considered “high risk” while those known to be alive after 800 days designated as “low risk.” Data from complete blood count, comprehensive metabolic panel, vital sign measurement, electrocardiogram, and echocardiogram, all performed within 7 days of the patients’ index HF event, was used to construct the model. The number of variables selected was determined by balancing inclusiveness so that patients with missing data would not be lost and the need to minimize overfitting that could result in over-training and loss of accuracy of the score in other populations. Ultimately, a composite of eight variables (diastolic blood pressure, creatinine, blood urea nitrogen, hemoglobin, white blood cell count, platelets, albumin, and red blood cell distribution width) that discriminated with a high degree of accuracy between patients at low and high risk was identified. Predictive accuracy was demonstrated across the entire spectrum of risk and was confirmed in the validation cohort. Comparison of MARKER-HF predicted survival in the training and validation cohorts is depicted in Fig. 2, Panel A. No evidence of over-training was detected.

The ability of a risk score to maintain its accuracy in clinically relevant subgroups of the HF population and in other independent populations is essential if it is to be widely used in clinical practice. Consequently, we assessed the predictive accuracy of MARKER-HF in patients according to sex, race, in-

patient vs out-patient status and acuity at the time of identification (determined by whether the index diagnosis was HF or pulmonary edema) and found it to perform equally well in all subgroups. External validation demonstrated that MARKER-HF maintained its predictive accuracy in populations followed in the University of California, San Francisco (UCSF) Healthcare System and the European based BIOSTAT-CHF Registry [32] (Fig. 2, Panel B).

Finally, to determine if MARKER-HF was superior to other predictors of risk we compared its performance to NT-proBNP, a well-validated HF biomarker, and to other risk scores used to predict mortality. While MARKER-HF scores tracked reasonably well with levels of NT-proBNP, it proved a much more reliable predictor of mortality than did the biomarker which had an AUC of 0.69 as shown in Fig. 2, Panel C. Initially, NT-proBNP was excluded as a covariate during the derivation of MARKER-HF due to its low availability (~50%). When it was added to the other eight variables in the MARKER-HF score, predictive accuracy was not significantly improved. MARKER-HF also proved to be superior to either the Intermountain Risk Score (IMRS) [7], the Get With the Guidelines-HF (GWTG-HF) score [11], or the Acute Decompensated Heart Failure Registry (ADHERE) risk score [8] in predicting mortality risk in the UCSD, UCSF, and BIOSTAT-CHF populations.

Future directions

By virtue of their ability to assess large data bases and find unsuspected associations between covariates, ML approaches offer a powerful new approach for risk assessment in patients with HF. The ability to accurately predict risk of future events is critical for providers in making recommendations about the advisability of specific therapies and for patients and their families who must plan for the future. In this review, we have outlined criteria for extracting and managing data that we believe are essential in developing ML based risk scores. In all cases, external validation in independent populations is needed to determine generalizability and comparison with available tools is required in order to know if they are superior to other approaches. For risk scores to be effective, they should

Fig. 2. Panel A. Comparison of 1-year Survival probabilities as a function of MARKER-HF for Training and Validation cohorts in the UCSD cohort. The values obtained in the Training and Validation cohorts do not differ significantly, indicating no significant overtraining. Panel B. The ROC curves for MARKER-HF in the University of California, San Diego (UCSD), University of California, San Francisco (UCSF) and BIOlogy Study to Tailored Treatment in Chronic Heart Failure (BIOSTAT-CHF) populations. The C-statistics obtained are consistent with each other within their 95% confidence intervals. Panel C. ROC Curves for NT-proBNP and MARKER-HF. MARKER-HF demonstrates superior prediction compared to the biomarker.

require input of a manageable number of variables that are easily accessible and widely collected in routine patient care, so as not to constrain their use.

Finally, although this review has focused on the use of ML to develop hospitalization and mortality risk scores for clinical events in patients with HF, they can be applied to a variety of situations in either this or other populations. There is great need for novel approaches for calculating risk of other adverse events known to occur in patients with HF such as stroke, atrial fibrillation and sudden cardiac death as well as the risk of adverse consequences of specific therapies designed to treat these conditions. Machine learning approaches may also be useful in future clinical trials in HF by helping to determine which patients to enroll.

Compliance with Ethical Standards

Human and Animal Rights and Informed Consent

This article does not contain any studies with human or animal subjects performed by any of the authors.

Conflict of Interest

Barry Greenberg, Alison Brann, Claudio Campagnari, Eric Adler, and Avi Yagil declare that they have no conflict of interest.

Open Access

This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References and Recommended Reading

Papers of particular interest, published recently, have been highlighted as:

- Of importance
 - Of major importance
1. Minoche AE, Horvat C, Johnson R, Gayevskiy V, Morton SU, Drew AP, et al. Genome sequencing as a first-line genetic test in familial dilated cardiomyopathy. *Genet Med*. 2019;21:650–62.
 2. Inspired by https://scikit-learn.org/stable/auto_examples/classification/plot_classifier_comparison.html
 3. Weng WH, Waghlikar KB, McCray AT, Szolovits P, Chueh HC. Medical subdomain classification of clinical notes using a machine learning-based natural language processing approach. *BMC Med Inform Decis Mak*. 2017;17:155.
 4. Xu Y, Goodacre R. On splitting training and validation set: a comparative study of cross-validation, bootstrap and systematic sampling for estimating the generalization performance of supervised learning. *J Anal Test*. 2018;2:249–62.
 5. Scikit-learn: Machine Learning in Python, Pedregosa et al., *Journal. of Machine Learning Research* 12, pp. 2825–2830, 2011.

6. Lang M, Binder M, Richter J, Schratz P, Pfisterer F, Coors S, et al. mlr3: a modern object-oriented machine learning framework in R. *Journal of Open Source Software*. 2019. <https://doi.org/10.21105/joss.01903>.
 7. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. New York, NY, USA: ACM; 2016. p. 785–94.
 8. Fonarow GC, Adams KF, Abraham WT, Yancy CW, Boscardin WJ. Risk stratification for in-hospital mortality in acutely decompensated heart failure: classification and regression tree analysis. *J Am Med Assoc*. 2005;293(5):572–80.
 9. Levy WC, Mozaffarian D, Linker DT, Sutradhar SC, Anker SD, Cropp AB, et al. The Seattle heart failure model: prediction of survival in heart failure. *Circulation*. 2006;113(11):1424–33.
 10. Peterson PN, Rumsfeld JS, Liang L, Albert NM, Hernandez AF, Peterson ED, et al. 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(1):25–32.
 11. Horne BD, May HT, Kfoury AG, Renlund DG, Muhlestein JB, Lappé DL, et al. The Intermountain Risk Score (including the red cell distribution width) predicts heart failure and other morbidity endpoints. *Eur J Heart Fail*. 2010;12(11):1203–13.
 12. Pocock SJ, Ariti CA, McMurray JJV, Maggioni A, Køber L, Squire IB, et al. Predicting survival in heart failure: a risk score based on 39 372 patients from 30 studies. *Eur Heart J*. 2013;34(19):1404–13.
 13. Sartipy U, Dahlström 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(2):173–9.
 14. Ouwerkerk W, Voors AA, Zwinderman AH. Factors influencing the predictive power of models for predicting mortality and/or heart failure hospitalization inpatients with heart failure. *JACC Hear Fail*. 2014;2(5):429–36.
 15. Allen LA, Matlock DD, Shetterly SM, Xu S, Levy WC, Portalupi LB, et al. Use of risk models to predict death in the next year among individual ambulatory patients with heart failure. *JAMA Cardiol*. 2017;2(4):435–41.
 16. Freitas P, Aguiar C, Ferreira A, Tralhão 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(3):443–9.
 17. Lanfear DE, Levy WC, Stehlik J, et al. 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 Man. *Circ Hear Fail*. 2017;10(5).
 18. Canepa M, Fonseca C, Chioncel O, Laroche C, Crespo-Leiro MG, Coats AJS, et al. Performance of prognostic risk scores in chronic heart failure patients enrolled in the European Society of Cardiology Heart Failure Long-Term Registry. *JACC Hear Fail*. 2018;6(6):452–62.
 19. Ambale-Venkatesh B, Yang X, Wu CO, Liu K, Hundley WG, McClelland R, et al. Cardiovascular event prediction by machine learning: the multi-ethnic study of atherosclerosis. *Circ Res*. 2017;121(9):1092–101.
 20. Segar MW, Vaduganathan M, Patel KV, McGuire DK, Butler J, Fonarow GC, et al. Machine learning to predict the risk of incident heart failure hospitalization among patients with diabetes: the WATCH-DM risk score. *Diabetes Care*. 2019;42(12):2298–306.
 21. Furberg CD, Wright JT, Davis BR, et al. Major outcomes in high-risk hypertensive patients randomized to angiotensin-converting enzyme inhibitor or calcium channel blocker vs diuretic: the antihypertensive and lipid-lowering treatment to prevent heart attack trial (ALLHAT). *J Am Med Assoc*. 2002;288(23):2981–97.
 22. Frizzell JD, Liang L, Schulte PJ, Yancy CW, Heidenreich PA, Hernandez AF, et al. 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(2):204–9.
 23. Eapen ZJ, Liang L, Fonarow GC, Heidenreich PA, Curtis LH, Peterson ED, et al. Validated, electronic health record deployable prediction models for assessing patient risk of 30-day rehospitalization and mortality in older heart failure patients. *JACC Hear Fail*. 2013;1(3):245–51.
 24. Golas SB, Shibahara T, Agboola S, et al. A machine learning model to predict the risk of 30-day readmissions in patients with heart failure: a retrospective analysis of electronic medical records data. *BMC Med Inform Decis Mak*. 2018;18(1).
 25. Awan SE, Bennamoun M, Sohail F, Sanfilippo FM, Dwivedi G. Machine learning-based prediction of heart failure readmission or death: implications of choosing the right model and the right metrics. *ESC Hear Fail*. 2019;6(2):428–35.
 26. Angraal S, Mortazavi BJ, Gupta A, et al. Machine learning prediction of mortality and hospitalization in heart failure with preserved ejection fraction. *JACC Hear Fail*. 2020;8(1):12–2.
- Developed models for predicting mortality and heart failure hospitalization in patients with heart failure and preserved ejection fraction.**
27. Adler ED, Voors AA, Klein L, Macheret F, Braun OO, Urey MA, et al. Improving risk prediction in heart failure using machine learning. *Eur J Heart Fail*. 2020;22:139–4.
- This manuscript describes the generation and validation of a risk score for mortality in patients with heart failure based on extraction of data from the electronic health record. Criteria for data extraction and covariate selection are described.**

28. Ahmad T, Lund LH, Rao P, et al. Machine learning methods improve prognostication, identify clinically distinct phenotypes, and detect heterogeneity in response to therapy in a large cohort of heart failure patients. *J Am Heart Assoc.* 2018;7(8).
 29. • Kwon JM, Kim KH, Jeon KH, et al. Artificial intelligence algorithm for predicting mortality of patients with acute heart failure. *PLoS One.* 2019;14(7)
- Used machine learning approach to developing risk score for predicting outcome in patients hospitalized with acute heart failure.**
30. Samad MD, Ulloa A, Wehner GJ, Jing L, Hartzel D, Good CW, et al. Predicting survival from large echocardiography and electronic health record datasets: optimization with machine learning. *JACC Cardiovasc Imaging.* 2019;12(4):681–9.
 31. Jing L, Ulloa Cerna AE, Good CW, Sauers NM, Schneider G, Hartzel DN, et al. A machine learning approach to management of heart failure populations. *JACC Hear Fail.* 2020;8(7):578–87.
 32. Voors AA, Anker SD, Cleland JG, Dickstein K, Filippatos G, van der Harst P, et al. A systems BIOlogy Study to Tailored Treatment in Chronic Heart Failure: rationale, design, and baseline characteristics of BIOSTAT-CHF. *Eur J Heart Fail.* 2016;18:716–26.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.