International Journal of Cardiology 365 (2022) 78–84

Contents lists available at ScienceDirect

International Journal of Cardiology

journal homepage: www.elsevier.com/locate/ijcard

Short communication

Comparison of machine learning and the regression-based EHMRG model for predicting early mortality in acute heart failure

David E. Austin[a], Douglas S. Lee[a,b,c,1,*], Chloe X. Wang[a,f], Shihao Ma[a,d,e], Xuesong Wang[a], Joan Porter[a], Bo Wang[a,b,d,e,f,g,1]

[a] ICES, Institute for Clinical Evaluative Sciences, 2075 Bayview Ave, Toronto, ON, M4N3M5, Canada
[b] Peter Munk Cardiac Centre and Joint Department of Medical Imaging of University Health Network, 585 University Ave., Toronto, ON, M5G2N2, Canada
[c] Ted Rogers Centre for Heart Research, 661 University Ave., Toronto, ON M5G1X8, Canada
[d] Department of Computer Science, University of Toronto, 40 St. George St., Toronto, ON M5S2E4, Canada
[e] Vector Institute of Artificial Intelligence, 661 University Ave., Suite 710, Toronto, ON M5G1M1, Canada
[f] Division of Vascular Surgery, University Health Network, 190 Elizabeth St., Toronto, ON M5G2C4, Canada
[g] Department of Laboratory Medicine and Pathobiology, University of Toronto, Canada

ARTICLE INFO

Keywords:
Heart failure
Machine learning
Statistical models
Prognosis
Outcomes
Mortality

ABSTRACT

Background: Although risk stratification of patients with acute decompensated heart failure (HF) is important, it is unknown whether machine learning (ML) or conventional statistical models are optimal. We developed ML algorithms to predict 7-day and 30-day mortality in patients with acute HF and compared these with an existing logistic regression model at the same timepoints.
Methods: Patients presenting to one of 86 hospitals, who were either admitted to hospital or discharged home directly from the emergency department, were randomly selected using stratified random sampling. ML approaches, including neural networks, random forest, XGBoost, and the Lasso, were compared with a validated logistic regression model for discrimination and calibration.
Results: Among 12,608 patients in our analysis, lasso regression (c-statistic 0.774; 95% CI, 0.743, 0.806) performed better than other ML models for 7-day mortality but did not outperform the baseline logistic regression model (0.794; 95% CI, 0.789, 0.800). For 30-day mortality, XGBoost performed better than other ML models (c-statistic 0.759; 95% CI; 0.740, 0.779), but was not significantly better than logistic regression (c-statistic 0.755; 95% CI, 0.750, 0.762). Logistic regression demonstrated better calibration at 7 days (calibration-in-the-large 0.017; 95% CI, −0.657, 0.692, and calibration slope 0.954; 95% CI, 0.769, 1.139), and at 30 days (−0.026; 95% CI, −0.374, 0.322, and 0.964; 95% CI, 0.831, 1.098), and best Brier scores, compared to ML approaches.
Conclusions: Logistic regression was comparable to ML in discrimination, but was superior to ML algorithms in calibration overall. ML algorithms for prognosis should routinely report calibration metrics in addition to discrimination.

1. Background

Heart failure (HF) is a global public health issue which affects approximately 26 million people globally [1]. Many patients present to the emergency department acutely, where the initial diagnosis is often made, as the condition is frequently undiagnosed in the ambulatory office setting [2]. Acute HF is a syndrome defined by new or worsening signs and symptoms of HF, which are due to systemic congestion, and often leads to hospitalization [3]. However, once hospitalized, the 30-day mortality rate of patients with HF is as high as 10.6% at a population level, portending the poor prognosis of HF and underscoring the need for improved ways to estimate prognosis using risk stratification [4,5]. In earlier work, our group derived the Emergency Heart failure Mortality Risk Grade (EHMRG) 7-day and 30-day risk prediction algorithms, a multivariable risk scoring formula developed using multiple logistic regression [6–8]. Using only 10 predictor variables, the model

[*] Corresponding author at: ICES, Division of Cardiology, Peter Munk Cardiac Centre, University Health Network, Ted Rogers Chair in Heart Function Outcomes, Medicine, University of Toronto, Canada.
E-mail address: dlee@ices.on.ca (D.S. Lee).
[1] Equal contribution.

https://doi.org/10.1016/j.ijcard.2022.07.035
Received 21 January 2022; Received in revised form 26 June 2022; Accepted 17 July 2022
Available online 20 July 2022

can predict mortality outcomes at two time points, and identifies those who are low risk, with 0% mortality at 7- and 30-days in those in the lowest risk quintile [9].

While other predictive models have been derived for use in patients with HF in the emergent setting, few have been prospectively validated, and none are used routinely in clinical practice [10]. Only the EHMRG model has been compared with and found to be superior to physician estimated risk, with c-indices of 0.71 vs. 0.81 for prediction of 7-day mortality in favor of the mathematical model [9,11]. The EHMRG 7- and 30-day models are currently being evaluated in a randomized controlled trial, the Comparison of Outcomes and Access to Care for Heart failure (COACH) trial, whereas no other model has been tested as a randomized intervention [11]. There has been great interest in machine learning (ML) as an alternative analytic approach to development of risk prediction models. Harrell defines ML as 'an algorithmic procedure for prediction or classification that tends to be empirical, nonparametric, flexible, and does not capitalize on additivity of predictors'. [12] ML is considered to be advantageous because it does not make assumptions about data distribution, handles complex relationships between data with varying degrees of correlation, and is agnostic to a priori assumptions about the importance of specific variables. ML may also be useful when there are many variables to consider, when inputs consist of complex data or images, when there are potentially previously unrecognized interactions, and when the signal-to-noise ratio is high.

It is unknown whether ML approaches are superior to conventional statistical models (CSM). The advantages of CSM includes the ability to infer the relationship between the covariates and outcome and the interpretability of the associations, which are more difficult to discern using the so-called 'black box' of ML. In two prior systematic reviews comparing ML and CSM, we found that ML had, on average, slightly higher c-statistics than CSM, however few studies compared ML to extensively validated risk algorithms that were derived using rigorous statistical and methodological approaches [13,14]. Therefore, in this study, we developed ML algorithms for 7-day and 30-day mortality among patients with acute HF and compared these with CSM-derived EHMRG models.

2. Methods

Study setting and data sources. We studied patients ≥18 years (Ontario, Canada) who presented to an emergency department with acute HF, randomly selected from 86 hospitals using stratified random sampling from the Emergency Heart Failure Mortality Risk Grade (EHMRG) and Enhanced Feedback for Effective Cardiac Treatment (EFFECT) Phase II chart review cohorts as detailed previously [6,15]. To be eligible for the EHMRG and EFFECT studies, patients were required to have a clinical diagnosis of HF by meeting the Framingham criteria, and most responsible diagnosis of HF in the discharge abstract from hospital (Canadian Institute for Health Information Discharge Abstract Database and the National Ambulatory Care Reporting System) based on the totality of information available during the hospital stay. The EHMRG cohort was comprised of patients presenting to the ED with HF and discharged home, while EFFECT II was comprised of hospitalized patients with HF. Those who were palliative prior to ED arrival, dialysis-dependent, and transfers from other facilities were excluded from both cohorts. In this study, we combined EHMRG and EFFECT II datasets, and considered 111 candidate variables in the following groupings: demographic, presentation details, vital signs, laboratory results, medical history, and pre-admission medication (see Supplementary table 1). We linked these cohorts to clinical administrative datasets using unique encoded identifiers to obtain information on hospital visits in the previous 2 years to supplement the clinical variables that were used to develop the original EHMRG risk model. These additional variables, which were indicative of past hospital visits included: the number of previous hospital admissions, number of previous ED visits leading to or not leading to hospitalization, total number of previous ED visits, time since first ED visit leading to hospitalization, time since first ED or most recent ED visit that did or did not lead to hospitalization, time since first hospitalization or most recent hospitalization (see Supplementary table 1). These datasets were linked using unique encoded identifiers and analyzed at ICES.

Analysis. We compared the following methods: a) logistic regression using the 10 EHMRG risk model covariates, b) neural networks, c) random forest, d) XGBoost, and e) Lasso logistic regressions. Methods (b) to (e) considered all 111 candidate variables in the training set. The EHMRG model covariates are available on the web (https://ehmrg.ices.on.ca) and both 7-day and 30-day models have been previously published [6,7]. We provide a brief explanation of these methods. The lasso is a form of logistic regression such that the coefficients are estimated using a cost function and where some of the coefficients will be set to zero, resulting in a more parsimonious model than conventional logistic regression. This helps to improve interpretability, removes variables that are only weakly associated with the outcome, thereby performing variable selection. Random forests and XGBoost are based on decision trees [16]. Decision trees group patients based on a series of binary splits. The split chosen at each stage is the one that best separates the patients by the outcome. Random forests fit a large sequence of decision trees, each of which is trained on a different bootstrap sample from the data. At each split, the tree can only choose from a random subset of the variables. The random forest's prediction is the average of predictions across all the trees [16]. XGBoost is similar to random forests, but trees are fit sequentially, such that each tree is grown to minimize the errors of previous trees [17]. Neural networks are machine learning algorithms based on the architecture of the human brain, with layers of neurons connected by edges. As the neural network is exposed to training data, it refines the connections between neurons to reduce prediction error [18]. We included neural networks in order to consider a commonly-used deep learning algorithm that was not tree-based. These methods were chosen because they are common in the cardiology literature as demonstrated in two prior systematic reviews comparing machine learning and conventional statistical analysis [13,14]. Interested readers are referred to recently published reviews and applications of machine learning in medicine and clinical investigation for further details [19–22].

We used recursive feature elimination to select features for XGBoost and random forest models [23]. For neural networks, feature selection involved combining the feature importance results from an elastic net model and a random forest model. For each algorithm, we trained separate models for 7-day and 30-day mortality. To ensure that the ML models were derived and tested using a similar split to that employed in the derivation of the original EHMRG models, we randomly divided the combined EHMRG/EFFECT II sample in a 2:1 ratio into derivation and test samples [6], as has been suggested previously [24]. The ML derivation sample was split 80–20% into training and validation sets, which is a standard approach to model-building using ML [21]. We calculated the anticipated precision of the estimated c-statistic resulting from our derivation-test sample division using previously-published methods [25], given an anticipated event rate of 2% for 7-day mortality and 8% for 30-day mortality, and an anticipated c-statistic of 0.75. Our split of the dataset produced a test sample that allowed us to estimate the c-statistic for 7-day and 30-day mortality with standard errors of approximately 0.027 and 0.014, respectively. Using one-third of the dataset as a test sample for validation allowed us to estimate the calibration intercept with standard errors of 0.115 at 7 days and 0.061 at 30 days. Similarly, we could estimate calibration slope with standard errors of 0.123 and 0.067 at 7 and 30 days, respectively. We employed hyperparameter tuning on the validation sample using Bayesian optimization [26]. Model performance was assessed in the test sample using the c-statistic, calibration-in-the-large, calibration slope, and calibration plots. Calibration plots were constructed by obtaining model-predicted probabilities for each patient, stratifying patients into deciles based on predicted probability, and plotting mean predicted probabilities versus

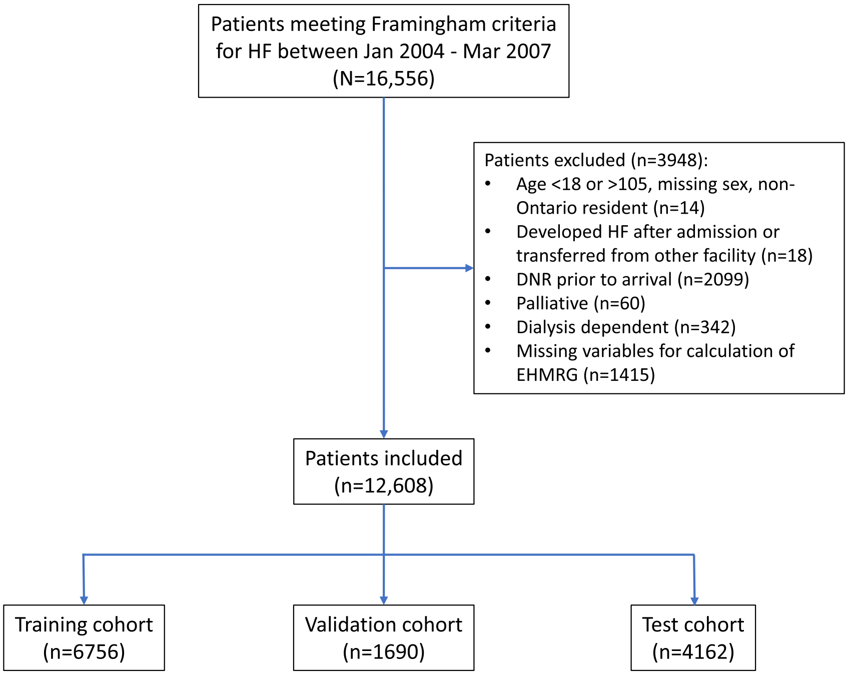

Patients meeting Framingham criteria for HF between Jan 2004 - Mar 2007 (N=16,556)
Patients excluded (n=3948):
• Age <18 or >105, missing sex, non-Ontario resident (n=14)
• Developed HF after admission or transferred from other facility (n=18)
• DNR prior to arrival (n=2099)
• Palliative (n=60)
• Dialysis dependent (n=342)
• Missing variables for calculation of EHMRG (n=1415)
Patients included (n=12,608)
Training cohort (n=6756)
Validation cohort (n=1690)
Test cohort (n=4162)

Fig. 1. Flow diagram.

Table 1
Machine learning and conventional statistical model comparison for prediction of 7- and 30-day mortality based on AUC (95%CI), and Δ AUC (95%CI).

Model type	7-day AUC	Δ AUC vs EHMRG 7	30-day AUC	Δ AUC vs EHMRG 30
EHMRG	0.794 (0.789, 0.800)	–	0.755 (0.750, 0.762)	–
Lasso	0.774 (0.743, 0.806)	−0.020 (−0.051, 0.012)	0.743 (0.725, 0.762)	−0.012 (−0.030, 0.007)
Neural Network	0.669 (0.613, 0.728)	−0.120 (−0.181, −0.066)	0.680 (0.650, 0.710)	−0.075 (−0.105, −0.045)
Random Forest	0.737 (0.698, 0.777)	−0.057 (−0.096, −0.017)	0.747 (0.728, 0.767)	−0.008 (−0.027, 0.012)
XGBoost	0.757 (0.721, 0.793)	−0.037 (−0.073, −0.001)	0.759 (0.740, 0.779)	0.004 (−0.015, 0.024)
Enhanced EHMRG	0.793 (0.789, 0.797)	−0.001 (−0.005, 0.003)	0.760 (0.756, 0.769)	0.005 (−0.004, 0.014)

AUC - area under the curve.
CI - confidence interval.
EHMRG - Emergency Heart failure Mortality Risk Grade risk prediction at 7 days or 30 days.

observed death rates. When working with neural networks, we employed random oversampling from the Imbalanced-Learn library to achieve a more balanced class distribution [27], because it has been shown that neural networks can perform poorly when dealing with class imbalances [28].

We used SHAP [29], a method that provides a measure of variable importance, to determine which features the top performing models found to be the most informative. We iteratively added informative features to the original EHMRG covariates and retained features that improved the AUC $\geq$ 0.005 in the validation set to construct an 'Enhanced EHMRG' logistic model comprised of the final set of such features.

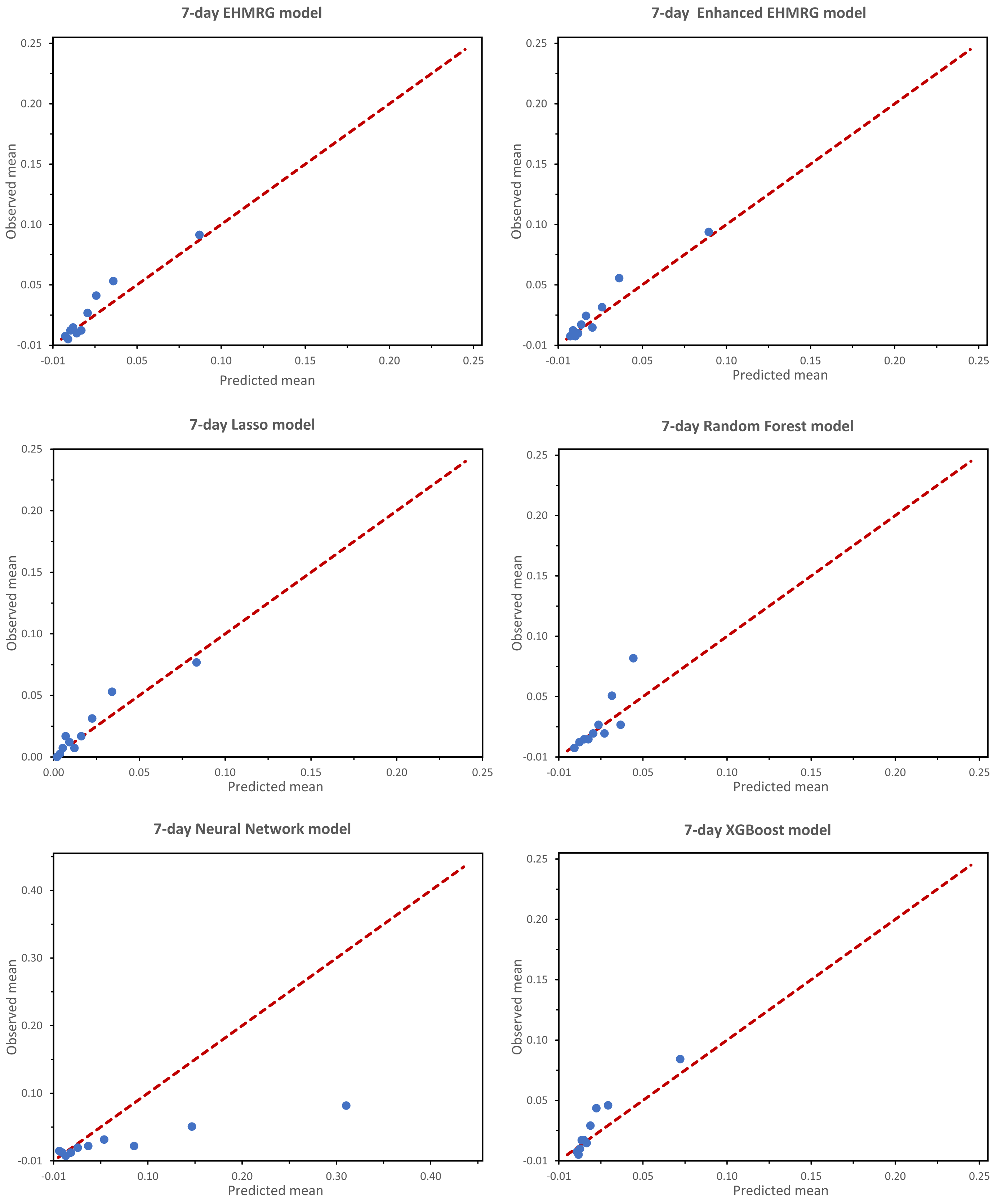
As these analyses were conducted at ICES, the use of these data was authorized under section 45 of Ontario's Personal Health Information Protection Act, which does not require review by a Research Ethics Board. Original chart review and data collection were performed with REB approval as previously described [6]. Analyses were performed using Python 3.7.4, Scikit-Learn 0.22.2.post1 [30], Pandas 0.25.0, Numpy 1.16.4, Shap 0.35.0, Matplotlib 3.1.3, and Imblearn 0.5.0.

3. Results

There were 12,608 patient records divided into training, validation and test sets (Fig. 1). Baseline characteristics are provided in Supplementary table 2. There were 244 deaths at 7 days and 761 deaths at 30 days overall. The number of 7-day deaths in the training, validation, and test sets were 125 (1.85%), 26 (1.54%), and 93 (2.23%) respectively. There were 400 (5.92%), 96 (5.68%) and 265 (6.37%) deaths at 30 days in the training, validation, and test sets, respectively. AUCs for the EHMRG 7- and 30-day models in the test sample were 0.794 (95% CI; 0.789, 0.800) and 0.755 (95% CI; 0.750, 0.762), respectively (Table 1).

Comparative performance of ML models. For 7-day and 30-day mortality, predictors derived from machine learning models were ranked according to variable importance using the lasso (Supplementary figs. 1 and 2), neural network (Supplementary figs. 3 and 4), random forest (Supplementary figs. 5 and 6), and XGBoost (Supplementary figs. 7 and 8). The list of variable names is described fully in Supplementary table 3. For 7-day mortality, the Lasso regression performed better than the other ML models but did not outperform the EHMRG 7-day model. For 30-day mortality, the AUC was higher for XGBoost than the other models, but it was not significantly different from the EHMRG 30-day model. Compared to the EHMRG 7-day model, ΔAUC was lower for the 7-day random forest (ΔAUC -0.057; 95% CI, −0.096, −0.017), and the 7-day XGBoost (ΔAUC -0.037; 95% CI, −0.073, −0.001). Compared to the EHMRG 30-day model, ΔAUC was substantially lower with the 30-day neural network (ΔAUC -0.075; 95% CI, −0.105, −0.045). Other ML models did not differ from EHMRG (Table 1).

Identifying new predictors of mortality using ML. For 7-day mortality, the only feature that increased the AUC by >0.005 compared to the EHMRG model was the time between the patient's earliest non-cardiovascular-related admission within the past two years and the index arrival date. This variable was the only one of 10 prior hospital



7-day EHMRG model

7-day Enhanced EHMRG model

7-day Lasso model

7-day Random Forest model

7-day Neural Network model

7-day XGBoost model

Fig. 2. Mean predicted probability vs. observed probability for 7-day mortality, by modeling approach.

Table 2
Brier scores, calibration-in-the-large, calibration slope, and 95%CI, for conventional statistical models versus machine learning.

	7 day Brier score	30 day Brier score	7 day calibration-in-the-large	7 day calibration slope	30 day calibration–in-the-large	30 day calibration slope
EHMRG	0.021 (0.017, 0.025)	0.056 (0.050, 0.062)	0.017 (−0.657, 0.692)	0.954 (0.769, 1.139)	−0.026 (−0.374, 0.322)	0.964 (0.831, 1.098)
Lasso	0.021 (0.017, 0.026)	0.056 (0.050, 0.062)	−0.264 (−0.946, 0.419)	0.882 (0.969, 1.068)	−0.285 (−0.610, 0.040)	0.861 (0.737, 0.985)
Neural Network	0.032 (0.027, 0.037)	0.093 (0.084, 0102)	−2.447 (−2.821, −2.072)	0.032 (0.023, 0.042)	−1.962 (−2.163, −1.761)	0.019 (0.014, 0.024)
Random Forest	0.021 (0.017, 0.026)	0.057 (0.051, 0.063)	1.679 (0.636, 2.721)	1.408 (1.120, 1.695)	1.979 (1.337, 2.622)	1.737 (1.486, 1.988)
XGBoost	0.023 (0.019, 0.026)	0.063 (0.059, 0.068)	−0.660 (−1.197, −0.122)	1.067 (0.868, 1.267)	−0.863 (−1.108, −0.617)	1.025 (0.884, 1.165)
Enhanced EHMRG	0.021 (0.017, 0.025)	0.056 (0.050, 0.062)	−0.128 (−0.785, 0.529)	0.914 (0.735, 1.093)	−0.019 (−0.358, 0.320)	0.963 (0.833, 1.093)

admission or ED visit variables that were included to supplement the ML model, indicating that there was limited discriminative ability of prior hospitalizations to predict acute HF mortality. For 30-day mortality, sodium concentration and long-term care residence status improved AUC marginally when added to the original EHMRG 30-day model. However, when the above variables were included in the enhanced 7- and 30-day models, there was no significant improvement in AUC compared to the base EHMRG models as shown in Table 1.

Calibration and Brier Score. As shown in the decile plots for 7-day mortality (Fig. 2), calibration was poor for neural networks when predicted risk was high and was poor for the random forest even at intermediate predicted risk. From calibration plots, the 7-day EHMRG and the remaining ML models for 7-day death were well calibrated (Fig. 2). For 30-day mortality, all approaches except neural networks were well-calibrated (Supplementary fig. 9). For 7-day prediction, EHMRG had the best calibration as demonstrated by both calibration-in-the-large being closest to zero and calibration slope being closest to one (Table 2). The next best model at 7 days was the logistic regression based enhanced EHMRG which used machine learning to identify new predictors and logistic regression to model mortality. EHMRG and enhanced EHMRG also had the best Brier Scores for 7-day mortality. For 30-day mortality, XGBoost had the best calibration slope, however, the calibration intercept was poor. EHMRG and the enhanced EHMRG had the best calibration-in-the-large (since calibration intercept was closest to zero) and the next highest calibration slope after XGBoost. For 30-day prediction, EHMRG, enhanced EHMRG, and the lasso had the best Brier Scores. Neural networks had the worst Brier Scores for both outcomes.

4. Discussion

We found that ML algorithms demonstrated comparable discrimination to the logistic regression-based EHMRG 7- and 30-day mortality models, although there was variability in performance of the different ML-based models. Neural networks had consistently lower discriminative ability and substantial miscalibration for predicting mortality. Despite considering over 100 variables, including medications prior to hospital arrival, using ML to identify new covariates to enhance the logistic regression models did not tend to improve AUC. We also found that comparing models using calibration provided insights into the optimal model. Using calibration indices, EHMRG was optimal for 7-day outcomes. For 30-day outcomes, the findings were more nuanced, with XGBoost demonstrating the best calibration slope but poor calibration intercept. EHMRG and the enhanced EHMRG were optimal since they had both good calibration intercept and slope.

Machine learning has the potential to personalize medical care by incorporating information and data that arise from 'collective experience' [31]. However, there is ongoing debate about the role of ML for predicting health outcomes. Some studies have reported higher AUCs with ML compared to CSM for complex problems such as predicting readmissions [32], outcomes in HF with preserved ejection fraction [33], and identifying incident atrial fibrillation [34]. Christodoulou et al. reported that ML was not superior to logistic regression-based clinical prediction models in 18 different medical fields [35]. In our recently-published systematic review of ML algorithms vs. CSM for prediction of outcomes after acute myocardial infarction, we identified several potential sources of bias in prior comparative studies [14]. With respect to the number of events per variable (EPV), a simulation study found that logistic regression models are optimal when the number of EPV exceeds a 10:1 ratio [36], whereas no such convention exists for machine learning. Research by van der Ploeg et al. showed that ML methods may require a larger sample size than CSM in order to have optimism below a given threshold [37]. Consequently in smaller sample sizes, ML methods may result in more optimistic estimates of performance than CSM. Thus, there is still equipoise and need for more rigorously-conducted studies comparing ML and CSM for prognosis-based research.

Our study adds to existing knowledge by demonstrating that validated conventional statistical models, despite greater parsimony, may be superior to more complex approaches that use ML. While ML models have unique strengths, our current study suggests that they should be compared with CSM, such as multiple logistic regression or Cox models whenever possible since a simpler approach with greater interpretability is preferred. Unless a comparison is performed, it cannot be easily predicted whether machine learning will be better than or inferior to conventional statistical models. Second, while many prior studies examined only the AUC when comparing ML vs. CSM [13,14], our current study demonstrates that examining calibration provides added benefits when trying to ascertain the value of ML algorithms. In a previously-published systematic review of ML vs. CSM in acute heart failure, all 20 studies reported AUCs but only two studies overall (and none utilizing neural networks) reported on calibration [13]. A similar lack of reporting of calibration of ML approaches was found in another systematic review of 24 studies of acute myocardial infarction [14]. Our study suggests that calibration should be routinely reported when machine learning is used for prediction. To this end, the Brier score (a measure of overall predictive accuracy) and measures of calibration (e.g., calibration intercept and slope) should be reported when assessing the performance of ML methods. Simply reporting the c-statistic is insufficient.

In conclusion, logistic regression-based mortality prediction in acute HF performed comparably to ML models. However, ML approaches may result in predictions that display poor calibration. Newly developed ML models for prognosis should be routinely compared with validated risk models for their discrimination and calibration.

Author contributions

Conceptualization: DEA, DSL, BW. Data curation: DEA, DSL, CXW, SM, XW, JP. Formal analysis: DEA, DSL, CXW, SM, XW, BW. Funding acquisition: DSL, BW. Investigation: All authors. Methodology: All authors. Project administration: DSL, JP, BW. Resources: DSL, JP, BW.

Software: DEA, CXW, SM, XW, BW. Supervision: DSL, CXW, XW, JP, BW. Validation: DEA, DSL, SM, XW, BW. Visualization: DEA, DSL, CXW, SM, BW. Writing original draft: DEA, JP. Writing review & editing: DSL, SM, BW

Funding/support

This study was funded by a Foundation grant from the Canadian Institutes of Health Research (CIHR) grant # FDN 148446 and a Strategic Innovation Grant from the Ted Rogers Centre for Heart Research. It is also supported by ICES, which is funded by an annual grant from the Ontario Ministry of Health (MOH) and the Ministry of Long-Term Care (MLTC). The analyses, conclusions, opinions and statements expressed herein are solely those of the authors and do not reflect those of the funding sources; no endorsement is intended or should be inferred. Dr. Lee is the Ted Rogers Chair in Heart Function Outcomes, University Health Network, University of Toronto.

Declaration of Competing Interest

There are no conflicts of interest to declare.

Acknowledgements

ICES is an independent, non-profit research institute whose legal status under Ontario's health information privacy law allows it to collect and analyze health care and demographic data, without consent, for health system evaluation and improvement. Parts of this material are based on data and/or information compiled and provided by the Canadian Institute for Health Information (CIHI). The analyses, conclusions, opinions and statements expressed herein are solely those of the authors and do not reflect those of CIHI; no endorsement is intended or should be inferred. We thank IQVIA Solutions Canada Inc. for use of their Drug Information File.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ijcard.2022.07.035.

References

[1] P.A. Heidenreich, N.M. Albert, L.A. Allen, et al., Forecasting the impact of heart failure in the United States: a policy statement from the American Heart Association, Circ. Heart Fail. 6 (3) (2013) 606–619, https://doi.org/10.1161/HHF.0b013e318291329a.

[2] K. Anderson, H.J. Ross, P.C. Austin, J. Fang, D.S. Lee, Health care use before first heart failure hospitalization: identifying opportunities to pre-emptively diagnose impending decompensation, JACC Heart Fail. 8 (12) (2020) 1024–1034, https://doi.org/10.1016/j.jchf.2020.07.008.

[3] M. Arrigo, M. Jessup, W. Mullens, et al., Acute heart failure, Nat. Rev. Dis. Primers 6 (1) (2020) 16, https://doi.org/10.1038/s41572-020-0151-7.

[4] I.R. Raslan, H.J. Ross, R.A. Fowler, et al., The associations between direct and delayed critical care unit admission with mortality and readmissions among patients with heart failure, Am. Heart J. 233 (2021) 20–38, https://doi.org/10.1016/j.ahj.2020.11.002.

[5] R. Dunbar-Yaffe, A. Stitt, J.J. Lee, S. Mohamed, D.S. Lee, Assessing risk and preventing 30-day readmissions in decompensated heart failure: opportunity to intervene? Curr. Heart Fail. Rep. 12 (5) (2015) 309–317, https://doi.org/10.1007/s11897-015-0266-4.

[6] D.S. Lee, A. Stitt, P.C. Austin, et al., Prediction of heart failure mortality in emergent care: a cohort study, Ann. Intern. Med. 156 (11) (2012) 767–775, https://doi.org/10.7326/0003-4819-156-11-201206050-00003.

[7] D. Greig, P.C. Austin, L. Zhou, et al., Ischemic electrocardiographic abnormalities and prognosis in decompensated heart failure, Circ. Heart Fail. 7 (6) (2014) 986–993, https://doi.org/10.1161/CIRCHEARTFAILURE.114.001460.

[8] D.S. Lee, S.E. Straus, P.C. Austin, et al., Rationale and design of the comparison of outcomes and access to care for heart failure (COACH) trial: a stepped wedge cluster randomized trial, Am. Heart J. 240 (2021) 1–10, https://doi.org/10.1016/j.ahj.2021.05.003.

[9] D.S. Lee, J.S. Lee, M.J. Schull, et al., Prospective validation of the emergency heart failure mortality risk grade for acute heart failure, Circulation. 139 (9) (2019) 1146–1156, https://doi.org/10.1161/CIRCULATIONAHA.118.035509.

[10] A.M. Michaud, S.I.A. Parker, H. Ganshorn, J.A. Ezekowitz, A.D. McRae, Prediction of early adverse events in emergency department patients with acute heart failure: a systematic review, Can. J. Cardiol. 34 (2) (2018) 168–179, https://doi.org/10.1016/j.cjca.2017.09.004.

[11] D.S. Lee, J.S. Lee, M.J. Schull, J.M. Grimshaw, P.C. Austin, J.V. Tu, Design and rationale for the acute congestive heart failure urgent care evaluation: the ACUTE study, Am. Heart J. 181 (2016) 60–65, https://doi.org/10.1016/j.ahj.2016.07.016.

[12] F.E. Harrell Jr., Glossary of statistical terms, Vanderbilt University School of Medicine, 2022. Accessed August 11, 2019, http://hbiostat.org/doc/glossary.pdf.

[13] S. Shin, P.C. Austin, H.J. Ross, et al., Machine learning vs. conventional statistical models for predicting heart failure readmission and mortality, ESC Heart Fail. 8 (1) (2021) 106–115, https://doi.org/10.1002/ehf2.13073.

[14] S.M. Cho, P.C. Austin, H.J. Ross, et al., Machine learning compared with conventional statistical models for predicting myocardial infarction readmission and mortality: a systematic review, Can. J. Cardiol. 37 (8) (2021) 1207–1214, https://doi.org/10.1016/j.cjca.2021.02.020.

[15] J.V. Tu, L.R. Donovan, D.S. Lee, et al., Effectiveness of public report cards for improving the quality of cardiac care: the EFFECT study: a randomized trial, JAMA 302 (21) (2009) 2330–2337, https://doi.org/10.1001/jama.2009.1731.

[16] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32, https://doi.org/10.1023/A:1010933404324.

[17] T. Chen, C. Guestrin, XGBoost: a scalable tree boosting system, in: KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785–794, https://doi.org/10.1145/2939672.2939785.

[18] D.E. Rumelhart, G.E. Hinton, R.J. Williams, Learning representations by back-propagating errors, Nature. 323 (1986) 533–536, https://doi.org/10.1038/323533a0.

[19] R. Nedadur, B. Wang, B. Yanagawa, The cardiac surgeon's guide to artificial intelligence, Curr. Opin. Cardiol. 36 (5) (2021) 637–643, https://doi.org/10.1097/HCO.0000000000000888.

[20] R. Nedadur, B. Wang, W. Tsang, Artificial intelligence for the echocardiographic assessment of valvular heart disease, Heart. (2022), https://doi.org/10.1136/heartjnl-2021-319725.

[21] D.S. Lee, S. Ma, A. Chu, et al., Predictors of mortality among long-term care residents with SARS-CoV-2 infection, J. Am. Geriatr. Soc. 69 (12) (2021) 3377–3388, https://doi.org/10.1111/jgs.17425.

[22] D.S. Lee, C.X. Wang, F.A. McAlister, et al., Factors associated with SARS-CoV-2 test positivity in long-term care homes: a population-based cohort analysis using machine learning, Lancet Reg. Health Am. 6 (2022), 100146, https://doi.org/10.1016/j.lana.2021.100146.

[23] I. Guyon, J. Weston, S. Barnhill, V. Vapnik, Gene selection for cancer classification using support vector machines, Mach. Learn. 46 (1–3) (2002) 389–422, https://doi.org/10.1023/a:1012487302797.

[24] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed., Springer Science and Business Media, 2017.

[25] M. Pavlou, C. Qu, R.Z. Omar, et al., Estimation of required sample size for external validation of risk models for binary outcomes, Stat. Methods Med. Res. 30 (10) (2021) 2187–2206, https://doi.org/10.1177/09622802211007522.

[26] J. Wu, X. Chen, H. Zhang, L. Xiong, H. Lei, S. Deng, Hyperparameter optimization for machine learning models based on Bayesian optimization, J. Electron. Sci. Technol. 17 (1) (2019) 26–40, https://doi.org/10.11989/JEST.1674-862X.80904120.

[27] G. Lemaitre, F. Nogueira, C. Aridas, Imbalanced-learn: a python toolbox to tackle the curse of imbalanced datasets in machine learning, J. Mach. Learn. Res. 18 (17) (2017) 1–5.

[28] R. Anand, K.G. Mehrotra, C.K. Mohan, S. Ranka, An improved algorithm for neural network classification of imbalanced training sets, IEEE Trans. Neural Netw. 4 (6) (1993) 962–969, https://doi.org/10.1109/72.286891.

[29] S.M. Lundberg, S.I. Lee, A unified approach to interpreting model predictions, in: 31st Conference on Neural Information Processing Systems (NIPS 2017), 2017, pp. 1–10, in: https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf.

[30] F. Pedregosa, G. Varoquaux, A. Gramfort, et al., Scikit-learn: machine learning in python, J. Mach. Learn. Res. 2011 (12) (2011) 2825–2830.

[31] A. Rajkomar, J. Dean, I. Kohane, Machine learning in medicine, N. Engl. J. Med. 380 (14) (2019) 1347–1358, https://doi.org/10.1056/NEJMra1814259.

[32] D.J. Morgan, B. Bame, P. Zimand, et al., Assessment of machine learning vs standard prediction rules for predicting hospital readmissions, JAMA Netw. Open 2 (3) (2019), e190348, https://doi.org/10.1001/jamanetworkopen.2019.0348.

[33] S. Angraal, B.J. Mortazavi, A. Gupta, et al., Machine learning prediction of mortality and hospitalization in heart failure with preserved ejection fraction, JACC Heart Fail. 8 (1) (2020) 12–21, https://doi.org/10.1016/j.jchf.2019.06.013.

[34] P. Tiwari, K.L. Colborn, D.E. Smith, F. Xing, D. Ghosh, M.A. Rosenberg, Assessment of a machine learning model applied to harmonized electronic health record data for the prediction of incident atrial fibrillation, JAMA Netw. Open 3 (1) (2020), e1919396, https://doi.org/10.1001/jamanetworkopen.2019.19396.

[35] E. Christodoulou, J. Ma, G.S. Collins, E.W. Steyerberg, J.Y. Verbakel, B. Van Calster, A systematic review shows no performance benefit of machine learning

over logistic regression for clinical prediction models, J. Clin. Epidemiol. 110 (2019) 12–22, https://doi.org/10.1016/j.jclinepi.2019.02.004.

[36] P. Peduzzi, J. Concato, E. Kemper, T.R. Holford, A.R. Feinstein, A simulation study of the number of events per variable in logistic regression analysis, J. Clin. Epidemiol. 49 (12) (1996) 1373–1379, https://doi.org/10.1016/s0895-4356(96)00236-3.

[37] T. van der Ploeg, P.C. Austin, E.W. Steyerberg, Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints, BMC Med. Res. Methodol. 14 (2014) 137, https://doi.org/10.1186/1471-2288-14-137.