European Heart Journal - Digital Health (2024) **5**, 324–334
https://doi.org/10.1093/ehjdh/ztae005

ORIGINAL ARTICLE

Predicting heart failure outcomes by integrating breath-by-breath measurements from cardiopulmonary exercise testing and clinical data through a deep learning survival neural network

Heather J. Ross [1], Mohammad Peikari[1], Julie K.K. Vishram-Nielsen[1], Chun-Po S. Fan[1], Jason Hearn[1], Mike Walker[1], Edgar Crowdy[1], Ana Carolina Alba[1], and Cedric Manlhiot [1,2,*]

[1]The Ted Rogers Centre for Heart Research, Peter Munk Cardiac Centre, University Health Network, Department of Medicine, University of Toronto, 200 Elizabeth Street, Toronto, ON M5G 2C4, Canada; and [2]The Blalock-Taussig-Thomas Pediatric and Congenital Heart Center, Department of Pediatrics, Johns Hopkins School of Medicine, Johns Hopkins University, 1800 Orleans Street, Baltimore, MD 21287, USA

Received 2 February 2023; revised 15 December 2023; accepted 2 January 2024; online publish-ahead-of-print 31 January 2024

Aims

Mathematical models previously developed to predict outcomes in patients with heart failure (HF) generally have limited performance and have yet to integrate complex data derived from cardiopulmonary exercise testing (CPET), including breath-by-breath data. We aimed to develop and validate a time-to-event prediction model using a deep learning framework using the DeepSurv algorithm to predict outcomes of HF.

Methods and results

Inception cohort of 2490 adult patients with high-risk cardiac conditions or HF underwent CPET with breath-by-breath measurements. Potential predictive features included known clinical indicators, standard summary statistics from CPETs, and mathematical features extracted from the breath-by-breath time series of 13 measurements. The primary outcome was a composite of death, heart transplant, or mechanical circulatory support treated as a time-to-event outcomes. Predictive features ranked as most important included many of the features engineered from the breath-by-breath data in addition to traditional clinical risk factors. The prediction model showed excellent performance in predicting the composite outcome with an area under the curve of 0.93 in the training and 0.87 in the validation data sets. Both the predicted vs. actual freedom from the composite outcome and the calibration of the prediction model were excellent. Model performance remained stable in multiple subgroups of patients.

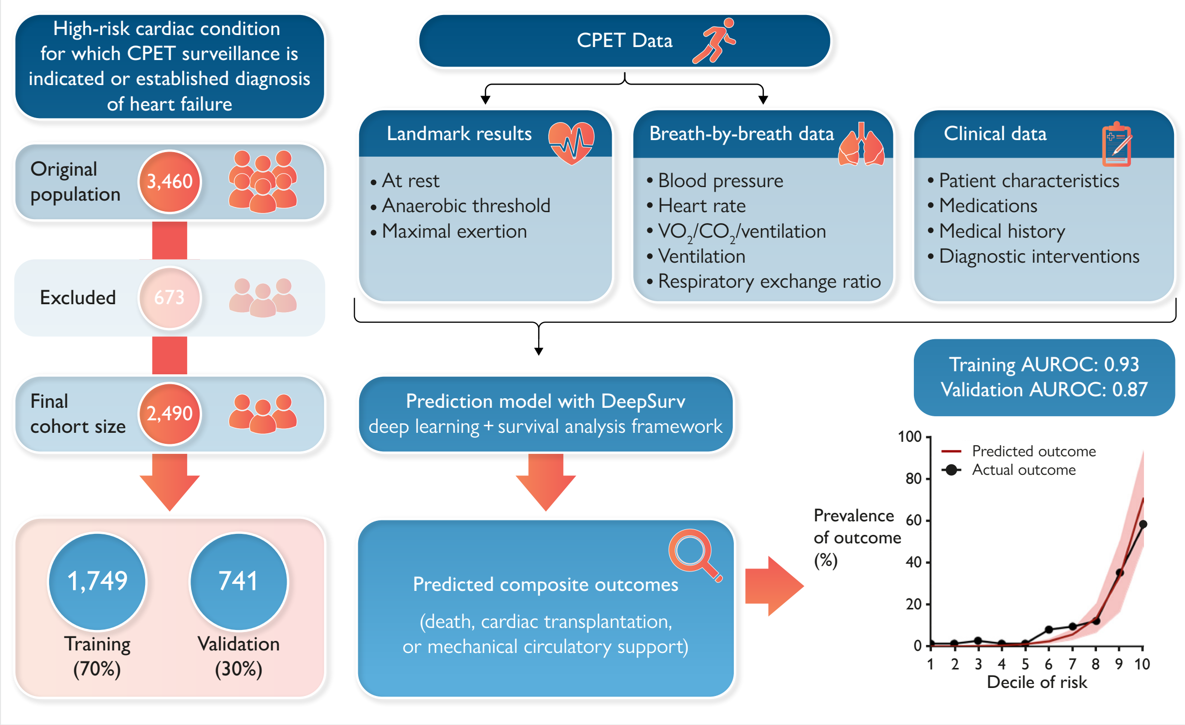
Conclusion

Using a combined deep learning and survival algorithm, integrating breath-by-breath data from CPETs resulted in improved predictive accuracy for long-term (up to 10 years) outcomes in HF. DeepSurv opens the door for future prediction models that are both highly performing and can more fully use the large and complex quantity of data generated during the care of patients with HF.

* Corresponding author. Tel: +(410) 955-5000, Email: cmanlhi1@jhmi.edu

© The Author(s) 2024. Published by Oxford University Press on behalf of the European Society of Cardiology.
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (https://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

Graphical Abstract



High-risk cardiac condition for which CPET surveillance is indicated or established diagnosis of heart failure

Original population 3,460

Excluded 673

Final cohort size 2,490

1,749 Training (70%)

741 Validation (30%)

CPET Data

Landmark results
- At rest
- Anaerobic threshold
- Maximal exertion

Breath-by-breath data
- Blood pressure
- Heart rate
- VO_2/CO_2/ventilation
- Ventilation
- Respiratory exchange ratio

Clinical data
- Patient characteristics
- Medications
- Medical history
- Diagnostic interventions

Prediction model with DeepSurv deep learning + survival analysis framework

Predicted composite outcomes (death, cardiac transplantation, or mechanical circulatory support)

Training AUROC: 0.93
Validation AUROC: 0.87

Keywords Heart failure • Prognosis • Machine learning • Survival analysis • Exercise testing

Background

The number of patients living with heart failure (HF) has been steadily increasing owing to an aging population, increased survival of patients at high risk for HF (e.g. those with ischaemic heart disease), and improved outcomes of those living with HF (secondary to the use of guideline-directed therapies).[1,2] The ability of predictive models to guide therapy, counsel patients, and anticipate disease progression is of critical importance. Indeed, many models have been developed to predict prognosis in patients with HF, often focusing exclusively on clinical risk factors.[3–5] Data from cardiopulmonary exercise testing (CPET) have been found to be an important prognostic factor for patients with heart failure, both as specific indices[6–8] and through composite scores derived from those indices.[9,10] These data have been included in predictive models[10–12]; however, these models have suffered from poor performance with c-statistics generally <0.75.[13]

In a typical CPET, various physiological parameters are measured either on a breath-by-breath basis or monitored continuously. From these data, measurements at pre-specified clinical landmarks, calculations of rate of change (slopes), or ratios between specific variables are used to summarize the results of the test and these are the data that are used to guide clinical care and are incorporated in clinical prediction models. Thus, the majority of the measurements generated during an exercise test is generally not used either clinically or in prediction models. However, we recently found breath-by-breath data substantially improved predictive model performance for 1-year outcomes in HF patients over single CPET indices or published composite scores derived from CPET.[14] The challenge with using these data is that without data reduction techniques (which substantially reduces the informativeness of the data), time series do not integrate well into classic probabilistic modelling methods. However, recent advances with the integration of time-to-event analysis in machine learning algorithms now make this possible.[15] Thus, the objective of this study was to use machine learning to create and internally validate a predictive model for a combined outcome of death, need for heart transplant, or mechanical circulatory support that integrates clinical risk factors, CPET indices, and breath-by-breath data.

Methods

Patient population

This single-centre retrospective study included consecutive ambulatory patients 18 years or older with a high-risk condition for which CPET surveillance was indicated (gene-carrying/heritable cardiomyopathy, cardiotoxic exposure) or with an established diagnosis of HF from any aetiology other than congenital heart disease. Patients with severe pulmonary disease were not included in this study. Patients were followed at the Peter Munk Cardiac Centre at University Health Network between December 2001 and December 2018. Patients were included if they had at least one CPET with available breath-by-breath data performed during the review period and followed for at least 12 months afterwards (unless they experienced an event during the 12-month observation period). The original patient population included 3460 unique patients/CPET pairs, of whom 673 were excluded because the breath-by-breath data were improperly saved at the time of the test, 255 had congenital heart disease, and 2 were excluded because of a previous ventricular assist device implantation (subsequently recovered), leaving a final cohort of 2490 patients. The Research Ethics Board of the University Health Network (Toronto, ON, Canada) approved this study. The requirement for patient consent was waived because of the retrospective nature of the study. C.M. had full access to all of the data in this study and takes responsibility for its integrity and for the data analysis.

Clinical exercise protocol

Clinical exercise protocol at our institution is standardized for all patients and is based on the 2002 American College of Cardiology (ACC)/American Heart Association (AHA) Guidelines for Exercise Testing.[16] Tests were performed by a single operator (M.W.) who decided on appropriate clinical deviations to the testing protocol for each individual patient as needed. Cardiopulmonary exercise testing was performed using the ramp protocol with a cycle ergometer (Lods MedGraphics) and a metabolic cart (MedGraphics CardioO$_2$ Ultima); equipment and software were updated over time as appropriate and as directed by the manufacturer. Tests start with an initial minute of rest in a seated position an addition minute of warm-up (at a load of 0 watts). Thereafter, an individualized (in duration and intensity) ramp protocol is used to achieve full exercise with increments of 10 watts per min. Ventilation (VE), VO$_2$, and VCO$_2$ were collected through the breath-by-breath analysis of expired gases. The average of the middle five of the last seven breaths was used to calculate peak VO$_2$. Oxygen uptake efficiency slope (OUES) was calculated from the following standardized formula (OUES indicated by a): VO$_2$ (mL/min) = $a[\log_{10}(\text{VE})] + b$. A least square mean regression fitted over the entire exercise test was used to calculate the VE/VCO$_2$ slope. Exercise tests were excluded from the analysis if the total test duration was <60 s, likely indicating a technical problem, or the test was deemed to be faulty based on an average respiratory rate <10 breaths per min. While the duration of CPETs varied between patients, <5% of patients of tests had a duration below 5 min, ~30% of tests had a duration >12 min, and 99% of patients had tests in the 4–19 min range. Test duration was included as a feature in the prediction model.

Study outcomes

The primary study outcome was a composite endpoint including death from any cause or need for heart transplantation or mechanical circulatory support [durable left ventricular assist device (most common), extracorporeal membrane oxygenation (ECMO), intra-aortic balloon pump, or Impella-type devices] for any duration. All-cause mortality was used as an outcome instead of mortality from cardiovascular cause only as many patients with HF die proximally of non-cardiac causes which are at least partially associated with the underlying HF; furthermore, in many cases, cause of death is not sufficiently well documented to adjudicate cause of death appropriately.[17] All analyses were performed as time-to-event analyses with the starting time being defined as the first qualifying CPET for each patient (in order to allow us to consider the largest possible time frame, subsequent studies were not considered in this analysis). Patients without outcomes were censored at the end of follow-up or on 31 December 2019 whichever came first. Outcome ascertainment was done through chart review, regular clinical follow-up, and active contact with patients by the HF clinical staff.

Predictive features

Predictive features in this study included both data generated from the CPETs and clinical data obtained at or around the test (corresponding clinical visit or ±3 months for laboratory, echocardiogram, and electrocardiogram findings). For the purpose of this study, we define clinical data as patient characteristics, medications, previous medical history, and diagnostic investigations. Three levels of data were extracted from the testing software: (i) summary data included in the standard exercise test report (which included both machine-generated and operator-acquired features), (ii) staged data which consisted of salient exercise performance indices measured at pre-specified exercise landmarks (at rest, at anaerobic threshold, and at maximal exertion), and (iii) breath-by-breath (i.e. measured for each breathing cycle) data. Indices measured on a breath-by-breath basis included end-tidal carbon dioxide (petCO$_2$) and oxygen (petO$_2$) tension, respiratory exchange ratio (RER), oxygen saturation (SpO$_2$), oxygen uptake (VO$_2$) efficiency slope, carbon dioxide production (VCO$_2$), minute ventilation (VE), and workload (watts) along with various ratios and indices derived from these measurements. Systolic and diastolic blood pressure and heart rate were monitored continuously throughout the CPET, and we aligned the blood pressure and heart rate time series to the breath-by-breath time series to generate heart rate and blood pressure measurements for each breath.

Clinical data (provided in *Table 1*) were extracted manually from the medical records and included patient demographics; aetiology of HF; co-morbidities; presence and type of pacemaker, cardiac resynchronization devices, or implantable cardioverter defibrillator (ICD); cardiac medications at the time of the test; laboratory investigations; New York Heart Association (NYHA) functional class; and ejection fraction and heart rhythm.

Data preprocessing

Cardiopulmonary exercise testing breath-by-breath data consist of a collection of time series variables captured on a breath-by-breath base. From these time series, over 2000 mathematical features are derived using the *tsfresh* python library.[18] For clinical variables and data generated from the CPET other than the breath-by-breath data, categorical fields were converted to binary fields using one-hot-encoding methods, and fields with continuous values were processed to remove outliers or irrelevant information. Patient records with ≥20% missing data were removed from the analysis, and variables with ≥35% missing values were not considered further in the analysis. The remainder of missing values was imputed using the R's Multivariate Imputation by Chained Equations (MICE) library.[19]

Algorithm development

The DeepSurv survival analysis method previously described[15] was used to model patients' outcome over their follow-up period. The method is essentially a multi-layer feed-forward neural network that models the effect of patients' covariates with their hazard rate using the network's weights. Each hidden layer of the network consists of a fully connected layer of nodes separated with a dropout layer.[20] The output of this model is a single node with linear activation which estimates the log-risk function in a Cox model. In order to train the network, modern deep learning techniques have been used including Scaled Exponential Linear Unit (SELU)[21] as activation functions, Adaptive Moment Estimation (Adam)[22] as gradient descent optimizer algorithm with Nesterov momentum,[23] and learning rate scheduling,[24] all of which are summarized in the original paper by Katzman *et al.*[15]

Data were randomly divided into mutually exclusive train (70%) and validation test (30%) sets. The training set was used to tune modelling parameters and generate a final model using all training data and with the best tuning parameters. The validation test set was then used to evaluate the performance of the final tuned model. A 5-fold cross-validation scheme was used to tune parameters using the training set. Hyperparameter tuning was done semi-automatically. That is, first, the number of layers and nodes was experimented, and second, the following parameters were tuned using cross-validation: learning rate, dropout rate, layer activation functions, feature dimensionality, and optimization method. The final network that was used in this study consisted of 2 hidden layers each with 100 nodes, with a learning and dropout rate of 0.1 and 0.4, respectively. In order to reduce dimensionality of the *tsfresh*-derived features from breath-by-breath data (initially over 2000 features), an initial feature selection technique was used. A systematic search of feature dimensionalities was performed with numbers (K) ranging from 5 to 500 on breath-by-breath–derived features only. For this study, we used analysis of variance (ANOVA) F-test method[25] for reducing the number of breath-by-breath–derived features. Specifically, the *P*-values for individual features were found and used to rank features in order; the top K features were taken as the feature subset. Finally, highly correlated variables were also removed from the selected set of breath-by-breath–derived features. The medical data were then added to the selected features.

Prediction of outcomes by the algorithm was performed using the same time interval from CPET as in the original data (e.g. if follow-up/event occurs x years after the qualifying CPET, then the same value of x was used for the timing of the prediction). To calibrate model probabilities and prediction threshold, a logistic regression (LR) model was used on the training probabilities and their labels to obtain coefficients and intercept of the LR model. Next, to calibrate test set probabilities, the found intercept and coefficient

Table 1 Patient characteristics and outcomes stratified in training vs. validation cohorts

	N	Training cohort	*N*	Validation cohort	*N*	All patients	*P*
Demographics and comorbidities							
Age at baseline (years)	1749	46.5 ± 16.1	741	46.3 ± 16.5	2490	46.4 ± 16.2	0.78
Female (vs. male)	1749	594 (34.0%)	741	273 (36.7%)	2490	867 (34.8%)	0.17
HF status/diagnosis	1749		741		2490		0.26
High-risk condition		472 (27.0%)		222 (30.0%)		694 (27.9%)	
Dilated cardiomyopathy		493 (28.2%)		190 (25.6%)		683 (27.4%)	
Ischaemic cardiomyopathy		287 (16.4%)		127 (17.1%)		414 (16.7%)	
Other aetiologies		295 (16.9%)		108 (14.6%)		403 (16.2%)	
Unknown		202 (11.6%)		94 (12.7%)		296 (11.9%)	
History of atrial fibrillation	1745	281 (16.1%)	738	121 (16.4%)	2483	402 (16.2%)	0.86
Chronic renal disease	1744	124 (7.1%)	738	44 (6.0%)	2482	168 (6.8%)	0.34
Diabetes	1744	235 (13.5%)	738	93 (12.6%)	2482	328 (13.2%)	0.60
Hypertension	1742	476 (27.3%)	737	181 (24.6%)	2479	657 (26.5%)	0.16
Previous malignancy	1747	121 (6.9%)	738	57 (7.7%)	2485	178 (7.2%)	0.50
Smoking status	1742		737		2479		
Current		191 (11.0%)		73 (9.9%)		264 (10.7%)	0.48
Former		337 (19.4%)		174 (23.6%)		511 (20.6%)	0.02
Body mass index (kg/m^2)	1746	27.2 ± 5.8	739	27.3 ± 5.8	2485	27.2 ± 5.8	0.80
Cardiac status							
Ejection fraction (%)	1580	44 ± 16	667	45 ± 16	2247	44 ± 16	0.26
≥50%		762 (48.2%)		345 (51.7%)		1140 (50.7%)	0.14
NYHA class	1749		741		2490		0.35
I		664 (38.0%)		290 (39.1%)		954 (38.3%)	
II		334 (19.1%)		142 (19.2%)		476 (19.1%)	
III		224 (12.8%)		79 (10.7%)		303 (12.2%)	
IV		25 (1.4%)		15 (2.0%)		40 (1.4%)	
Not documented		502 (28.7%)		215 (29.0%)		717 (28.8%)	
Conduction abnormalities	1482	457 (30.8%)	623	178 (28.6%)	2105	635 (30.2%)	0.32
Current atrial fibrillation/flutter	1535	141 (9.2%)	652	73 (11.2%)	2187	214 (9.8%)	0.16
Pacemaker/resynchronization device	1749	212 (12.1%)	741	112 (15.1%)	2490	324 (13.0%)	0.04
ICD	1749	328 (18.8%)	741	127 (17.1%)	2490	455 (18.3%)	0.36
QRS duration (ms)	1319	125 ± 34	562	127 ± 38	1881	126 ± 35	0.16
Heart rate (beats/min)	1603	70 ± 14	689	71 ± 13	2292	70 ± 13	0.43
Medications							
ACE inhibitors (all classes)	1678	757 (45.1%)	697	307 (44.1%)	2375	1064 (44.8%)	0.65
Angiotensin receptor blockers (all classes)	1678	221 (13.2%)	697	72 (10.3%)	2375	293 (12.3%)	0.06
Beta-blockers (all classes)	1678	1048 (62.5%)	698	415 (59.5%)	2376	1048 (62.5%)	0.18
Bisoprolol	1678	352 (21.0%)	698	131 (18.8%)	2376	483 (20.3%)	0.24
Carvedilol	1678	464 (27.7%)	698	178 (25.5%)	2376	642 (27.0%)	0.29
Metoprolol	1678	209 (12.5%)	698	97 (13.9%)	2376	306 (12.9%)	0.35
Aldosterone receptor antagonist (MRA)	1676	534 (31.9%)	697	218 (31.3%)	2373	752 (31.7%)	0.81
Antiarrhythmics (all classes)	1678	434 (25.9%)	697	176 (25.3%)	2375	610 (25.7%)	0.80
Digoxin	1677	296 (17.7%)	697	118 (16.9%)	2374	414 (17.4%)	0.72
Anticoagulants (all classes)	1676	465 (27.7%)	697	190 (27.3%)	2373	655 (27.6%)	0.84
Diuretics (all classes)	1678	671 (40.0%)	697	271 (38.9%)	2375	942 (39.7%)	0.65
Thiazide	1674	75 (4.5%)	697	32 (4.6%)	2371	107 (4.5%)	0.91
Loop diuretics (all classes)	1674	641 (38.3%)	696	257 (36.9%)	2370	898 (37.9%)	0.55
Furosemide	1674	633 (37.8%)	695	255 (36.7%)	2369	888 (37.5%)	0.64
Lipid lowering medications (all classes)	1678	525 (31.3%)	697	228 (32.7%)	2375	753 (31.7%)	0.50
Platelet inhibitors	1678	504 (30.0%)	698	204 (29.2%)	2376	708 (29.8%)	0.73

Continued

Table 1 ***Continued***

	N	Training cohort	*N*	Validation cohort	*N*	All patients	*P*
Laboratory investigations							
BNP (pg/mL)	1077	103 (31–326)	431	97 (32–282)	1508	101 (31–315)	0.77
White blood cell count (×10^9 cells/L)	1178	7.3 ± 2.3	470	7.4 ± 2.5	1648	7.3 ± 2.4	0.29
Basophils (×10^9 cells/L)	1171	0.03 (0.01–0.06)	470	0.04 (0.01–0.06)	1641	0.03 (0.01–0.06)	0.10
Eosinophils (×10^9 cells/L)	1173	0.15 (0.01–0.24)	470	0.16 (0.10–0.25)	1643	0.15 (0.10–0.24)	0.64
Lymphocytes (×10^9 cells/L)	1169	1.75 (1.30–2.20)	470	1.63 (1.29–2.24)	1639	1.72 (1.30–2.21)	0.71
Monocytes (×10^9 cells/L)	1173	0.56 (0.44–0.70)	470	0.59 (0.47–0.73)	1643	0.57 (0.45–0.71)	0.08
Neutrophils (×10^9 cells/L)	1173	4.31 (3.40–5.56)	470	4.43 (3.45–5.69)	1643	4.34 (3.40–5.61)	0.21
Haematocrit	1178	0.42 ± 0.05	471	0.42 ± 0.05	1649	0.42 ± 0.05	0.74
Haemoglobin (g/L)	1175	142 ± 17	471	142 ± 17	1646	142 ± 17	0.82
Platelet count (×10^9 cells/L)	1175	218 ± 65	467	221 ± 72	1642	219 ± 67	0.45
Red blood cell count (×10^{12} cells/L)	1173	4.7 ± 0.6	471	4.7 ± 0.7	1644	4.7 ± 0.6	0.66
Chloride (mmol/L)	1160	103 ± 4	468	103 ± 4	1628	103 ± 4	0.99
Potassium	1183	4.2 ± 0.4	480	4.2 ± 0.4	1663	4.2 ± 0.4	0.29
Sodium (mmol/L)	1183	139 ± 4	482	138 ± 4	1665	138 ± 4	0.36
Serum creatinine (umol/L)	1184	99 ± 69	488	96 ± 43	1672	98 ± 62	0.28
Glomerular filtration rate (mL/min/1.73 m^2)	1172	77 ± 24	484	77 ± 24	1656	77 ± 24	0.93
Outcome							
Combined outcome	1749	226 (12.9%)	741	97 (13.1%)	2490	323 (13.0%)	0.90
Mechanical circulatory support		53 (3.0%)		14 (1.9%)		67 (2.7%)	0.14
Heart transplantation		57 (3.3%)		26 (3.5%)		83 (3.3%)	0.81
Death		116 (6.6%)		57 (7.7%)		173 (7.0%)	0.35
Duration of follow-up (months)	1749	56.0 ± 33.3	741	55.6 ± 32.9	2490	55.9 ± 33.2	0.78

Data reported as means ± standard deviations, medians with interquartile range or frequencies as appropriate.

ACC, American College of Cardiology; AHA, American Heart Association; BNP, beta-natriuretic peptide; ICD, implantable cardioverter defibrillators; MRA, mineralocorticoid receptor antagonist.

along with test probabilities were substituted in the regression equation. The new threshold was then chosen in such a way that the same proportion of positive and negative cases (compared with training cohort) was found using the calibrated probabilities.

Data analyses

Data are described using means with standard deviations, median with 25th and 75th percentiles, and frequencies as appropriate. Comparisons between the training and validation sets were performed using Student's *t*-test assuming unequal variance between groups and Fisher's exact test. All performance and calibration metrics are reported separately for the training and validation sets. Feature importance was calculated by taking coefficients of a ridge regression model fitted on the data samples and their predicted survival probability into consideration. A scree plot was used to illustrate the ranking of features by importance separated between clinical markers and classic CPET indices vs. advanced CPET indices based on the breath-by-breath analysis. Finally, the performance of the prediction model in various subsets of patients in the validation set was evaluated. All analyses were performed using R 3.5.3 and Python 3.6.9.

Results

A total of 2490 patients were included in this analysis of which 741 (30%) were randomly segregated in the validation data set and 1749 (70%) were used for model training and internal cross-validation. Patient characteristic and exercise test results at baseline and incidence of outcomes over time were similar between the training and validation data sets (*Tables 1* and *2*).

Model performance metrics for both the training and the validation data sets are reported in *Table 3* with the corresponding area under the curve (AUCs) reported in *Figure 1* and predicted vs. actual freedom from the composite endpoint reported in *Figure 2*. In the validation data set, the AUC of the prediction model was 0.87. We explored more comprehensive model performance metrics using three potential cut-off points: (i) to maximize raw accuracy, (ii) to match the prevalence of outcome in both the training and hold out set, and (iii) to maximize sensitivity and specificity. The decision point based on maximizing accuracy had a sensitivity of 0.58 and a specificity of 0.94. Matching the prevalence of outcomes in the training set did not substantially affect accuracy (90% vs. 91%), sensitivity (0.58), or specificity (0.94). However, using a decision point maximizing specificity and sensitivity reduced overall accuracy (78%) and specificity (0.80) but substantially increased sensitivity to 0.72. Performance metrics in the training data set were marginally higher (AUC of 0.93), but the difference in effect did not suggest overfitting. There was substantial concordance between actual vs. predicted freedom from the composite endpoint in the training and validation data sets.

An examination of the scree plot of coefficient of variable importance (*Figure 3*) shows that the variables with the highest importance to generate predictions were, in descending order of importance, minute ventilation/carbon dioxide production ratio, minute ventilation/oxygen intake ratio and expiration volume, and finally heart rate recovery. Clinical features of high importance for the prediction model, in

Table 2 Cardiopulmonary parameters stratified by training vs. validation cohorts

	N	Training cohort	*N*	Validation cohort	*N*	All patients	*P*
Systolic blood pressure at rest (mmHg)	1744	115 (104–126)	739	114 (105–125)	2483	115 (104–126)	0.48
Diastolic blood pressure at rest (mmHg)	1748	72 (66–79)	739	72 (66–79)	2487	72 (66–79)	0.41
Heart rate at rest (b.p.m.)	1746	70 (61–78)	740	71 (62–80)	2486	70 (61–79)	0.10
O_2 saturation at rest (%)	1745	98 (98–99)	740	99 (98–99)	2485	98 (98–99)	0.58
Forced vital capacity at rest (L)	1729	3.4 (2.8–4.2)	728	3.4 (2.7–4.1)	2457	3.4 (2.7–4.2)	0.29
PP forced vital capacity at rest (%)	1729	79 (66–90)	728	79 (67–89)	2457	79 (67–90)	0.43
Forced expiratory capacity at rest (L)	1726	2.74 (2.19–3.36)	727	2.69 (2.15–3.31)	2453	2.73 (2.17–3.34)	0.29
PP forced expiratory capacity at rest (%)	1729	80 (67–92)	729	80 (67–91)	2458	80 (67–91)	0.69
Peak systolic blood pressure (mmHg)	1745	146 (130–162)	737	144 (127–166)	2482	145 (128–163)	0.90
Peak diastolic blood pressure (mmHg)	1745	77 (70–80)	737	78 (70–82)	2482	78 (70–80)	0.75
Peak heart rate (b.p.m.)	1747	125 (102–150)	738	122 (100–151)	2485	123 (102–150)	0.58
Heart rate—1 min post peak (b.p.m.)	1742	101 (85–122)	738	101 (83–127)	2480	101 (84–123)	0.90
Heart rate recovery 1 min (b.p.m.)	1746	21 (13–29)	740	20 (13–28)	2486	20 (13–28)	0.11
O_2 saturation at peak (%)	1746	98 (98–99)	739	98 (98–99)	2485	98 (98–99)	0.56
Exercise time (s)	1744	603 (453–781)	736	584 (422–760)	2480	597 (440–774)	0.18
Workload (watts)	1701	90 (70–122)	722	90 (60–120)	2423	90 (70–120)	0.08
PP workload (%)	1698	62 (47–78)	719	61 (47–76)	2417	61 (47–78)	0.24
Peak indexed VO_2 (mL/kg/min)	1748	17.0 (13.0–22.0)	740	16.2 (12.6–21.9)	2488	16.7 (12.8–22.0)	0.34
PP peak indexed VO_2 (%)	1697	58 (46–72)	721	57 (46–71)	2418	58 (46–72)	0.40
Peak VO_2 (L/min)	1747	1.33 (1.02–1.76)	739	1.28 (0.98–1.71)	2486	1.35 (1.01–1.75)	018
PP peak VO_2 (%)	1749	61 (50–75)	741	60 (49–74)	2490	61 (50–75)	0.50
Peak ventilation (L/min)	1749	46.1 (35.4–58.8)	741	45.8 (34.1–57.5)	2490	46.0 (35.1–58.7)	0.63
Peak VCO_2 (L/min)	1744	1.45 (1.10–1.94)	740	1.42 (1.05–1.92)	2484	1.45 (1.08–1.94)	0.18
VE/VCO_2 slope	1576	31 (27–35)	677	31 (28–35)	2253	31 (27–35)	0.28
VE/VCO_2 at anaerobic threshold	1695	30 (27–34)	717	30 (27–34)	2412	30 (27–34)	0.16
Anaerobic threshold (mL/kg/min)	1693	10.8 (8.5–13.8)	710	10.7 (8.4–13.7)	2403	10.7 (8.5–13.8)	0.98
Per cent of peak VO_2 at AT (%)	1688	63 (58–69)	708	64 (59–69)	2396	64 (58–69)	0.17
PP of peak VO_2 at AT (%)	1693	38 (30–46)	710	37 (31–45)	2403	37 (30–46)	0.79
Peak respiratory exchange ratio	1746	1.10 (1.03–1.16)	740	1.09 (1.03–1.16)	2486	1.10 (1.03–1.16)	0.34
End-tidal partial pressure of CO_2 (mmHg)	736	35 (32–39)	312	36 (33–39)	1048	35 (32–39)	0.67
Oxygen uptake efficiency slope	737	1.60 (1.20–2.01)	310	1.63 (1.23–2.02)	1047	1.61 (1.21–2.02)	0.81

AT, anaerobic threshold; CO_2, carbon dioxide; O_2, oxygen; PP, per cent predicted; VCO_2, carbon dioxide production; VE, ventilation; VO_2, oxygen consumption.

decreasing order of importance, were body mass index (BMI), the use of diuretics, presence of ICD, use of antiarrhythmic medications, worsening NYHA class, blood urea, leucocyte count, the presence of atrial fibrillation, and the use of angiotensin receptor blockers. Many engineered mathematical features derived from the breath-by-breath measurements were highly ranked features confirming their prognostic value in patients with HF.

Calibration of the prediction model (*Figure 4*) was excellent in the training data set (average absolute difference of 0.6%) and remained very good in the validation data set (average absolute difference of 3.2%). In the validation data set, the prediction model slightly underestimated risk of outcomes in the 6th, 7th, and 10th decile of risk, but the magnitude of the differences is unlikely to be clinically important. Finally, *Figure 5* reports model AUC in various subgroups of patients in the validation cohort. The results show that model AUC in all subgroups assessed remained above 0.75, with the exception of patients with ischaemic cardiomyopathy for whom the model AUC only reached ~0.70.

Discussion

In this study, we have shown that by using a deep survival network, we could effectively incorporate breath-by-breath data generated during CPET to improve the accuracy of prediction model for a composite endpoint of HF outcomes defined as death, need for heart transplantation, or mechanical circulatory support. Our final prediction model was able to accurately predict patients who are at high risk of the composite outcome over a 10-year period. It is important to note, however, that the prediction model is built so that it is capable of predicting risk of outcome for any horizon up to 10 years. Therefore, an end-user could select a short (1–2 years) prediction horizon for sick/elderly patients and a longer (5–10 years for younger high-risk patients).

Although not directly comparable because of the inclusion of high-risk patients, performance metrics of this new model was excellent and above the usual threshold for it to be used clinically for the prognostication of patients with HF.[13] Furthermore, the underlying hazard function of the prediction model could be used to generate predictions over one

Table 3 Comparison of model performance metrics

	Optimized for highest accuracy	Optimized to match prevalence	Optimized for highest Sn/Sp
Training cohort			
AUC: 0.928 (0.008)			
Cut-off probability	0.573	0.383	0.154
Accuracy (%)	0.91	0.90	0.85
Sensitivity	0.77	0.63	0.85
Specificity	0.93	0.95	0.85
False positive rate	0.07	0.05	0.15
False negative rate	0.23	0.37	0.15
Validation cohort			
AUC: 0.865 (0.021)			
Cut-off probability	0.408	0.399	0.096
Accuracy (%)	0.89	0.89	0.78
Sensitivity	0.58	0.58	0.72
Specificity	0.94	0.94	0.80
False positive rate	0.06	0.06	0.21
False negative rate	0.42	0.42	0.28

AUC, area under the curve; Sn, sensitivity; Sp, specificity.

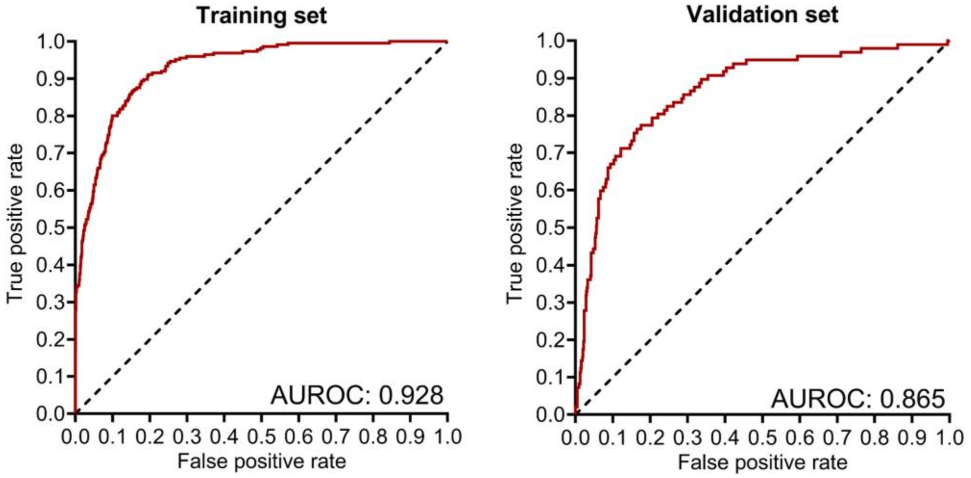


Figure 1 Area under the curve (AUC) for prediction models in the training vs. validation cohorts.

or more specific horizons, thus further improving utility compared with traditional models which predict the outcome at a single point in the future (i.e. survival at *x* years). This study is novel in two respects. First, it is the first clinical prediction model for HF outcomes to integrate mathematical features derived from breath-by-breath data generated during CPETs as opposed to relying on classic summary indices, which are easier to obtain but likely less informative. Second, it uses a novel deep learning framework which integrates survival analysis as opposed to relying on a binary, time-delimited outcome. Both of these characteristics are important advances as they open the door for future prediction models in the field of HF that are both better performing and can more fully use the large quantity of data that are generated in the care of patients with HF, particularly diagnostic investigations such as laboratory values and cardiac imaging.

Previous studies have shown the utility of using machine learning for the long-term prognosis of patients with HF. Myers *et al.*[10] showed that the use of neural networks marginally increased the performance of predictive models generated from summary CPET data over LR. In this study, the use of an artificial neural network to predict death from cardiac mortality in patients with HF resulted in an increase in AUC from 0.70 with LR to 0.72 when five classic CPET summary indices were used as predictors. A recent review of machine learning–based prediction models developed for HF showed that the majority of previous attempts has used the strictly binary, time-restricted, confines that are necessary for most classic supervised machine learning algorithms.[26] In the case of time-to-event outcomes, this strategy requires the creation of a time-landmarked version of the outcome

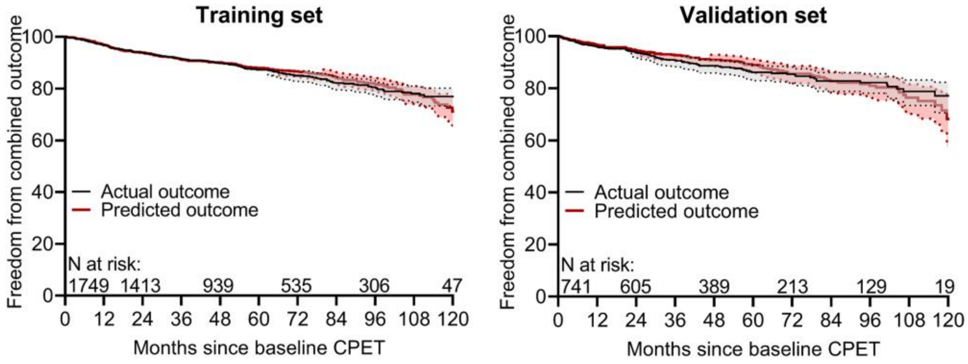


Figure 2 Actual vs. predicted freedom from the combined heart failure outcome in the training vs. validation cohorts. CPET, cardiopulmonary exercise test.

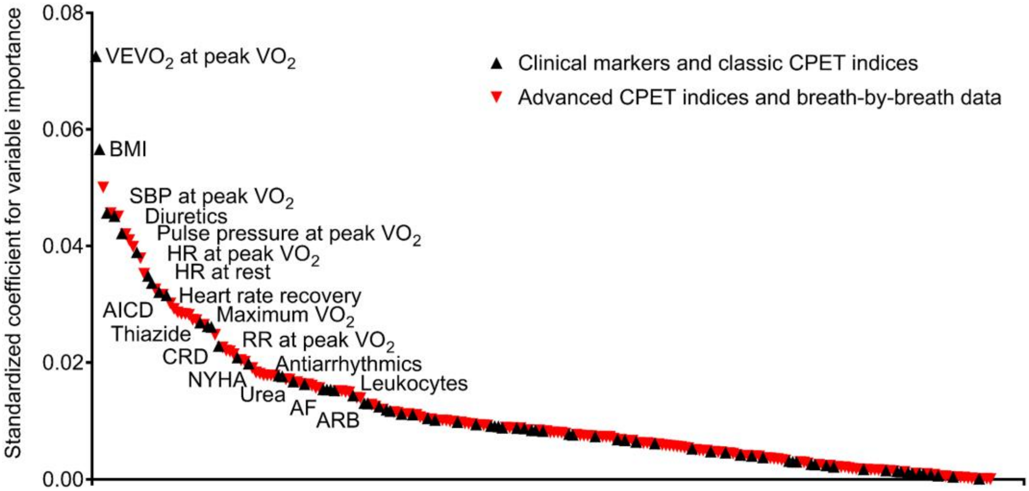


Figure 3 Feature importance ranking for prediction model for combined outcome. Abbreviations: AF, atrial fibrillation; AICD, automatic implantable cardioverter defibrillators; ARB, angiotensin receptor blockers; BMI, body mass index; CRT, cardiac resynchronization therapy; HR, heart rate; CPET, cardiopulmonary exercise test; NYHA, New York heart association; max, maximum; SBP, systolic blood pressure; VE, ventilation; VO$_2$, volume oxygen.

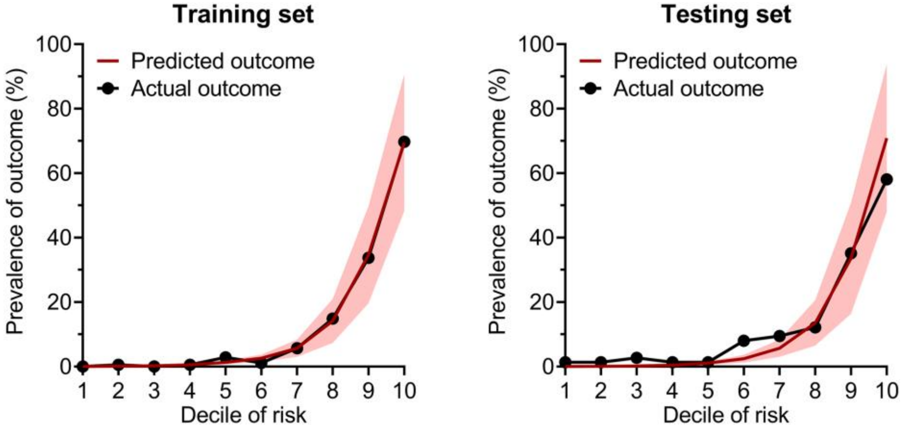


Figure 4 Calibration curves in the training vs. validation cohorts.

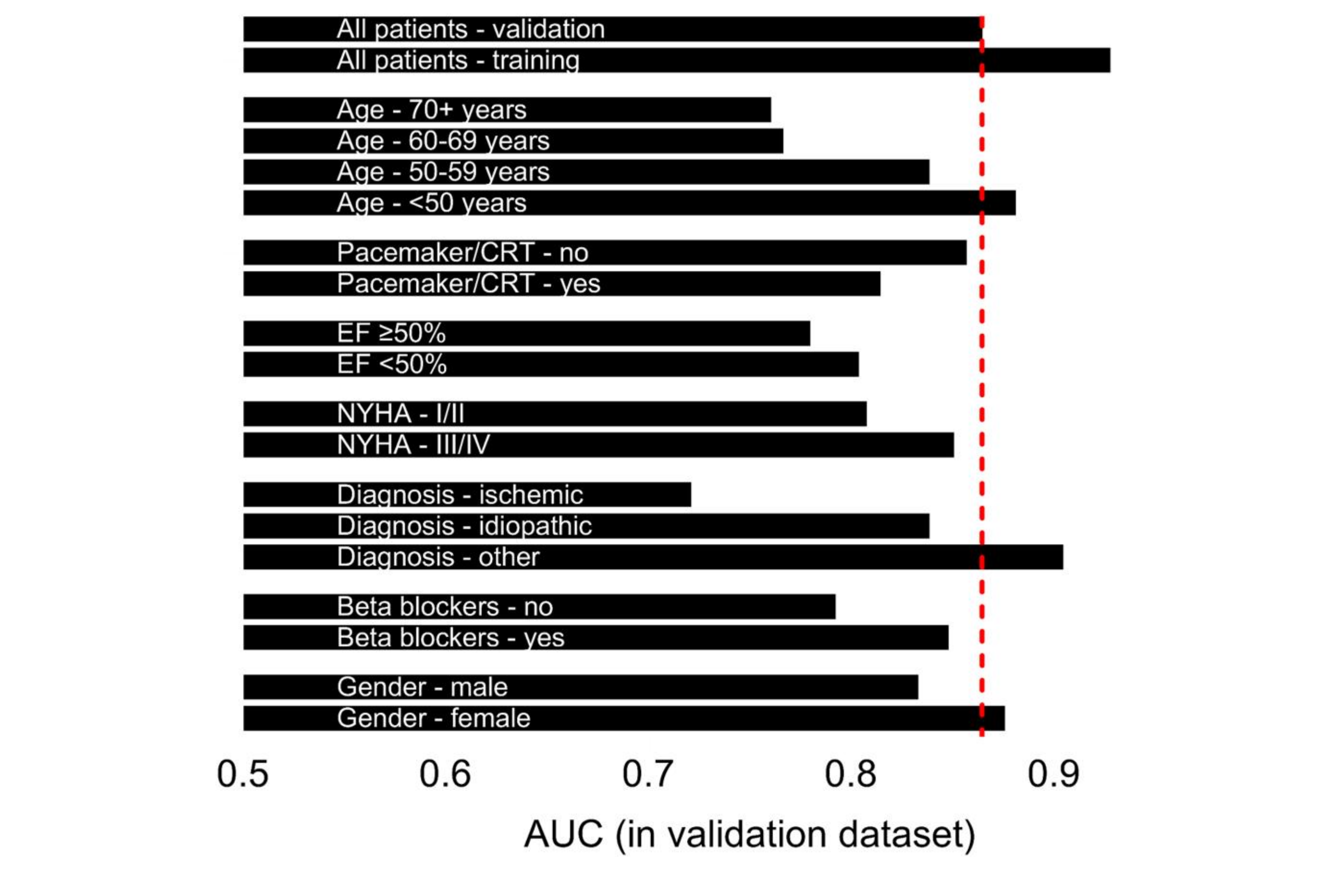


Figure 5 Model performance metrics in various patient subgroups in the validation cohort.

(e.g. outcomes at *x* years after time 0) as opposed to using the right censoring methodology which is normally used for such outcomes. This strategy results in a loss of information (and often a reduction in sample size) but allows the use of supervised machine learning models in this context. Random survival forest has been used in studies,[27,28] and this method is also not ideal as it can only approximate the framework of survival analysis and as such still present substantial limitations. This is why the use of DeepSurv,[15] as implemented in this study, is novel and a substantial advance given that it does not require the use of an alternative or approximation to survival models for the prediction of long-term outcomes. Moreover, it enables the use of complex predictive features for prediction through deep learning.

Predicting outcomes in patients with HF, either through classic probabilistic models or more recently through machine learning, has historically been a challenge given the heterogeneity of the patient population, the complex interrelation between numerous risk factors, the large spectrum of clinical severity, and varying treatments received.[26,29,30] Studies using machine learning algorithms over conventional methods have shown slightly better performance in predicting mortality and hospitalization in HF patients.[31] However, the extent to which these minor improvements in performance further improve clinical prediction remains uncertain. Generally speaking, risk factors for adverse outcomes in HF patients include age, functional class, ejection fraction, BMI, blood pressure, heart rate, renal and liver function, and natriuretic peptide levels.[32]

Summary indices from CPETs have been associated with outcomes in patients with HF but historically have only marginally improved the performance of prediction models over those including only clinical features.[3,33] However, recently, we showed a substantial improvement in prediction of 1-year adverse outcomes in patients with HF when mathematical features derived from breath-by-breath data were included in a neural network over summary CPET indices and basic clinical data.[14] In the current study, we have further demonstrated that these complex features can be integrated in clinical prediction models for long-term outcomes in survival analysis through deep learning.

Historically, the majority of prediction models developed for HF has shown AUCs in the low to mid 0.70 s.[13,34,35] Many of the most common risk scores for mortality in patients with chronic HF fall in this category; this includes models such as the Seattle Heart Failure (AUC = 0.73),[36] CORONA (AUC = 0.72),[37] MAGGIC (AUC = 0.74),[38] and CHARMS (AUC = 0.75)[39] models. These models all use combinations of clinical data, medical history, and echocardiography, but not exercise testing, to predict medium-term mortality. Two additional models integrated these features and added the results of exercise testing with marginal improvement: HF-ACTION (AUC = 0.73)[40] and the MECKI score (AUC = 0.76–0.80 over 1–4 year horizons).[41] In a recent review of 40 prediction models developed in patients with HF between 2013 and 2018, only 15% of models reached an AUC between 0.80 and 0.85 in external validation cohorts and none reached an AUC above 0.85.[42] When considering only prediction models using a composite outcome such as the one in this study, only 1 of 13 models reached an AUC above 0.80.[42] As such, the algorithm presented here represents a substantial improvement with an AUC of 0.87 in the validation data set [mild HF (NYHA I/II): 0.81, severe HF (NYHA III/IV): 0.85]. Nevertheless, it is worth noting that our population represents a younger age group and as such comparisons with other prediction models developed on different HF populations might not be entirely accurate.

There are a number of important technical considerations about our algorithm that should be mentioned. The stratified performance information shows that performance was maintained in all subgroups of patients, including various diagnoses, patients with abnormal or paced rhythm, and patients on beta-blockers. The lower AUC in patients with ischaemic cardiomyopathy is likely a reflection of the smaller sample size and the higher event rate (26%) in this group than in the other groups (18%), suggesting that group-specific segmentation of the algorithm might be necessary in future iterations. Traditionally, patients with abnormal or paced heart rhythm represent a challenge to the integration of some CPET indices in HF prediction models and have either been excluded or considered separately in this context. The fact that they could be included in the current algorithm without diminishing performance is an important improvement over previous studies. The ability of the algorithm to handle a heterogeneous patient population is also evidenced by the generally consistent performance across diagnoses, albeit with a small reduction in performance for patients with ischaemic cardiomyopathy. Patients with congenital heart disease were excluded from the study for logistic reasons and require future analysis.

We elected to train the model using patients with all stages of HF. This approach had several advantages; first, it increased the size of the training data set and provided a good number of training cases with a 'normal' exercise response, something that would be rare in patients with advanced HF, thus preventing training bias towards sicker patients. Second, this strategy allows the model to be used for all CPET indications (diagnostic confirmation, monitoring of high-risk but stable patients, and prognostication for patients with advanced disease). We were able to demonstrate that model performance was preserved across NYHA functional classes, thus confirming that our strategy of including all patients regardless of stage of HF did not come to the detriment of either ends of the spectrum.

For this study to be feasible, we used a composite outcome of death, cardiac transplantation, or mechanical circulatory support although previous studies have shown that models focusing on a single outcome in patients with HF tend to perform better.[42] It is expected that future iterations of this algorithm will be able to separate each outcome and consider them distinctly and that this strategy will result in improved algorithm performance. Finally, the calibration of the algorithm shows clear concordance between predicted and actual risk of outcomes. Thus, rather than using a single cut-off point to predict a binary outcome, the accurate calibration suggests that the expected probability of adverse outcomes can be used to guide clinical care and, therefore, to improve clinical utility.

This study should be considered in light of some limitations. First, it is a single-centre study with a retrospective design and a younger patient population; thus, we cannot fully establish the generalizability of our findings or extrapolate future performance in an external validation cohort. Second, data regarding the diagnosis of patients in regard to reduced vs. preserved ejection fraction are not available for the patients included in this study; as such, we were not able to assess the event prevalence, contribution to the prediction model, and performance of the prediction in patients with preserved vs. reduced ejection fraction. Third, given that heart transplantation is included in the composite outcome and that exercise testing is one of the indications for heart transplantation, the model performance might be overestimated because of target leaking; however, this is a common problem for all such models in patients with HF.

While the algorithm development is still at the prototype stage, future versions of this algorithm could be deployed through an application programming interface integrated within the user interface and reporting system of standard CPET systems, thus facilitating their use by clinicians despite the complex underlying computational infrastructure needed to execute the algorithm.[43] It is important to note that, while complete breath-by-breath data are not currently routinely stored by most CPET systems, the changes needed for this algorithm to be available in other institutions are minimal. Cardiopulmonary exercise testing system needs to be modified to standardize file naming and storage location and for the source data to be mapped to the algorithm's input format. Those changes can easily be done by information technology staff at local sites as part of the routine configuration and maintenance of the CPET systems.

In conclusion, using a survival model integrated in a deep learning framework, we were able to create a prediction model for a composite endpoint (death, heart transplant, or mechanical circulatory support) in patients with HF that incorporated clinical data, classic summary indices from CPETs, and mathematical features derived from the breath-by-breath data generated during CPETs. Model performance was characterized by high discrimination with excellent calibration. This level of performance is superior to other similar models for HF that have been previously published and indicates a high potential for clinical utility in future iterations.

Funding

This study was funded by the Ted Rogers Centre for Heart Research, University Health Network, Toronto, Ontario, Canada. The study funder had no involvement in study design, data analysis, or in the decision to publish the results of this study.

Conflict of interest: none declared.

Data availability

Data used in this study contain confidential health information, and as such, under the Ontario Personal Health Information Protection Act (PHIPA), there are legal restrictions on disclosure and distribution of these data, even in an anonymized format. Data used in this study can be accessed by qualified researchers who meet the criteria for access to confidential health information. In addition to contacting the principal investigator to access the data, requestors will be required to obtain approval from the Research Ethics Board at University Health Network.

References

1. Braunwald E. The war against heart failure: the Lancet lecture. *Lancet* 2015;**385**: 812–824.
2. Mirkin B, Weinberger M. The demography of population ageing. *Popul Bullet UN* 2001; **42**:41–48.
3. Alba AC, Adamson MW, MacIsaac J, Lalonde SD, Chan WS, Delgado DH, *et al.* The added value of exercise variables in heart failure prognosis. *J Card Fail* 2016;**22**:492–497.
4. Alba AC, Walter SD, Guyatt GH, Levy WC, Fang J, Ross HJ, *et al.* Predicting survival in patients with heart failure with an implantable cardioverter defibrillator: the heart failure meta-score. *J Card Fail* 2018;**24**:735–745.
5. Buchan TA, Ching C, Foroutan F, Malik A, Daza JF, Hing NNF, *et al.* Prognostic value of natriuretic peptides in heart failure: systematic review and meta-analysis. *Heart Fail Rev* 2021;**27:**2022, 645–654.
6. Guazzi M, Dickstein K, Vicenzi M, Arena R. Six-minute walk test and cardiopulmonary exercise testing in patients with chronic heart failure: a comparative analysis on clinical and prognostic insights. *Circ Heart Fail* 2009;**2**:549–555.
7. Milani RV, Lavie CJ, Mehra MR, Ventura HO. Understanding the basics of cardiopulmonary exercise testing. *Mayo Clin Proc* 2006;**81**:1603–1611.
8. Corra U, Piepoli MF, Adamopoulos S, Agostoni P, Coats AJ, Conraads V, *et al.* Cardiopulmonary exercise testing in systolic heart failure in 2014: the evolving prognostic role: a position paper from the Committee on Exercise Physiology and Training of the Heart Failure Association of the ESC. *Eur J Heart Fail* 2014;**16**:929–941.
9. Myers J, Arena R, Dewey F, Bensimhon D, Abella J, Hsu L, *et al.* A cardiopulmonary exercise testing score for predicting outcomes in patients with heart failure. *Am Heart J* 2008;**156**:1177–1183.
10. Myers J, de Souza CR, Borghi-Silva A, Guazzi M, Chase P, Bensimhon D, *et al.* A neural network approach to predicting outcomes in heart failure using cardiopulmonary exercise testing. *Int J Cardiol* 2014;**171**:265–269.

11. Metra M, Faggiano P, D'Aloia A, Nodari S, Gualeni A, Raccagni D, *et al.* Use of cardiopulmonary exercise testing with hemodynamic monitoring in the prognostic assessment of ambulatory patients with chronic heart failure. *J Am Coll Cardiol* 1999;**33**:943–950.
12. Aaronson KD, Schwartz JS, Chen TM, Wong KL, Goin JE, Mancini DM. Development and prospective validation of a clinical index to predict survival in ambulatory patients referred for cardiac transplant evaluation. *Circulation* 1997;**95**:2660–2667.
13. Alba AC, Agoritsas T, Jankowski M, Courvoisier D, Walter SD, Guyatt GH, *et al.* Risk prediction models for mortality in ambulatory patients with heart failure: a systematic review. *Circ Heart Fail* 2013;**6**:881–889.
14. Hearn J, Ross HJ, Mueller B, Fan CP, Crowdy E, Duhamel J, *et al.* Neural networks for prognostication of patients with heart failure. *Circ Heart Fail* 2018;**11**:e005193.
15. Katzman JL, Shaham U, Cloninger A, Bates J, Jiang T, Kluger Y. DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC Med Res Methodol* 2018;**18**:24.
16. Gibbons RJ, Balady GJ, Bricker JT, Chaitman BR, Fletcher GF, Froelicher VF, *et al.* ACC/AHA 2002 guideline update for exercise testing: summary article: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines (Committee to Update the 1997 Exercise Testing Guidelines). *Circulation* 2002;**106**:1883–1892.
17. Abraham WT, Psotka MA, Fiuzat M, Filippatos G, Lindenfeld J, Mehran R, *et al.* Standardized definitions for evaluation of heart failure therapies: scientific expert panel from the Heart Failure Collaboratory and Academic Research Consortium. *JACC Heart Fail* 2020;**8**:961–972.
18. Christ M, Braun N, Neuffer J, Kempa-Liehr AW. Time series feature extraction on basis of scalable hypothesis tests (tsfresh—a Python package). *Neurocomputing* 2018;**307**:72–77.
19. Azur MJ, Stuart EA, Frangakis C, Leaf PJ. Multiple imputation by chained equations: what is it and how does it work? *Int J Methods Psychiatr Res* 2011;**20**:40–49.
20. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. *J Mach Learn Res* 2014;**15**:1929–1958.
21. Klambauer G, Unterthiner T, Mayr A, Hochreiter S. Self-Normalizing Neural Networks. Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17), Long Beach, 4-9 December 2017, 972–981. arXiv:1706.02515. 2017.
22. Kingma D, Adam BJ. Adam: a method for stochastic optimization. International Conference on Learning Representation. San Diego, May 2015, 7–9. arXiv:1412.6980. 2015.
23. Nesterov Y. Gradient methods for minimizing composite functions. *Math Program* 2013;**140**:125–161.
24. Senior A, Heigold G, Ranzato M, Yang K. An empirical study of learning rates in deep neural networks for speech recognition. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013:6724–6728.
25. Elssied NOF, Ibrahim O, Osman A. A novel feature selection based on one-way ANOVA F-test for e-mail spam classification. *Res J Appl Sci Eng Technol* 2014;**7**:625–638.
26. Shin S, Austin PC, Ross HJ, Abdel-Qadir H, Freitas C, Tomlinson G, *et al.* Machine learning vs. conventional statistical models for predicting heart failure readmission and mortality. *ESC Heart Fail* 2021;**8**:106–115.
27. Miao F, Cai YP, Zhang YX, Fan XM, Li Y. Predictive modeling of hospital mortality for patients with heart failure by using an improved random survival forest. *IEEE Access* 2018;**6**:7244–7253.
28. Padhukasahasram B, Reddy CK, Li Y, Lanfear DE. Joint impact of clinical and behavioral variables on the risk of unplanned readmission and death after a heart failure hospitalization. *PLoS One* 2015;**10**:e0129553.
29. Angraal S, Mortazavi BJ, Gupta A, Khera R, Ahmad T, Desai NR, *et al.* Machine learning prediction of mortality and hospitalization in heart failure with preserved ejection fraction. *JACC Heart Fail* 2020;**8**:12–21.
30. 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**:139–147.
31. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *J Clin Epidemiol* 2019;**110**:12–22.
32. Lau K, Malik A, Foroutan F, Buchan TA, Daza JF, Sekercioglu N, *et al.* Resting heart rate as an important predictor of mortality and morbidity in ambulatory patients with heart failure: a systematic review and meta-analysis. *J Card Fail* 2021;**27**:349–363.
33. Dardas T, Li Y, Reed SD, O'Connor CM, Whellan DJ, Ellis SJ, *et al.* Incremental and independent value of cardiopulmonary exercise test measures and the Seattle Heart Failure Model for prediction of risk in patients with heart failure. *J Heart Lung Transplant* 2015;**34**:1017–1023.
34. 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**:435–441.
35. Rahimi K, Bennett D, Conrad N, Williams TM, Basu J, Dwight J, *et al.* Risk prediction in patients with heart failure: a systematic review and analysis. *JACC Heart Fail* 2014;**2**:440–446.
36. 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**:1424–1433.
37. Wedel H, McMurray JJ, Lindberg M, Wikstrand J, Cleland JG, Cornel JH, *et al.* Predictors of fatal and non-fatal outcomes in the Controlled Rosuvastatin Multinational Trial in Heart Failure (CORONA): incremental value of apolipoprotein A-1, high-sensitivity C-reactive peptide and N-terminal pro B-type natriuretic peptide. *Eur J Heart Fail* 2009;**11**:281–291.
38. 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.
39. Pocock SJ, Wang D, Pfeffer MA, Yusuf S, McMurray JJ, Swedberg KB, *et al.* Predictors of mortality and morbidity in patients with chronic heart failure. *Eur Heart J* 2006;**27**:65–75.
40. O'Connor CM, Whellan DJ, Wojdyla D, Leifer E, Clare RM, Ellis SJ, *et al.* Factors related to morbidity and mortality in patients with chronic heart failure with systolic dysfunction: the HF-ACTION predictive risk score model. *Circ Heart Fail* 2012;**5**:63–71.
41. Agostoni P, Corra U, Cattadori G, Veglia F, La Gioia R, Scardovi AB, *et al.* Metabolic exercise test data combined with cardiac and kidney indexes, the MECKI score: a multiparametric approach to heart failure prognosis. *Int J Cardiol* 2013;**167**:2710–2718.
42. Di Tanna GL, Wirtz H, Burrows KL, Globe G. Evaluating risk prediction models for adults with heart failure: a systematic literature review. *PLoS One* 2020;**15**:e0224135.
43. Manlhiot C, van den Eynde J, Kutty S, Ross HJ. A primer on the present state and future prospects for machine learning and artificial intelligence applications in cardiology. *Can J Cardiol* 2022;**38**:169–184.