BMJ Open

Machine learning for prediction of sudden cardiac death in heart failure patients with low left ventricular ejection fraction: study protocol for a retroprospective multicentre registry in China

Fanqi Meng,[1,2] Zhihua Zhang,[1,3] Xiaofeng Hou,[1] Zhiyong Qian,[1] Yao Wang,[1] Yanhong Chen,[4] Yilian Wang,[5] Ye Zhou,[6] Zhen Chen,[7] Xiwen Zhang,[8] Jing Yang,[8] Jinlong Zhang,[9] Jianghong Guo,[10] Kebei Li,[11] Lu Chen,[12] Ruijuan Zhuang,[13] Hai Jiang,[14] Weihua Zhou,[15] Shaowen Tang,[16] Yongyue Wei,[17] Jiangang Zou[1,18]

To cite: Meng F, Zhang Z, Hou X, *et al*. Machine learning for prediction of sudden cardiac death in heart failure patients with low left ventricular ejection fraction: study protocol for a retroprospective multicentre registry in China. *BMJ Open* 2019;**9**:e023724. doi:10.1136/bmjopen-2018-023724

► Prepublication history and additional material for this paper are available online. To view these files, please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2018-023724).

Received 24 April 2018
Revised 4 February 2019
Accepted 13 March 2019

For numbered affiliations see end of article.

Correspondence to
Dr Jiangang Zou;
jgzou@njmu.edu.cn

ABSTRACT

Introduction Left ventricular ejection fraction (LVEF) ≤35%, as current significant implantable cardioverter-defibrillator (ICD) indication for primary prevention of sudden cardiac death (SCD) in heart failure (HF) patients, has been widely recognised to be inefficient. Improvement of patient selection for low LVEF (≤35%) is needed to optimise deployment of ICD. Most of the existing prediction models are not appropriate to identify ICD candidates at high risk of SCD in HF patients with low LVEF. Compared with traditional statistical analysis, machine learning (ML) can employ computer algorithms to identify patterns in large datasets, analyse rules automatically and build both linear and non-linear models in order to make data-driven predictions. This study is aimed to develop and validate new models using ML to improve the prediction of SCD in HF patients with low LVEF.

Methods and analysis We will conduct a retroprospective, multicentre, observational registry of Chinese HF patients with low LVEF. The HF patients with LVEF ≤35% after optimised medication at least 3 months will be enrolled in this study. The primary endpoints are all-cause death and SCD. The secondary endpoints are malignant arrhythmia, sudden cardiac arrest, cardiopulmonary resuscitation and rehospitalisation due to HF. The baseline demographic, clinical, biological, electrophysiological, social and psychological variables will be collected. Both ML and traditional multivariable Cox proportional hazards regression models will be developed and compared in the prediction of SCD. Moreover, the ML model will be validated in a prospective study.

Ethics and dissemination The study protocol has been approved by the Ethics Committee of the First Affiliated Hospital of Nanjing Medical University (2017-SR-06). All results of this study will be published in international peer-reviewed journals and presented at relevant conferences.

Trial registration number ChiCTR-POC-17011842; Pre-results.

Strengths and limitations of this study

- This study is the first multicentre registry study in China, aimed to investigate the feasibility and accuracy of applying machine learning (ML) to predict sudden cardiac death (SCD) in heart failure (HF) patients with low left ventricular ejection fraction (LVEF).
- A broad range of outcomes, including SCD, all-cause death, lethal arrhythmia, sudden cardiac arrest, cardiopulmonary resuscitation and rehospitalisation due to HF, will be evaluated in this study, and the corresponding prognostic models will be developed.
- ML and the traditional multivariable Cox proportional hazards regression model will be derived from the same database and be compared.
- HF patients with LVEF >35% will not be included based on the design of this study, which will restrict the application of the results of this study to the HF with low LVEF.
- It might be difficult to determine the endpoint of this study sometimes for some patients, when dealing with SCD, lethal arrhythmia and sudden cardiac arrest, especially when outside the hospital.

INTRODUCTION

Heart failure (HF) has become a major public health problem with increased prevalence in both Asia and Western countries. The prevalence of HF in Asia is 1.2%–6.7% depending on the population studied.[1] In China, there are 4.2 million HF patients, and 500 000 new cases are being diagnosed each year.[1] Although the survival rate after HF diagnosis has been increased due to improvement in medical therapy, the mortality of HF

remains high. Around 50% of people diagnosed with HF will die within 5 years.[2] The two most common causes of death in patients with HF are sudden cardiac death (SCD) and progressive pump failure. SCD in HF patients is usually caused by lethal arrhythmias such as ventricular tachycardia or ventricular fibrillation, and is reported to be responsible for ~50% of all cardiovascular death in HF patients.[3 4]

The most effective strategy for prevention of SCD in patients with HF is the implantable cardioverter-defibrillator (ICD), associated with 54% relative risk reduction in primary prevention,[5] and 50% relative risk reduction in arrhythmia-related death in secondary prevention.[6] There is a higher risk of SCD in patients with left ventricular ejection fraction (LVEF) ≤35% than with LVEF >35%.[7] At present, LVEF ≤35% is the major ICD indication for primary prevention of SCD.[8] However, real-world data show that only 3%–5% of ICD patients for primary prevention with LVEF ≤35% receive shock therapies on an annual basis,[9] whereas some SCD victims have LVEF >35%.[10 11] Identifying the patients who will be most likely to benefit from primary prevention ICD is urgently needed. Based on the latest literature, LVEF ≤35% is still an independent predictor of all-cause and cardiovascular mortality in chronic systolic HF, and displays a better combination of sensitivity and specificity than 40% cut-off.[12] Finding ways to evaluate the SCD risk in patients with lower EF will be more efficient and economically significant. Furthermore, a noticeable decline in the rate of SCD for HF patients with reduced LVEF has been observed, which was consistent with the cumulative benefit of optimising medication including ACE inhibitor (ACEI) or angiotensin receptor blocker (ARB), beta-blocker and mineralocorticoid receptor antagonist (MRA).[13] Therefore, it is imperative to update the criterion for ICD implantation.

Over the last decade, lots of multivariate prognostic models derived for chronic HF patients have been proposed (table 1).[14–25] However, these models are not appropriate to identify ICD candidates at high risk of SCD in HF patients with low LVEF. Most above prognostic scores were developed form trial databases, and the subjects included various types of HF. There is no specific study for the prognosis of low LVEF population. Additionally, although all the scores are 'not parsimonious', some critical factors are not incorporated into the prognostic models, for example, medications are contained in Irbesartan in Heart Failure With Preserved Ejection Fraction Study (I-PRESERVE),[17] Meta-Analysis Global Group in Chronic Heart Failure (MAGGIC)[21] and Cardiac and Comorbid Conditions HF (3C-HF).[23] Optimised medication was not required as inclusion criteria in all 12 studies. Furthermore, the most above prognostic models are not able to predict SCD risk. In recent years, the advances in strain echocardiography,[26 27] cardiac magnetic resonance[26 27] and cardiac radionuclide imaging[28 29] have provided essential insights into the mechanisms of

Table 1 The risk model for HF in the literature

Author	Database	Year	Variables (n)	Patients (n)	Endpoints
Agostoni[14]	MECKI	2012	6	2716	Cardiovascular death; urgent cardiac transplant
Barlera[15]	GISSI-HF	2013	14	6975	All-cause mortality
Collier[16]	EMPHASIS-HF	2013	10	2737	All-cause mortality
Komajda[17]	I-PRESERVE	2011	12	4128	All-cause mortality
Levy[18]	SHFM	2006	14	1125	Survival
O'Connor[19]	HF-ACTION	2012	4	2331	All-cause mortality
Pocock[20]	CHARM	2006	21	7599	All-cause mortality
Pocock[21]	MAGGIC	2012	13	39372	All-cause mortality
Senni[22]	CVM-HF	2006	13	292	All-cause mortality
Senni[23]	3C-HF	2013	11	2016	All-cause mortality; urgent heart transplant (1 year)
Vazquez[24]	MUSIC	2009	10	992	All-cause mortality; cardiac mortality; pump failure death, sudden death
Uszko-Lencer[25]	BARDICHE-index	2017	8	1811	All-cause mortality; all-cause hospitalisation; CHF-related hospitalisation

BARDICHE, Body mass index (B), Age (A), Resting systolic blood pressure (R), Dyspnea (D), N-terminal pro brain natriuretic peptide (NT-proBNP) (I), Cockroft-Gault equation to estimate glomerular filtration rate (C), resting Heart rate (H), and Exercise performance using 6-min walk test (E); CHARM, the Candesartan in Heart Failure: Assessment of Reduction in Mortality and morbidityj; CVM-HF, CardioVascular Medicine Heart Failure index; EMPHASIS-HF, the Eplerenone in Mild Patients Hospitalization and Survival Study in Heart Failure trial; GISSI-HF, Gruppo Italiano per lo Studio della Streptochinasi nell'Infarto Miocardico-Heart failure Trial; HF, heart failure; HF-ACTION, A Controlled Trial Investigating Outcomes of Exercise TraiNing trial; MECKI, Metabolic exercise test data combined with cardiac and kidney indexes; MUSIC, MUerte Subita en Insuficiencia Cardiaca study; SHFM, the Seattle Heart Failure Model.

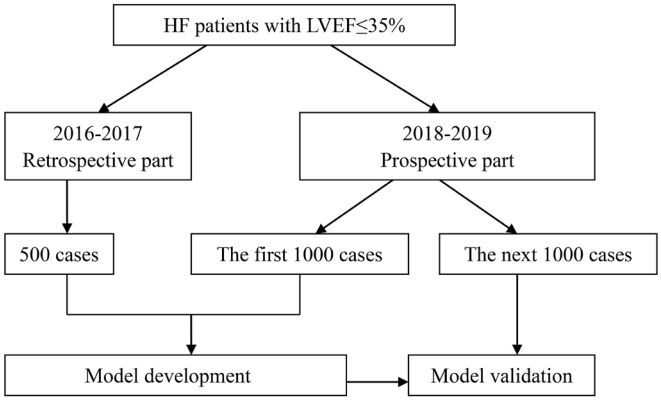

Figure 1 Flow diagram of progress. HF, heart failure; LVEF, left ventricular ejection fraction.

ventricular arrhythmias, and have been recommended to predict the SCD in patients with HF. Although these new methods are effective and non-invasive, the widespread use in large HF population to predict SCD is difficult, due to high equipment and technical requirements. Resting 12-lead ECG and Holter, as the longest surviving, broadly available, quickly deployed and inexpensive tests, can provide a measure of cumulative electrical risk, which may be combined with other factors to improve the SCD risk prediction.[30]

Based on above reasons, the novel risk assessment tools should meet the following requirements: (1) the risk model should be developed from the population with low LVEF (≤35%) to accelerate its clinical application and promote the accuracy of ICD indications for primary prevention. (2) More cardiac and non-cardiac factors beyond LVEF should be included. (3) Electrical risk factors should be included as candidate predictors to evaluate the risk of sudden arrhythmic death. (4) Although sometimes it is not easy to determine the cause of death, SCD as the primary endpoint should be defined whenever possible.

Data processing is the crucial step to develop the prognostic models. This study involves non-linear prediction models, a large number of patients and numerous predictors with complicated correlations. Traditional hypothesis-driven statistical analysis is difficult to overcome these challenges. The machine learning (ML) approaches have great potential to improve the solution. They employ computer algorithms to identify patterns in large datasets with a large number of variables, analyse rules automatically and build both linear and non-linear models in order to make data-driven predictions or decisions.[31] Weng *et al*[32] found that ML significantly improved the accuracy of cardiovascular risk prediction, increased the number of patients who could benefit from preventive treatment and avoided unnecessary treatment. Recent studies have shown that the application of ML techniques may have the potential to improve HF outcomes and management, including cost savings by improving existing diagnostic and treatment support systems.[33] ML algorithms also have been applied to predict SCD in some recent studies and results indicate their significant advantages for predicting SCD.[34 35] However, more studies based on large-scale cohort are needed to evaluate ML for prediction of SCD in HF patients. Therefore, the application of ML for the prediction of SCD in HF patients with low LVEF is technically innovative and clinically significant.

AIMS

The purpose of our study is to develop and validate new models to improve the prediction of SCD in HF patients with low LVEF. The new strategies of identifying HF patients most likely to benefit from primary prevention ICD will improve the revolution of ICD indications. The specific research objective is to develop prediction models to evaluate prognosis and SCD risk, respectively, by ML methods and traditional Cox proportional hazard regression in HF patients with low LVEF (≤35%).

METHODS AND ANALYSIS

Study design

This study is a retroprospective, multicentre, non-interventional, observational clinical registry. The primary sponsor is The First Affiliated Hospital of Nanjing Medical University. The study will be conducted across 14 cardiovascular departments in tertiary A hospitals throughout the People's Republic of China (see online supplementary file 1).

The cases from January 2016 to December 2017 in the First Affiliated Hospital of Nanjing Medical University and Xiamen Cardiovascular Hospital Xiamen University will be collected retrospectively and followed-up prospectively. About 500 retrospective cases meet the inclusion criteria according to preliminary estimation. The prospective recruitment has started in the above 14 hospitals since January 2018. The retrospective cases and the first 1000 prospective cases will be used to develop the prediction models. And the next 1000 prospective cases will be used for model validation. The flow diagram of the progress is illustrated in figure 1.

Inclusion criteria

To participate in this study, patients must comply with all of the following.

1. Diagnosis of heart failure with reduced EF (HFrEF) according to the 2016 European Society of Cardiology (ESC) HF guideline.[8]
2. LVEF ≤35% (measured by Simpson's methods) after optimised medication including ACEI or ARB, beta-blocker and MRA if available and not contraindicated at least 3 months.
3. Signed informed consent.

Exclusion criteria

The patient with any of the following will be excluded.

1. Hypertrophic cardiomyopathy.
2. Rheumatic heart disease.

3. Congenital heart disease.
4. Pulmonary heart disease.
5. Pericardial diseases and myocarditis.
6. Acute myocardial infarction in recent 3 months, including ST segment elevated myocardio infarction (STEMI) and NSTEMI.
7. Aortic dissection.
8. Severe haematological disease including leukaemia, lymphoma, aplastic anaemia.
9. Autoimmune disease.
10. Malignant tumour.
11. Hormone replacement.
12. Application of other interventional clinical trials.
13. Non-drug therapies for improving heart function: cardiac resynchronization therapy with or without implantable cardioverter-defibrillator (CRT-P/D), ICD, heart transplantation, surgical resection of ventricular aneurysm, interventional left ventricular restoration with Revivent/Parachute system), MitraClip therapy for recurrent mitral regurgitation.

Endpoints

Primary endpoint

All-cause death and SCD, including cardiac death and death from other causes.

Secondary endpoint

Lethal arrhythmia, sudden cardiac arrest, cardiopulmonary resuscitation, rehospitalisation due to HF.

Recruitment and consent

Participants will be identified and recruited at each of the participating centres. The clinical status of potential participants will be assessed, and their medical records will also be reviewed to confirm the eligibility according to the inclusion and exclusion criteria.

The study details will be explained to all potentially eligible and interesting subjects. The patients who agree to attend this study will sign the informed consent form (ICF) indicating that they fully understand the study and their rights of confidentiality and withdrawal from the study without giving a reason.

Baseline evaluation

Prognostic models of HF in the last 10 years have been reviewed, and the associated risk factors have been ranked according to their corresponding HR in respective risk models (table 1, figure 2). Age, sex, New York Heart Association (NYHA) class, LVEF, prior HF hospitalisation, course of HF, severe valvular heart disease, atrial fibrillation, prior myocardial infarction/coronary artery bypass grafting (CABG), renal dysfunction, chronic obstructive pulmonary disease (COPD), diabetes mellitus (DM), ischaemic aetiology, decreased systolic pressure, low body mass index, anaemia, hyponatremia, high N-terminal probrain natriuretic peptide (NT-proBNP), uricemia and current smoker were included. Variables which were not listed in previous models but appear relevant to higher risk of SCD in HF patients, and would therefore, merit consideration, including syncope or presyncope, frequent premature ventricular beat, non-sustained ventricular tachycardia, complete left bundle branch block, long QT interval and increased QT dispersion. In addition, self-care ability, social support and psychological state including depression and anxiety, are also predictors for subsequent poor prognosis in HF patients. The above risk factors have been assessed and confirmed by an expert panel of cardiologists and statisticians and will be collected in this study particularly.

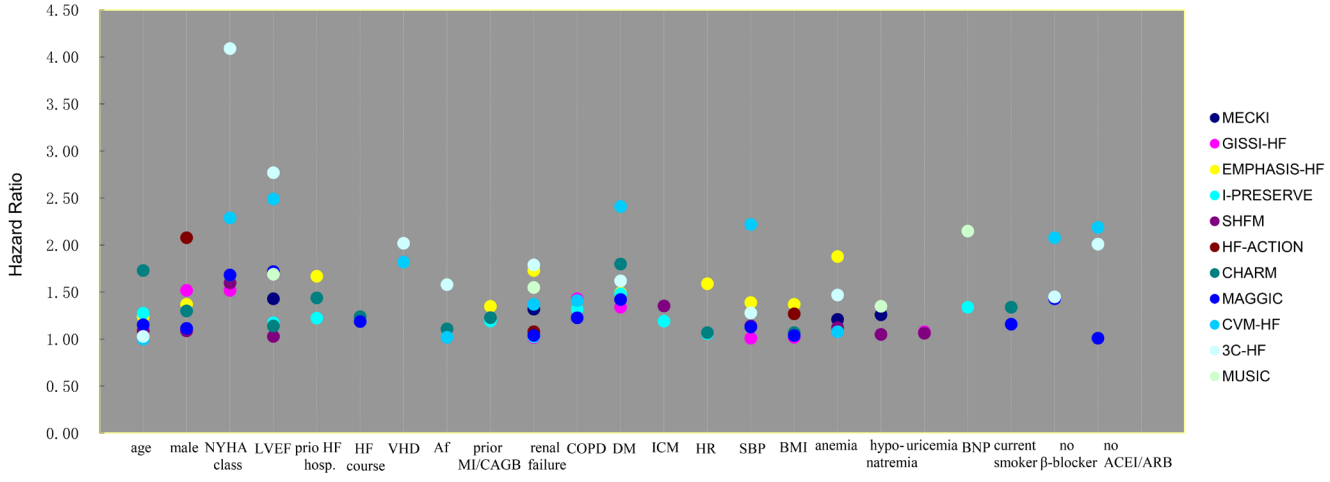

Figure 2 HR of variables in different risk models. Af, atrial fibrillation; BMI, body mass index; BNP, brain natriuretic peptide; CABGB, coronary artery bypass grafting; COPD, chronic obstructive pulmonary disease; DM, diabetes mellitus; HR, heart rate; ICM, ischaemic cardiomyopathy; MI, myocardial infarction; NYHA, New York Heart Association; SBP, systolic blood pressure; VHD, valvular heart disease.

Table 2 The checklist for data collection

Data collection	Baseline		Regular visit	Withdraw/death
	Retrospective cases	Prospective cases		
Informed consent	√	√		
Quantification verification (inclusion and exclusion)	√	√		
Baseline evaluation	√	√		
Medication	√	√		
Questionnaires 9-EHFScBS SSRS HAMD HAMA socioeconomic and educational status		√		
Regular follow-up visit (every 3 months)			√	
Survival state	√		√	√
Adverse event	Once happen √			
Study bias	Once happen √			
Withdraw from the study	Once happen √			
Death	Once happen √			

9-EHFScBS, 9-item European Heart Failure Self-care Behaviour Scale; HAMA, Hamilton Anxiety Scale; HAMD, Hamilton Depression Scale; SSRS, Social Support Rating Scale.

The baseline data that will be collected in all eligible subjects are as follows.

- Demographic characteristics: date of birth, gender, height and weight.
- Lifestyle behaviour: smoking and drinking status.
- Vital signs: blood pressure and heart rate.
- NYHA class.
- Aetiology of HF: the ischaemic aetiology will be confirmed if any following point is met: (a) prior myocardial infarction or revascularisation history (CABG/percutaneous coronary intervention); (b) left main or proximal segment of the left anterior descending artery stenosis ≥75% showed by coronary angiogram (CAG); (c) at least two main coronary artery branches stenosis ≥75% showed by CAG. Otherwise, non-ischaemic HF should be identified.
- Prior HF hospitalisation history: first HF hospitalisation or not, times of prior HF hospitalisation, the course of HF (since the HF symptoms appear; if unavailable, since the decreased EF was found).
- Coronary heart disease history: myocardial infarction or angina history, CAG result, revascularisation history, recent angina.
- Arrhythmia history: atrial fibrillation, atrial flutter, premature atrial contraction (PAC), premature ventricular contraction (PVC), non-sustained VT (NSVT), sustained VT, ventricular fibrillation and some bradyarrhythmias.
- Syncope or presyncope history.
- Cardiac arrest/cardiopulmonary resuscitation history.
- Other histories: hypertension, DM, COPD.
- Echocardiography: LV end-diastolic volume, LV end-systolic volume and LVEF measured by Simpson's method; left atrial diameter, LV end-diastolic diameter and LV end-systolic diameter, pulmonary artery systolic pressure. The status of valve regurgitation will be evaluated (0-none; 1-mild; 2-mild to moderate; 3-moderate; 4-severe).
- ECG: left/right bundle branch block will be recorded. QRS duration and QT interval will be tested, and QT dispersion will be calculated.
- Holter: total heartbeat of the whole day, minimum/maximum/average HR, onset of PVC, PAC, NSVT, VT, atrial fibrillation/flutter.
- Laboratory tests results: serum creatinine, blood urea nitrogen, serum natrium, haemoglobin, thyroid-stimulating hormone, free triiodothyronine, free thyroxine, NT-proBNP.
- Medication: ACEI/ARB, beta-blocker, aldosterone antagonist, diuretic, digoxin, antiplatelet agent, anticoagulant, statin, calcium channel blocker, antiarrhythmics, Ivabradine and angiotensin receptor blocker-neprilysin inhibitor.
- Evaluation of self-care behaviour and social support: 9-item European Heart Failure Self-care Behaviour Scale (9-EHFScBS)[36] will be used to determine the self-care levels in HF patients. Social Support Rating Scale (SSRS)[37] will be used to evaluate the social support condition in HF patients.
- Assessment of psychological status: Hamilton Depression Scale (HAMD) and Hamilton Anxiety Scale (HAMA).

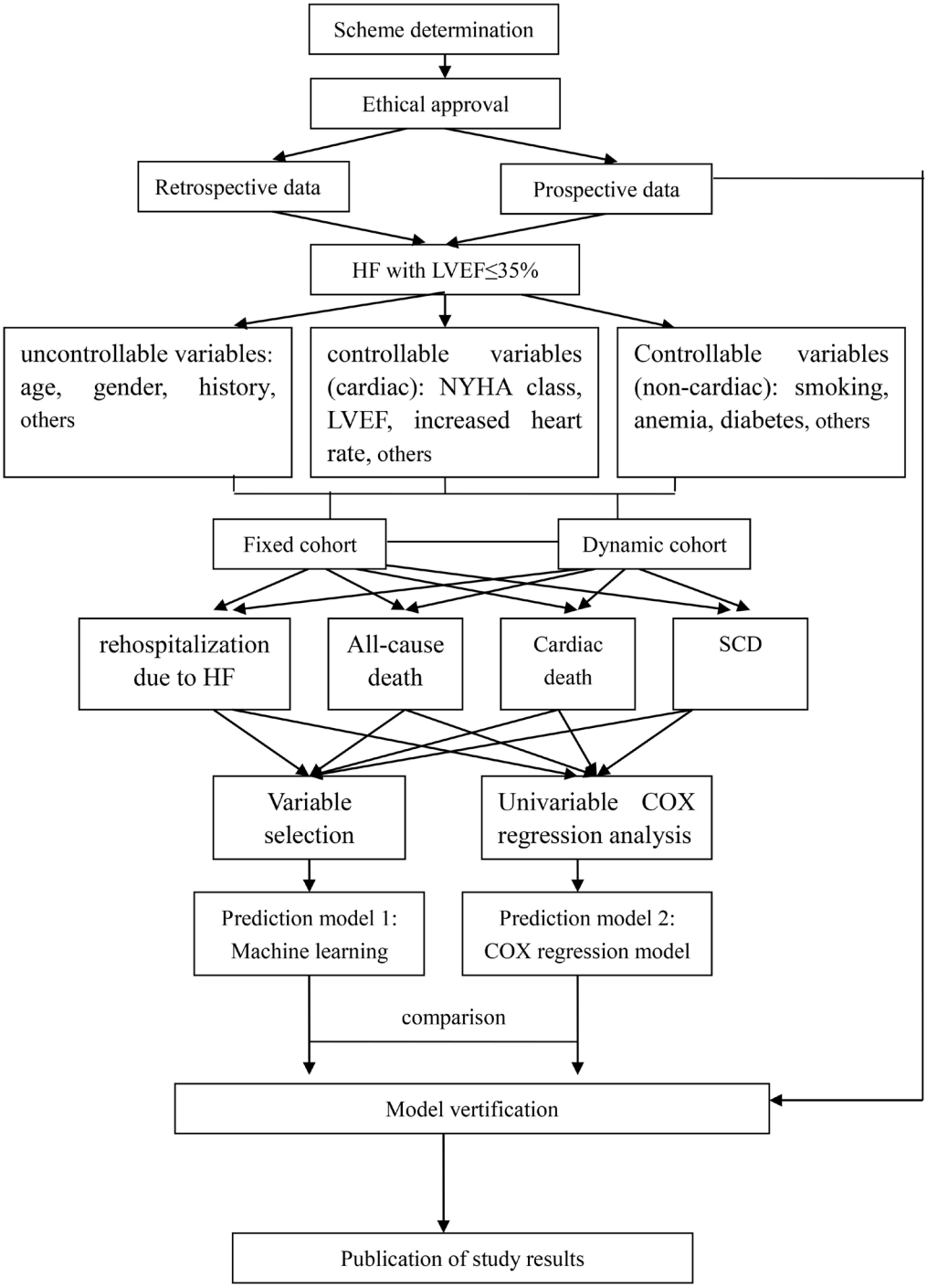

Figure 3 Study framework and process. HF, heart failure; LVEF, left ventricular ejection fraction; NYHA, New York Heart Association; SCD, sudden cardiac death.

- Socioeconomic and educational status: marital status, educational status, monthly income, sources of medical expenses, medical insurance.

Patient visits

After being enrolled in this research, all the subjects will be followed-up periodically in the outpatient department or by telephone interview every 3 months. The compliance with medications will be evaluated. As the primary endpoint, all-cause death and SCD will be focused. Cause of death will be analysed in detail. SCD is defined by the WHO as unexpected death that occurs within 1 hour from the onset of new or worsening symptoms (witnessed arrest) or, if unwitnessed, within 24 hours from when the individual was last observed alive and asymptomatic.[38] The lethal arrhythmia including ventricular tachycardia/ ventricular fibrillation (VT/VF), sudden cardiac arrest, cardiopulmonary resuscitation and rehospitalisation due to HF will be recorded carefully.

During follow-up, lethal arrhythmia will be recognised more precisely for patients who receive ICD or cardiac resynchronization therapy with implantable cardioverter-defibrillator (CRT/D) implantation, and will be recorded as an adverse event (AE). The patients, who receive CRT-P/D, heart transplantation, surgical resection of a ventricular aneurysm, interventional left ventricular restoration with Revivent/Parachute system, MitraClip therapy for recurrent mitral regurgitation, or some other non-drug therapy to improve heart function, will be followed up as usual.

Data collection

In the prospective part, clinical data of subjects will be collected and filled in the electrical data capture (EDC) system at baseline and particular follow-up visit. In the retrospective part, the same baseline information, except for 9-EHFScBS, SSRS, HAMD and HAMA questionnaires, will also be captured and input into the EDC system. The following prospective visits (every 3 months) will be conducted regularly and will be recorded in the EDC system. Investigators will record all the information of AEs, study bias, withdrawal from the study or death in EDC system. In this study, the participants will be identified by study codes, and their names will not appear in the EDC system. All the personal information including contact information, medical record and outcome will not be revealed to any person who has not been authorised by a principal investigator. Professional staffs are responsible for database management, data maintenance and regular data backup. Data quality will be monitored regularly. The data collection checklist is showed in table 2.

Data preprocessing

All above-collected variables, which might be predictors of all adverse prognosis of HF described in endpoint events, will be classified as uncontrollable variables (eg, age, gender, history), controllable variables associated with heart (eg, NYHA class, LVEF, increased heart rate) and controllable variables beyond heart (eg, smoking, anaemia, DM). Appropriate dummy variables will be used for binary variables and categorical variables, and quantitative variables will be fitted as a single continuous measurement (eg, age, heart rate, NT-proBNP), unless there is clear evidence of non-linearity. In order to create a practice simple risk score, some continuous variables will also be categorised into several groups according to both common clinical cut points and expert advice.

Machine learning

Variable selection is the process of selecting a subset of relevant variables for use in model construction, which can substantially reduce the abundant information and decrease the number of variables that are input to the prediction model. In this study, the technique named as 'information gain ranking' will be used to select appropriate variables. Information gain represents the effectiveness of a variable based on entropy, which characterises the unpredictability of a system. The information gain of a variable is evaluated as the entropy difference of the system when including and excluding this variable. Then, the variables whose information gain scores are less than a threshold are considered to be insignificant and will be excluded from the prediction.

Prediction models for SCD in HF patients will be developed by the following classification algorithms, respectively: decision trees, logistic regression, support vector machine, random forest and artificial neural network.[29] The performance and general error estimation of these ML models will be assessed by 10-fold cross-validation. The dataset will be randomly divided into 10 equal folds. Ninefolds will be used as the training set with the remaining onefold as the validation set. The validation results from 10 repeats will be combined to provide a measure of the overall performance. The prediction models derived from the above classification algorithms above will be evaluated based on the accuracy, sensitivities, specificities and the area under the receiver-operating characteristic (ROC) curve. Finally, clinical experts and computer specialists will discuss and choose the best model to predict the prognosis of SCD in HF patients and then perform further validation with the prospective dataset.

Cox proportional hazards regression

Univariable Cox proportional hazards modelling will be used to identify strong independent baseline candidate predictors for the primary and secondary outcomes. We will use both forward and backward stepwise procedure to derive the multivariable Cox proportional hazards model with $p<0.05$ as the inclusion criterion. Every variable in the model will be multiplied by its β-coefficient, and the products will be summed to calculate the risk score. Risk function will be used to estimate the level of risk. The calculating formula is as follows.[39]

$$P=h(t_j; X_k)=h_0(t_j)\exp(\mathrm{SCORE})$$

$$\mathrm{SCORE}=X_k \beta_k = \beta_0 + \beta_1 \times_1 + \beta_2 \times_2 + \ldots\ldots \beta_p x_p$$

Model validation

The dynamic prospective cases will be used for external validation of the optimal ML and Coxproportional hazards models. The validation will be performed using the models to calculate the probability of the outcome of interest occurring for each individual included in the validation sample when compared with the events actually observed to occur in this sample. The discrimination of each model will be estimated by ROC curve. The calibration of the models will be assessed by the Hosmer-Lemeshow goodness-of-fit test. The ML prediction model will be compared with the Cox proportional hazards regression model.

Patient and public involvement

During the design of this study, a survey of patient requirements, including communication needs, follow-up frequency and visit cost, was conducted in population

of potential HF participants, which provided important evidence for drawing up this study protocol to meet most of the patients' needs, build close contact with patients, enhance the overall adherence and improve the accuracy of endpoint event. This study is not a patient-led research, and patients are not involved in the recruitment of the study. The study results will be informed to the participants by phone at the end of this study. The alive patients will be evaluated with the new prediction model, and the ICD intervention will be recommended to the high SCD risk patients.

Study timeframe

The retrospective data collection in the two subcentres started in March 2017, and prospective enrolment in all 14 subcentres has started in January 2018. The follow-up period is scheduled to end in December 2019. The major part of data analysis will be performed from January to June 2020. The study framework and process is summarised in figure 3.

ETHICS AND DISSEMINATION

All necessary information about this study will be disclosed to the patients. Every subject will be asked to sign the ICF, indicating that they fully understand the study and voluntarily participate in this study. All results of this study will be published in international peer-reviewed journals and presented at relevant conferences.

DISCUSSION

The evaluation of SCD risk in HF patients is a problem that urgently needed to be solved. The existing prediction strategies for the SCD risk in HF patients lack clinical practice value for various reasons. ICD indication for primary prevention of SCD could be optimised by identifying the high SCD risk patients in HF with low LVEF (≤35%). It is of great practical value and economic significance.

We reviewed some predictive studies of HF in the past years and ranked the risk factors according to their corresponding HR, which have been included in our study as candidate risk factors. Otherwise, some other variables which appear relevant to risk of SCD in HF patients are also collected. Therefore, the efficiency and practicality of predictive model development has been highly improved.

This study is the first multicentre registry study in China, aimed to investigate the feasibility and accuracy of applying ML to predict SCD in HF patients with low LVEF. A broad range of outcomes, including SCD, all-cause death, lethal arrhythmia, sudden cardiac arrest, cardiopulmonary resuscitation and rehospitalisation due to HF, will be evaluated in this study, and the corresponding prognostic models will be developed. ML and the traditional multivariable Cox proportional hazards regression model will be derived from the same database and will be compared.

The limitations of this study are as follows: (1) HF patients with LVEF >35% will not be included based on the design of this study, which will restrict the application of the results of this study to the HF with low LVEF. (2) It might be difficult to determine the endpoint of this study sometimes for some patients, when dealing with SCD, lethal arrhythmia and sudden cardiac arrest, especially when outside the hospital.

Author affiliations
[1]Department of Cardiology, The First Affiliated Hospital of Nanjing Medical University, Nanjing, Jiangsu, China
[2]Department of Cardiology, Xiamen Cardiovascular Hospital, Xiamen University, Xiamen, Fujian, China
[3]Department of Cardiology, Jiangning Hospital Affiliated to Nanjing Medical University, Nanjing, Jiangsu, China
[4]Department of Cardiology, Wuhan Asia Heart Hospital, Wuhan, Hubei, China
[5]Department of Cardiology, The Second People's Hospital of Lianyungang, Lianyungang, Jiangsu, China
[6]Department of Cardiology, The Affiliated Hospital of Jiangsu University, Zhenjiang, Jiangsu, China
[7]Department of Cardiology, Taixing People's Hospital, Taixing, Jiangsu, China
[8]Department of Cardiology, The First People's Hospital of Huaian, Huaian, Jiangsu, China
[9]Department of Cardiology, The First People's Hospital of Yancheng, Yancheng, Jiangsu, China
[10]Department of Cardiology, Rugao People's Hospital, Rugao, Jiangsu, China
[11]Department of Cardiology, The First People's Hospital of Zhangjiagang, Zhangjiagang, Jiangsu, China
[12]Department of Cardiology, The Third People's Hospital of Suzhou, Suzhou, Jiangsu, China
[13]Department of Cardiology, The Third People's Hospital of Wuxi, Wuxi, Jiangsu, China
[14]Department of Cardiology, The Second Affiliated Hospital of Nanjing Medical University, Nanjing, Jiangsu, China
[15]School of Computing, University of Southern Mississippi, Hattiesburg, Mississippi, USA
[16]Department of Epidemiology, Nanjing Medical University, Nanjing, Jiangsu, China
[17]Department of Biostatistics, Nanjing Medical University, Nanjing, Jiangsu, China
[18]Key Laboratory of Targeted Intervention of Cardiovascular Disease, Collaborative Innovation Center for Cardiovascular Disease Translational Medicine, Nanjing Medical University, Nanjing, Jiangsu, China

Acknowledgements The authors thank Xiamen Cardiovascular Hospital, Xiamen University (Xiamen, China), Wuhan Asia Heart Hospital (Wuhan, China), Jiangning Hospital Affiliated to Nanjing Medical University, (Nanjing, China), The Second People's Hospital of Lianyungang (Lianyungang, China), The Affiliated Hospital of Jiangsu University (Zhenjiang, China), Taixing People's Hospital (Taixing, China), The First People's Hospital of Huaian (Huaian, China), The First People's Hospital of Yancheng (Yancheng, China), Rugao People's Hospital (Rugao, China), The First People's Hospital of Zhangjiagang (Zhangjiagang, China), The Third People's Hospital of Suzhou (Suzhou, China), The Third People's Hospital of Wuxi (Wuxi, China) and The Second Affiliated Hospital of Nanjing Medical University (Nanjing, China) for collaboration including recruitment and follow-up of HF patients. The authors also thank the HF patients who participated in the survey of patient requirements during the design of this study.

Contributors JGZ and FM conceived and designed the study. ZZ, XH, ZQ, YW, YC, YLW, YZ, ZC, XZ, JY, JLZ, JG, KL, LC, RZ and HJ participated in different phases of the protocol design. WZ provided expertise in data processing and machine learning. ST and YW provided their expertise for traditional statistical analysis. JGZ obtained funding. FM drafted the final manuscript. All authors have read the manuscript and provided feedback. JGZ approved the final version of the manuscript before submission. FM took responsibility for the submission process.

Funding This study was sponsored partly by the grant of clinical frontier technology from Jiangsu Science and Technology Agency (BE2016764).

Competing interests None declared.

Patient consent for publication Obtained.

Ethics approval The study protocol has been approved by the Ethics Committee of The First Affiliated Hospital of Nanjing Medical University (2017-SR-06).

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1. Sato N. Epidemiology of Heart Failure in Asia. *Heart Fail Clin* 2015;11:573–9.
2. Mozaffarian D, Benjamin EJ, Go AS, *et al*. American Heart Association Statistics Committee and Stroke Statistics Subcommittee. Heart disease and stroke statistics–2015 update: a report from the American Heart Association. *Circulation* 2015;131:e29–e322.
3. Tomaselli GF, Zipes DP. What causes sudden death in heart failure? *Circ Res* 2004;95:754–63.
4. Solomon SD, Wang D, Finn P, *et al*. Effect of candesartan on cause-specific mortality in heart failure patients: the Candesartan in Heart failure Assessment of Reduction in Mortality and morbidity (CHARM) program. *Circulation* 2004;110:2180–3.
5. Yousuf O, Chrispin J, Tomaselli GF, *et al*. Clinical Management and Prevention of Sudden Cardiac Death. *Circ Res* 2015;116:2020–40.
6. Connolly SJ, Hallstrom AP, Cappato R, *et al*. Meta-analysis of the implantable cardioverter defibrillator secondary prevention trials. AVID, CASH and CIDS studies. Antiarrhythmics vs Implantable Defibrillator study. Cardiac Arrest Study Hamburg. Canadian Implantable Defibrillator Study. *Eur Heart J* 2000;21:2071–8.
7. Myerburg R, Spooner PM. Opportunities for sudden death prevention: Directions for new clinical and basic research. *Cardiovasc Res* 2001;50:177–85.
8. Ponikowski P, Voors AA, Anker SD, *et al*. ESC Scientific Document Group. 2016 ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure: The Task Force for the diagnosis and treatment of acute and chronic heart failure of the European Society of Cardiology (ESC)Developed with the special contribution of the Heart Failure Association (HFA) of the ESC. *Eur Heart J* 2016;37:2129–200.
9. Køber L, Thune JJ, Nielsen JC, *et al*. Defibrillator Implantation in Patients with Nonischemic Systolic Heart Failure. *N Engl J Med* 2016;375:1221–30.
10. Myerburg RJ, Spooner PM. Opportunities for sudden death prevention: directions for new clinical and basic research. *Cardiovasc Res* 2001;50:177–85.
11. Stecker EC, Vickers C, Waltz J, *et al*. Population-based analysis of sudden cardiac death with and without left ventricular systolic dysfunction: two-year findings from the Oregon Sudden Unexpected Death Study. *J Am Coll Cardiol* 2006;47:1161–6.
12. Aimo A, Januzzi JJ, Vergaro G, *et al*. Left ventricular ejection fraction for risk stratification in chronic systolic heart failure[J]. *Int J Cardiol* 2018.
13. Shen L, Jhund PS, Petrie MC, *et al*. Declining Risk of Sudden Death in Heart Failure. *N Engl J Med* 2017;377:41–51.
14. Agostoni P, Corrà U, Cattadori G, *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;167:2710–8.
15. Barlera S, Tavazzi L, Franzosi MG, *et al*. Predictors of mortality in 6975 patients with chronic heart failure in the Gruppo Italiano per lo Studio della Streptochinasi nell'Infarto Miocardico-Heart Failure trial: proposal for a nomogram. *Circ Heart Fail* 2013;6:31–9.
16. Collier TJ, Pocock SJ, McMurray JJ, *et al*. The impact of eplerenone at different levels of risk in patients with systolic heart failure and mild symptoms: insight from a novel risk score for prognosis derived from the EMPHASIS-HF trial. *Eur Heart J* 2013;34:2823–9.
17. Komajda M, Carson PE, Hetzel S, *et al*. Factors associated with outcome in heart failure with preserved ejection fraction: findings from the Irbesartan in Heart Failure with Preserved Ejection Fraction Study (I-PRESERVE). *Circ Heart Fail* 2011;4:27–35.
18. Levy WC, Mozaffarian D, Linker DT, *et al*. The Seattle Heart Failure Model: prediction of survival in heart failure. *Circulation* 2006;113:1424–33.
19. O'Connor CM, Whellan DJ, Wojdyla D, *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.
20. Pocock SJ, Wang D, Pfeffer MA, *et al*. Predictors of mortality and morbidity in patients with chronic heart failure. *Eur Heart J* 2006;27:65–75.
21. Pocock SJ, Ariti CA, McMurray JJ, *et al*. Predicting survival in heart failure: a risk score based on 39 372 patients from 30 studies. *Eur Heart J* 2013;34:1404–13.
22. Senni M, Santilli G, Parrella P, *et al*. A novel prognostic index to determine the impact of cardiac conditions and co-morbidities on one-year outcome in patients with heart failure. *Am J Cardiol* 2006;98:1076–82.
23. Senni M, Parrella P, De Maria R, *et al*. Predicting heart failure outcome from cardiac and comorbid conditions: the 3C-HF score. *Int J Cardiol* 2013;163:206–11.
24. Vazquez R, Bayes-Genis A, Cygankiewicz I, *et al*. The MUSIC Risk score: a simple method for predicting mortality in ambulatory patients with chronic heart failure. *Eur Heart J* 2009;30:1088–96.
25. Uszko-Lencer N, Frankenstein L, Spruit MA, *et al*. Predicting hospitalization and mortality in patients with heart failure: The BARDICHE-index. *Int J Cardiol* 2017;227:901–7.
26. Delgado V, Bucciarelli-Ducci C, Bax JJ. Diagnostic and prognostic roles of echocardiography and cardiac magnetic resonance. *J Nucl Cardiol* 2016;23:1399–410.
27. Halliday BP, Cleland JGF, Goldberger JJ, *et al*. Personalizing Risk Stratification for Sudden Death in Dilated Cardiomyopathy: The Past, Present, and Future. *Circulation* 2017;136:215–31.
28. Kelesidis I, Travin MI. Use of cardiac radionuclide imaging to identify patients at risk for arrhythmic sudden cardiac death. *J Nucl Cardiol* 2012;19:142–52.
29. Martins da Silva MI, Vidigal Ferreira MJ, Morão Moreira AP. Iodine-123-metaiodobenzylguanidine scintigraphy in risk stratification of sudden death in heart failure. *Rev Port Cardiol* 2013;32:509–16.
30. Aro AL, Reinier K, Rusinaru C, *et al*. Electrical risk score beyond the left ventricular ejection fraction: prediction of sudden cardiac death in the Oregon Sudden Unexpected Death Study and the Atherosclerosis Risk in Communities Study. *Eur Heart J* 2017;38:3017–25.
31. Quinlan JR. Induction of decision trees. *Mach Learn* 1986;1:81–106.
32. Weng SF, Reps J, Kai J, *et al*. Can machine-learning improve cardiovascular risk prediction using routine clinical data? *PLoS One* 2017;12:e174944.
33. Awan SE, Sohel F, Sanfilippo FM, *et al*. Machine learning in heart failure: ready for prime time. *Curr Opin Cardiol* 2018;33:190–5.
34. Ebrahimzadeh E, Foroutan A, Shams M, *et al*. An optimal strategy for prediction of sudden cardiac death through a pioneering feature-selection approach from HRV signal. *Comput Methods Programs Biomed* 2019;169:19–36.
35. Au-Yeung W-TM, Reinhall PG, Bardy GH, *et al*. Development and validation of warning system of ventricular tachyarrhythmia in patients with heart failure with heart rate variability data. *PLoS One* 2018;13:e207215.
36. Jaarsma T, Arestedt KF, Mårtensson J, *et al*. The European Heart Failure Self-care Behaviour scale revised into a nine-item scale (EHFScB-9): a reliable and valid international instrument. *Eur J Heart Fail* 2009;11:99–105.
37. Hu X, Hu X, Su Y, *et al*. The changes and factors associated with post-discharge self-care behaviors among Chinese patients with heart failure. *Patient Prefer Adherence* 2015;9:1593–601.
38. Yousuf O, Chrispin J, Tomaselli GF, *et al*. Clinical management and prevention of sudden cardiac death. *Circ Res* 2015;116:2020–40.
39. Harrell FE, Lee KL, Califf RM, *et al*. Regression modelling strategies for improved prognostic prediction. *Stat Med* 1984;3:143–52.