





OPEN

DATA DESCRIPTOR

Electronic healthcare records and external outcome data for hospitalized patients with heart failure

Zhongheng Zhang^{1,2,6}  , Linghong Cao^{3,6}, Rangui Chen³, Yan Zhao³, Lukai Lv³, Ziyin Xu³ & Ping Xu^{3,4,5} 

Heart failure is one of the most important reasons for hospitalization among elderly individuals and is associated with significant mortality and morbidity. Epidemiological studies require the establishment of high-quality databases. Several datasets that primarily involve heart failure populations have been established in Western countries and have generated many high-quality studies. However, no such dataset is available from China. Due to differences in genetic background and healthcare systems between China and Western countries, the establishment of a heart failure database for the Chinese population is urgently needed. We performed a retrospective single-center observational study to collect data regarding the characteristics of heart failure patients in China by integrating electronic healthcare records and follow-up outcome data. The study collected information for a total of 2,008 patients with heart failure, containing 166 attributes.

Background & Summary

Heart failure (HF) affects over 6 million people in the United States, with an estimated incidence of 21 per 1000 people in the elderly population¹. By using mathematical prediction models, heart failure is estimated to affect 8 million people over the age of 18². Heart failure is one of the most important reasons for hospitalization among elderly individuals and is also associated with significant mortality and morbidity. It has been reported that the mortality ranges from 20% to 60% one year after hospitalization for acute HF³⁻⁶, depending on comorbidities and coexisting medical conditions. Many cohort studies have been carried out for epidemiological investigations of hospitalized patients with HF. For instance, the Cleveland Heart Disease dataset, which contains 75 variables for 303 patients, is mainly used for practising machine learning algorithms⁷. The Nationwide Inpatient Sample (NIS) is a publicly available database from the Healthcare Utilization Project (HCUP) that is supported by the Agency for Healthcare Research and Quality. This dataset contains many patients with heart diseases, but the variables/attributes included in this dataset are not specifically designed for HF⁸⁻¹⁰. The Medical Information Mart for Intensive Care (MIMIC) database contains data associated with >60,000 distinct hospital admissions to critical care units between 2001 and 2012. Many of the patients have a HF diagnosis, and thus MIMIC is a good resource for testing research hypotheses related to critically ill HF patients¹¹. However, these studies are either designed with attributes that are limited in number or not specific for HF. In other words, data collection was dictated by expert knowledge, and only variables deemed important were entered into the data collection form. The determination of feature variable inclusion/exclusion is largely driven by expertise and previous studies. Such a dataset can be used only to address a limited number of clinical questions. For example, the ESC-Heart Failure Association (HFA) EURObservational Research Programme (EORP) generated a large dataset that contained

¹Department of Emergency Medicine, Sir Run Run Shaw Hospital, Zhejiang University School of Medicine, Hangzhou, 310016, Zhejiang, China. ²Key Laboratory of Emergency and Trauma, Ministry of Education, College of Emergency and Trauma, Hainan Medical University, Haikou, 571199, China. ³Emergency Department, Zigong Fourth People's Hospital, 19 Tanmulin Road, Zigong, Sichuan, China. ⁴Artificial Intelligence Key Laboratory of Sichuan Province, Zigong, 643000, China. ⁵Medical Big Data and Artificial Intelligence Laboratory of Zigong Fourth People's Hospital, Zigong, 643000, China. ⁶These authors contributed equally: Zhongheng Zhang, Linghong Cao. ✉e-mail: zh_zhang1984@zju.edu.cn; xp1657@126.com

Code	Description
428	Heart failure
4280	Congestive heart failure, unspecified
4281	Left heart failure
4282	Systolic heart failure
42820	Systolic heart failure, unspecified
42821	Acute systolic heart failure
42822	Chronic systolic heart failure
42823	Acute on chronic systolic heart failure
4283	Diastolic heart failure
42830	Diastolic heart failure, unspecified
42831	Acute diastolic heart failure
42832	Chronic diastolic heart failure
42833	Acute on chronic diastolic heart failure
4284	Combined systolic and diastolic heart failure
42840	Combined systolic and diastolic heart failure, unspecified
42841	Acute combined systolic and diastolic heart failure
42842	Chronic combined systolic and diastolic heart failure
42843	Acute on chronic combined systolic and diastolic heart failure
4289	Heart failure, unspecified

Table 1. ICD-9 code for the diagnosis of heart failure.

specifically HF patients, and a large amount of data that are routinely collected during clinical practice were abandoned. To the best of our knowledge, this is the largest HF dataset in the world, including 337 cardiology centres from 33 ESC Member countries¹². In essence, many trivial attributes may work together to influence the clinical outcome. Thus, a dataset including all aspects of individual patient-level data can help disentangle complex relationships among attributes. In the era of big data, the electronic healthcare records are able to produce a large amount of data related to a given HF patient. These multiparameter relational databases may or may not be related to a given research question. Different studies and analyses require different variables. Making such a publicly available dataset can help to encourage data reuse, thereby promoting more medical knowledge discovery.

Our study aimed to establish a HF database based on electronic healthcare records. Data on subsequent hospital admissions and mortality were obtained at mandatory follow-up visits at 28 days, 3 months and 6 months (if the patient was unable to reach the clinical centre, the follow-up visit was replaced by a telephone call). The study was a retrospective study enrolling hospitalized patients with heart failure from December 2016 to June 2019. Patients were enrolled from Zigong Fourth People's Hospital. Data were extracted from electronic healthcare records. However, this is a single-centre dataset, covering only Chinese patients. Findings with these data alone may not have convincing generalizability. Researchers may combine this dataset with other heart failure cohort data for a larger-scale study.

Methods

Study setting and population. The study was conducted at Zigong Fourth People's Hospital, Sichuan, China from December 2016 to June 2019, and was approved by the ethics committee of Zigong Fourth People's Hospital (Approval Number: 2020-010). Informed consent was waived due to the retrospective design of the study. The study complies with the Declaration of Helsinki.

Electronic healthcare records of consecutive patients with a diagnosis of HF were reviewed. We included all types of heart failure including acute HF, chronic HF, left HF, right HF, or a mixture of all. Heart failure was defined according to the European Society of Cardiology (ESC) criteria¹³:

- 1) The presence of symptoms and/or signs of HF. Typical symptoms include breathlessness, orthopnoea, paroxysmal nocturnal dyspnea, reduced exercise tolerance, fatigue, tiredness, increased time to recover after exercise and ankle swelling. Typical signs include elevated jugular venous pressure, hepatojugular reflux, third heart sound (gallop rhythm) and laterally displaced apical impulse.
- 2) Elevated levels of BNP (BNP > 35 pg/mL and/or NT-proBNP > 125 pg/mL)
- 3) Objective evidence of other cardiac functional and structural alterations underlying HF.
- 4) In case of uncertainty, a stress test or invasively measured elevated LV filling pressure may be needed to confirm the diagnosis.

Patients who had a diagnosis of heart failure on hospital admission were enrolled in our study. The diagnosis was recorded with ICD-9 in the EHR (Table 1).

Variables and attributes. Data collected for the dataset included three broad categories: demographic data, baseline clinical characteristics, comorbidities, laboratory findings, drugs and outcomes. Demographic data were entered manually into the EMR system by the nurses on admission if a patient first visited our hospital.

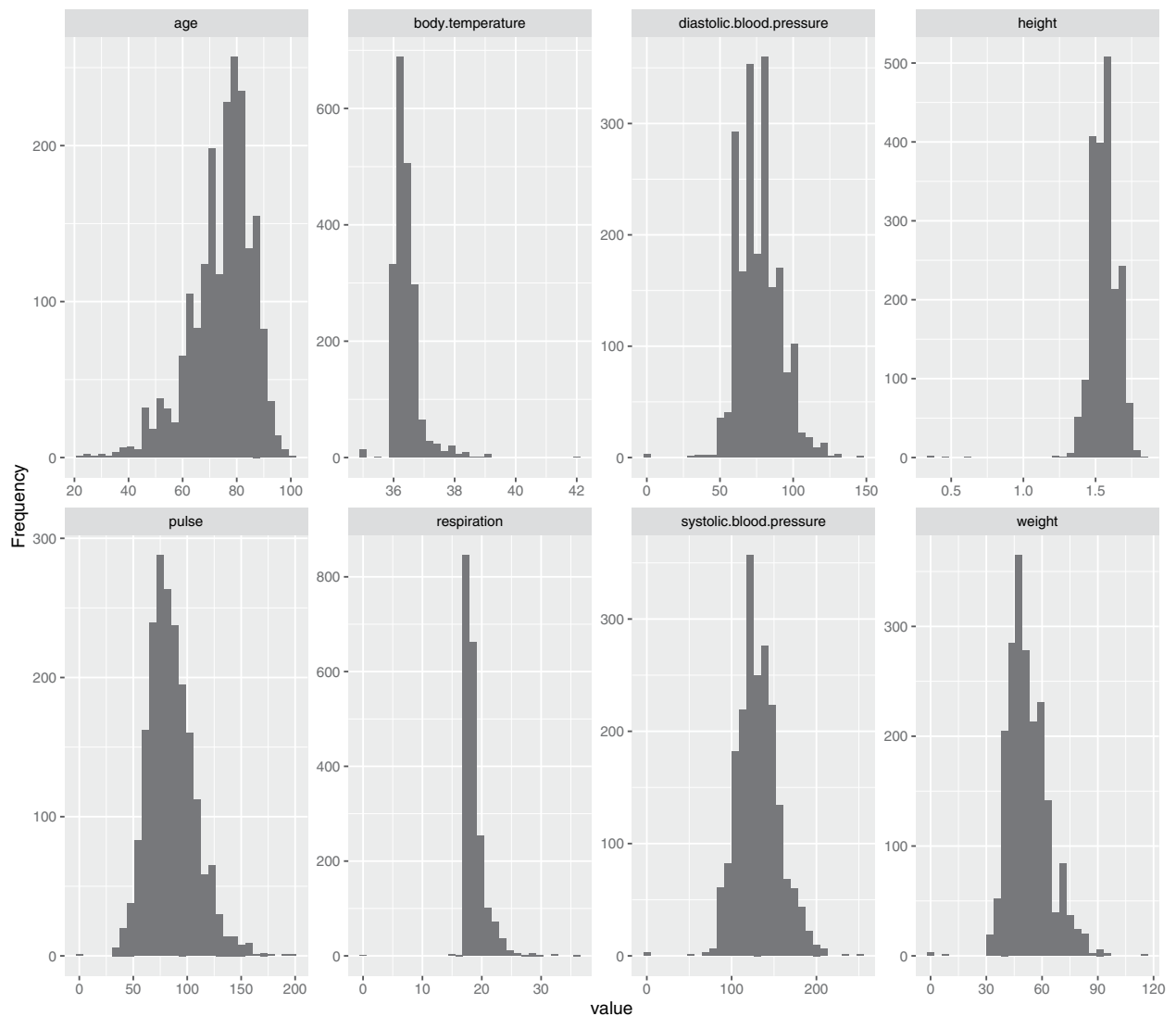


Fig. 1 Histogram showing the distribution of numeric attributes at baseline.

Otherwise, demographic data could be automatically extracted from previous visits. Some missing or error data were checked if they were identified by the nurses. To ensure the accuracy and consistency of data entry, a drop-down list was used for some variables in our EMR system, such as sex, department of admission and occupation. Laboratory tests and drugs were electronically entered by physicians and/or lab workers. Data in the EMR were extracted by SQL query to establish the current database. The accuracy of the SQL query was then checked manually by randomly selecting 50 patients. Many data items were recorded in Chinese in the electronic healthcare record database, thus the largest challenge is the language barrier. All the lab test items, examinations, drug names and diagnoses were recorded in Chinese in the electronic healthcare record database. To address this problem, all Chinese terms were translated to English by the principal investigators (Z.Z., P.X. and L.C.).

The demographic data were obtained from the first sheet of the medical records and included age, sex, height, body weight, admission ward, type of admission (emergency vs. nonemergency), occupation, discharge department, admission date, visit times, and marital status.

Baseline clinical characteristics were measured on the day of hospital admission and included body temperature, pulse, respiration rate, systolic blood pressure, diastolic blood pressure, mean arterial blood pressure, weight, height, body mass index (BMI), type of heart failure, New York Heart Association (NYHA) cardiac function, Killip Grade (Class 1 No rales, no 3rd heart sound; Class 2 Rales in $<1/2$ lung field or presence of a 3rd heart sound; Class 3 Rales in $>1/2$ lung field–pulmonary oedema; Class 4 Cardiogenic shock–determined clinically), and Glasgow Coma Scale (GCS) score. Echocardiographic findings included left ventricular ejection fraction (LVEF), left ventricular end diastolic diameter, mitral valve peak E wave velocity (m/s), mitral valve peak A wave velocity (m/s), E/A, tricuspid valve regurgitation velocity, and tricuspid valve regurgitation pressure.

Comorbidities included a medical history of myocardial infarction, congestive heart failure, peripheral vascular disease, cerebrovascular disease, dementia, chronic obstructive pulmonary disease (COPD), connective tissue disease, peptic ulcer disease, diabetes, moderate-to-severe chronic kidney disease, hemiplegia, leukaemia,

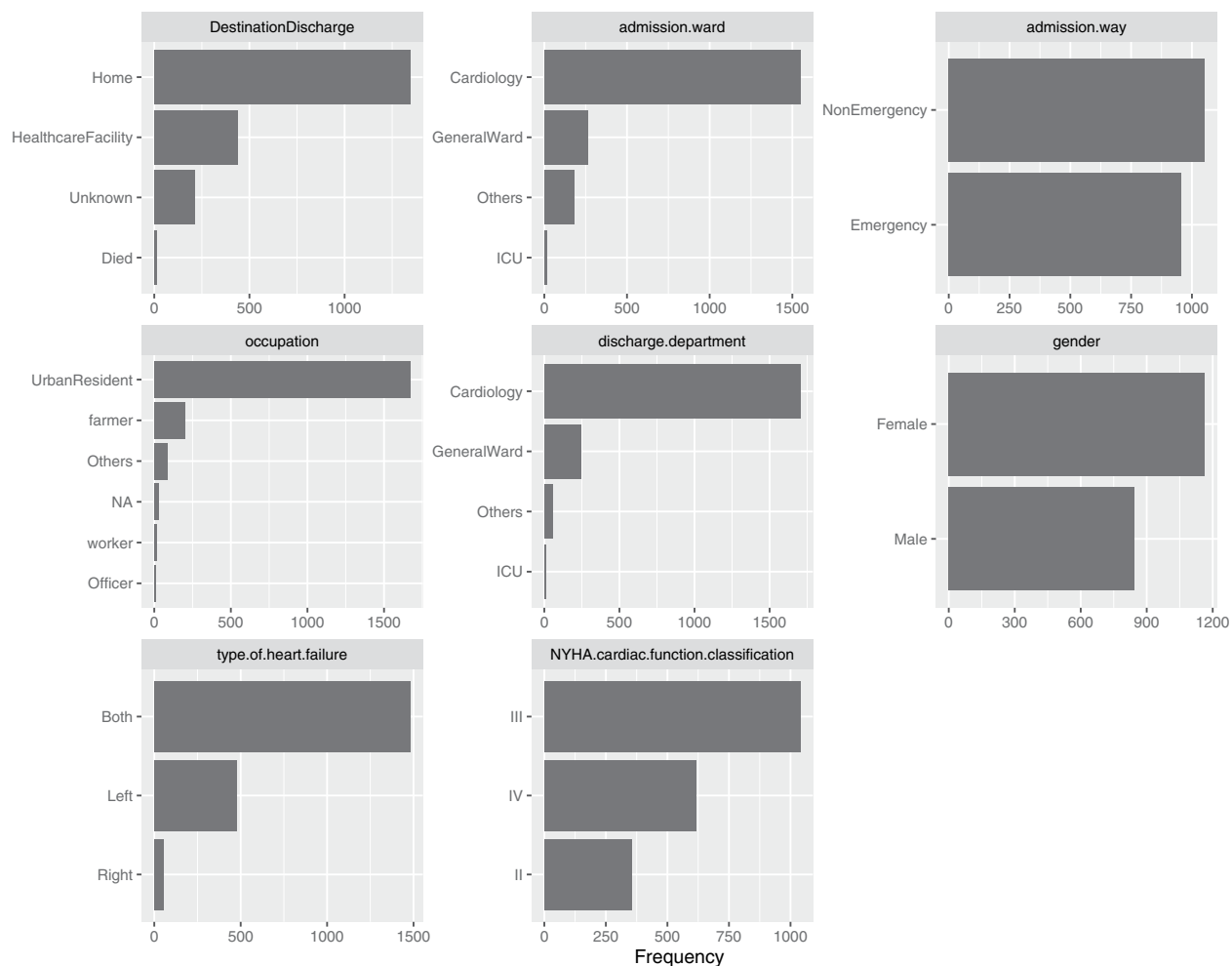


Fig. 2 Bar chart showing the distribution of discrete attributes at baseline.

malignant lymphoma, solid tumour, liver disease and AIDS. The Charlson Comorbidity Index (CCI) was calculated by summing all comorbidity points described above¹⁴. A minority of patients were not coded as having a diagnosis of “congestive heart failure” in the comorbidity list because they did not have a past history of congestive heart failure on admission. They were diagnosed with HF for the first time in the index hospitalization. The comorbidities were taken from the admission notes.

Laboratory findings were obtained from day one of hospital admission, including serum creatinine, urea, uric acid, glomerular filtration rate, cystatin, white blood cell count, monocyte ratio, monocyte count, red blood cell count, coefficient of variation of red blood cell distribution width, standard deviation of red blood cell distribution width, mean corpuscular volume, haematocrit, lymphocyte count, mean haemoglobin volume, mean haemoglobin concentration, mean platelet volume, basophil ratio, basophil count, eosinophil ratio, eosinophil count, haemoglobin, platelet, platelet distribution width, platelet haematocrit, neutrophil ratio, neutrophil count, D-dimer, international normalized ratio, activated partial thromboplastin time, thrombin time, prothrombin activity, prothrombin time ratio, fibrinogen, high sensitivity troponin, myoglobin, carbon dioxide binding capacity, calcium, potassium, chloride, sodium, inorganic phosphorus, serum magnesium, creatine kinase isoenzyme to creatine kinase, hydroxybutyrate dehydrogenase to lactate dehydrogenase, hydroxybutyrate dehydrogenase, glutamic oxaloacetic transaminase, creatine kinase, creatine kinase isoenzyme, lactate dehydrogenase, brain natriuretic peptide, high sensitivity protein, nucleotidase, fucosidase, albumin, albumin/globulin ratio, cholinesterase, glutamyltranspeptidase, glutamic pyruvic transaminase, glutamic oxalylplatin, indirect bilirubin, alkaline phosphatase, globulin, direct bilirubin, total bilirubin, total bile acid, total protein, erythrocyte sedimentation rate, cholesterol, low-density lipoprotein cholesterol, triglyceride, high-density lipoprotein cholesterol, homocysteine, apolipoprotein A, apolipoprotein B, lipoprotein, pH, standard residual base, standard bicarbonate, partial pressure of carbon dioxide, total carbon dioxide, methemoglobin, haematocrit blood gas, reduced haemoglobin, potassium ion, chloride ion, sodium ion, glucose blood gas, lactate, measured residual base, measured bicarbonate, carboxyhaemoglobin, body temperature blood gas, oxygen saturation, partial oxygen pressure, oxyhaemoglobin, anion gap, free calcium, and total haemoglobin.

Primary drug categories included in our dataset were diuretics, inotropes, and vasodilators. The diuretics included furosemide, torasemide and spironolactone. Inotropes included deslanoside, dobutamine, digoxin, isoprenaline and milrinone. Vasodilators included isosorbide mononitrate and nitroglycerin.

Outcome variables included discharge date of the index hospital, vital status at hospital discharge, death within 28 days, readmission within 28 days, death within 3 months, readmission within 3 months, death within 6 months, readmission within 6 months, time to death (days from index hospital admission), time to readmission (days from index hospital admission), return to emergency department within 6 months, and time to visit emergency department within 6 months. The variable “*DestinationDischarge*” was recorded after hospital discharge, and the variable “*outcome.during.hospitalization*” was recorded after the decision to discharge was made.

Data Records

The study generated a single dataset, that contained information on 166 attributes of 2008 hospitalized patients from December 2016 to June 2019. The dataset is available at PhysioNet (<https://doi.org/10.13026/8a9e-w734>)¹⁵. Missing values are indicated with blanks. Detailed information on variable specifications is included in a variable description file.

Technical Validation

The present study was a retrospective design. Information on eligible patients was collected at Zigong Fourth People’s Hospital. First, the required data were exported from the electronic healthcare database with the assistance of the information technology technician. The exported data were then checked by expert emergency and critical care physicians; if outliers in each variable and contradictions within data were detected, data were validated by another investigator. The outliers and contradictions were judged by expert emergency and critical care physicians. Data on subsequent hospital admissions and mortality were obtained at mandatory follow-up visit at 28 days, 3 months and 6 months (if the patient was unable to reach the clinical centre, the follow-up visit was replaced by a telephone call).

Data were finalized and fully anonymized on June 8, 2020.

Baseline characteristics of included patients. The overall mortality rate at hospital discharge was 1% (14/2008). A total of 212 patients were discharged to unknown places (212/2008, 11%), 1344 patients were discharged home (67%) and 438 patients were discharged to healthcare facilities (22%). Most patients were admitted to the department of cardiology (1547/2008, 77%), followed by the general ward (265/2008, 13%), others (181/2008, 9%) and the ICU (15/2008, 1%). There was also a significant difference between emergency and none-emergency patients (Online-only Table 1). The distributions of the baseline characteristics are shown in Fig. 1, Fig. 2 and Online-only Table 1.

Received: 14 September 2020; Accepted: 18 January 2021;

Published online: 05 February 2021

References

1. Benjamin, E. J. *et al.* Heart Disease and Stroke Statistics-2017 Update: A Report From the American Heart Association. *Circulation* **135**, e146–e603 (2017).
2. Fang, N., Jiang, M. & Fan, Y. Ideal cardiovascular health metrics and risk of cardiovascular disease or mortality: A meta-analysis. *Int. J. Cardiol.* **214**, 279–283 (2016).
3. Lombardi, C. *et al.* In-hospital and long-term mortality for acute heart failure: analysis at the time of admission to the emergency department. *ESC Heart Fail* <https://doi.org/10.1002/ehf2.12847> (2020).
4. Ye, S.-D. *et al.* Association between anemia and outcome in patients hospitalized for acute heart failure syndromes: findings from Beijing Acute Heart Failure Registry (Beijing AHF Registry). *Intern Emerg Med* **151**, 457 (2020).
5. Braunschweig, F. *et al.* New York Heart Association functional class, QRS duration, and survival in heart failure with reduced ejection fraction: implications for cardiac resynchronization therapy. *Eur. J. Heart Fail.* **19**, 366–376 (2017).
6. Al-Jarallah, M. *et al.* Incidence and impact of cardiorenal anaemia syndrome on all-cause mortality in acute heart failure patients stratified by left ventricular ejection fraction in the Middle East. *ESC Heart Fail* **6**, 103–110 (2019).
7. Detrano, R. *et al.* International application of a new probability algorithm for the diagnosis of coronary artery disease. *Am. J. Cardiol.* **64**, 304–310 (1989).
8. Shah, R. U. & Merz, C. N. B. Publicly Available Data: Crowd Sourcing to Identify and Reduce Disparities. *J. Am. Coll. Cardiol.* **66**, 1973–1975 (2015).
9. Khera, S. *et al.* Temporal Trends and Sex Differences in Revascularization and Outcomes of ST-Segment Elevation Myocardial Infarction in Younger Adults in the United States. *J. Am. Coll. Cardiol.* **66**, 1961–1972 (2015).
10. Stretch, R., Sauer, C. M., Yuh, D. D. & Bonde, P. National trends in the utilization of short-term mechanical circulatory support: incidence, outcomes, and cost analysis. *J. Am. Coll. Cardiol.* **64**, 1407–1415 (2014).
11. Johnson, A. E. W. *et al.* MIMIC-III, a freely accessible critical care database. *Sci Data* **3**, 160035 (2016).
12. Kapłon-Cieślicka, A. *et al.* Is heart failure misdiagnosed in hospitalized patients with preserved ejection fraction? From the European Society of Cardiology - Heart Failure Association EURObservational Research Programme Heart Failure Long-Term Registry. *ESC Heart Fail* **2**, 235 (2020).
13. Ponikowski, P. *et al.* 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. *European journal of heart failure* **18**, 891–975 (2016).
14. Charlson, M. E., Pompei, P., Ales, K. L. & MacKenzie, C. R. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. *J Chronic Dis* **40**, 373–383 (1987).
15. Zhang, Z. *et al.* Hospitalized patients with heart failure: integrating electronic healthcare records and external outcome data. *PhysioNet* <https://doi.org/10.13026/8a9e-w734> (2020).

Acknowledgements

P.X. received funding from Research project of Zigong City Science & Technology and Intellectual Property Right Bureau (2018SF04), Research project of Health, Open Foundation of Artificial Intelligence Key Laboratory of Sichuan Province and Family Planning Commission Of Sichuan Province (17PJ136) and Research project of Zigong City Science & Technology and Intellectual Property Right Bureau (2017SF04). Z.Z. received funding from Key Laboratory of Emergency and Trauma (Hainan Medical University), Ministry of Education (Grant.KLET-202017).

Author contributions

Z.Z. and P.X. conceived the idea; L.C. and R.C. curated data; Y.Z. and L.L. checked accuracy of the data; Z.X. performed patient follow up.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Z.Z. or P.X.

Reprints and permissions information is available at www.nature.com/reprints.

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



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

The Creative Commons Public Domain Dedication waiver <http://creativecommons.org/publicdomain/zero/1.0/> applies to the metadata files associated with this article.

© The Author(s) 2021