


BMJ Open Predicting incident heart failure from population-based nationwide electronic health records: protocol for a model development and validation study

Yoko M Nakao ^{1,2,3} Ramesh Nadarajah ^{1,2,4} Farag Shuweihdi,¹ Kazuhiro Nakao,^{1,2,5} Ahmet Fuat,⁶ Jim Moore,⁷ Christopher Bates,⁸ Jianhua Wu,⁹ Chris Gale^{1,2,4}

To cite: Nakao YM, Nadarajah R, Shuweihdi F, *et al*. Predicting incident heart failure from population-based nationwide electronic health records: protocol for a model development and validation study. *BMJ Open* 2024;**14**:e073455. doi:10.1136/bmjopen-2023-073455

► Prepublication history for this paper is available online. To view these files, please visit the journal online (<http://dx.doi.org/10.1136/bmjopen-2023-073455>).

YMN and RN are joint first authors.

Received 06 March 2023
Accepted 29 June 2023



© Author(s) (or their employer(s)) 2024. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ.

For numbered affiliations see end of article.

Correspondence to
Dr Yoko M Nakao;
y.nakao@leeds.ac.uk

ABSTRACT

Introduction Heart failure (HF) is increasingly common and associated with excess morbidity, mortality, and healthcare costs. Treatment of HF can alter the disease trajectory and reduce clinical events in HF. However, many cases of HF remain undetected until presentation with more advanced symptoms, often requiring hospitalisation. Predicting incident HF is challenging and statistical models are limited by performance and scalability in routine clinical practice. An HF prediction model implementable in nationwide electronic health records (EHRs) could enable targeted diagnostics to enable earlier identification of HF.

Methods and analysis We will investigate a range of development techniques (including logistic regression and supervised machine learning methods) on routinely collected primary care EHRs to predict risk of new-onset HF over 1, 5 and 10 years prediction horizons. The Clinical Practice Research Datalink (CPRD)-GOLD dataset will be used for derivation (training and testing) and the CPRD-AURUM dataset for external validation. Both comprise large cohorts of patients, representative of the population of England in terms of age, sex and ethnicity. Primary care records are linked at patient level to secondary care and mortality data. The performance of the prediction model will be assessed by discrimination, calibration and clinical utility. We will only use variables routinely accessible in primary care.

Ethics and dissemination Permissions for CPRD-GOLD and CPRD-AURUM datasets were obtained from CPRD (ref no: 21_000324). The CPRD ethical approval committee approved the study. The results will be submitted as a research paper for publication to a peer-reviewed journal and presented at peer-reviewed conferences.

Trial registration details The study was registered on Clinical Trials.gov (NCT 05756127). A systematic review for the project was registered on PROSPERO (registration number: CRD42022380892).

INTRODUCTION

An estimated 64.3 million people are living with heart failure (HF) worldwide,¹ and the prevalence of HF is projected to increase.² HF is the most common cause of unplanned hospital admissions in older persons, and is

STRENGTHS AND LIMITATIONS OF THIS STUDY

- ⇒ Large and nationwide dataset representative of the UK primary care population.
- ⇒ Investigation of regression and machine learning techniques for the derivation of a prediction model for incident heart failure in the short and long term.
- ⇒ Candidate variable data types are deliberately limited to ensure widespread applicability of the model given the reality of ‘missing’ data in routinely collected electronic health records.
- ⇒ This study is designed to fill an implementation gap to enable electronic health records to provide decision support to primary care physicians.
- ⇒ The derivation and validation work will be undertaken in datasets collected in the UK; therefore, further validation work may be pursued for international contexts.

associated with reduced quality of life and premature mortality.^{3–6} Advances in the treatment of HF have offered improvements in prognosis,^{7–9} however, many cases of HF present and are diagnosed and treated late in course of the disease.^{2 10}

International guidelines define four stages of HF: Stage A HF (at-risk for HF), Stage B HF (pre-HF; structural heart disease without symptoms), Stage C HF (symptomatic HF), and Stage D HF (advanced HF).^{7 11} Mortality increases with progression through the stages. Accordingly, guidelines recommend initiatives to identify individuals with Stage A and Stage B HF as evidence supports that the onset of symptomatic HF can be delayed or prevented by targeting modifiable risk factors.¹²

In the UK, 98% of the populace are registered in primary care and have electronic health records (EHRs).¹³ A decision tool that exploits routinely collected EHR data across a population to calculate HF risk could offer

a scalable, efficient and cost-effective approach to identifying individuals with Stage A/B HF.¹⁴ Previous models applicable to community-based EHRs to predict HF risk have been limited. Models have seldom been externally validated,^{15 16} which prohibits an understanding of their generalisability. Many have been developed in curated prospective cohort studies, and their performance may not translate to EHR data.^{16 17} Others include laboratory results (eg, natriuretic peptide measurement),¹⁸ specialist investigations (eg, cardiac magnetic resonance (CMR))¹⁹ or observations (eg, blood pressure and body mass index)^{17 20} that are missing in the majority of primary care EHRs and which may limit their scalability and applicability across the population.²¹ Predictive models developed using deep learning have yet to report calibration performance and may be limited in clinical application by explainability.²² Furthermore, models have either provided risk prediction over short (6 months) or long prediction horizons (10 years),^{16 22} and therefore may not be used to both inform targeting of diagnostics and primary prevention initiatives.

The Clinical Practice Research Datalink (CPRD) is an ongoing primary care database, established in 1987, that comprises anonymised medical records and prescribing data from a network of General Practices (GPs) across the UK.¹³ CPRD undertakes over 900 checks covering the integrity, structure and format of the daily GP data collection and is an optimal tool for undertaking real-world, population-based evaluations of healthcare as well as the development of clinical prediction models.^{13 23}

Developing a prediction model for HF from routinely collected primary care EHR data could offer several advantages. A model created from widely available data in routinely collected EHRs could be translated into clinical practice by being embedded into existing clinical EHRs. Furthermore, a model embedded in EHRs could give risk prediction for incident HF over the next 1–10 years that is updated each time an individual's clinical situation changes (eg, age, diagnoses recorded), which more accurately reflects the dynamic nature of disease pathogenesis and clinical decision making.

RESEARCH AIM

The aim of the study is to develop and validate a model for predicting incident HF from national primary care EHRs (figure 1). Specifically, we wish to develop a model that is widely applicable and scalable in routinely collected community-based EHRs, test its performance across a range of prediction horizons, and externally validate it in a geographically distinct primary care EHR dataset.

METHODS AND ANALYSIS

Data sources and permissions

The derivation dataset for training and testing the model will be the CPRD-GOLD dataset. This is an ongoing primary care database, established in 1987, that comprises anonymised medical records and prescribing data contributed by GPs using Vision software.¹³ It contains data for

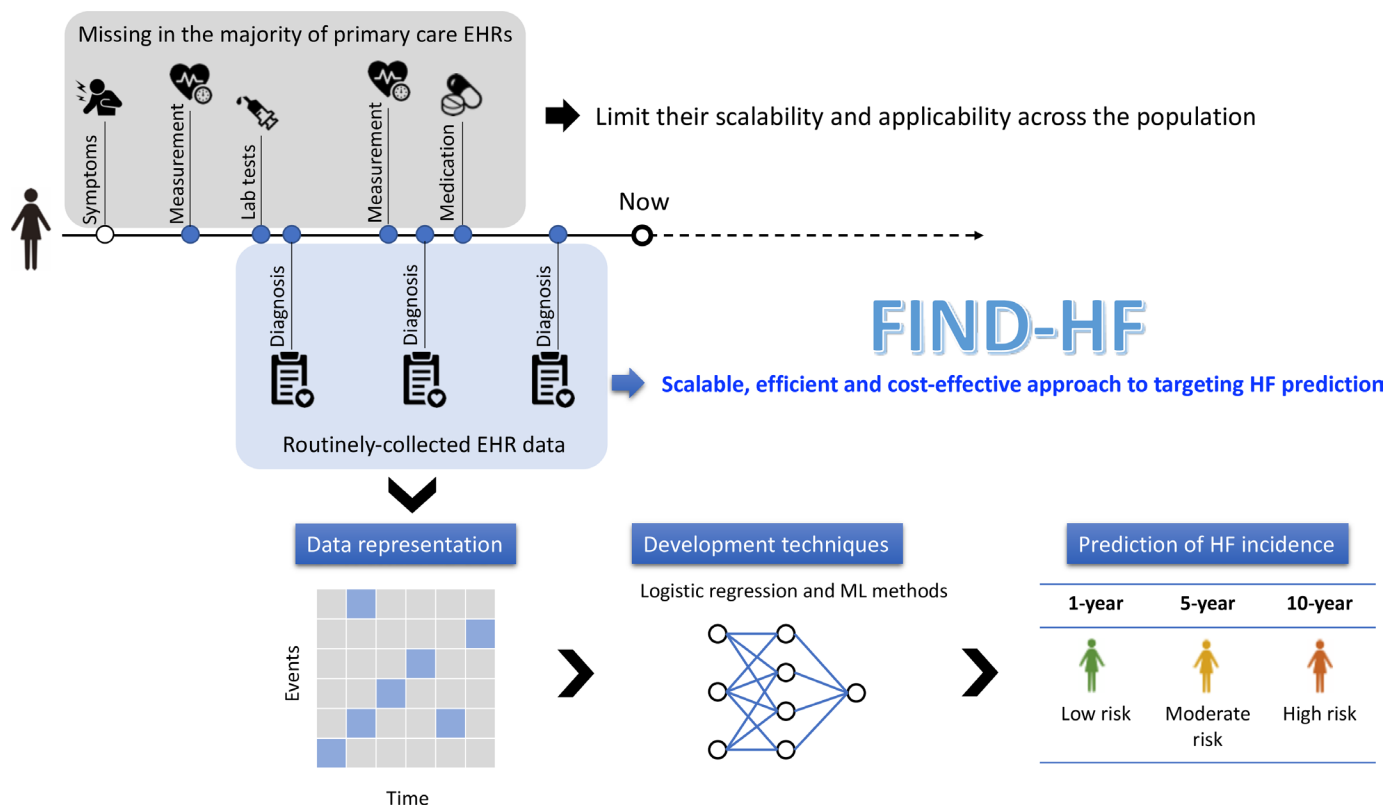


Figure 1 Schematic representation of the FIND-HF study.

approximately 17.5 million patients, with 30% of contributing practices in England.¹³ The included patients are broadly representative of the UK general population regarding age, sex and ethnicity.¹³ In order to contribute to the database, GPs and other health centres must meet prespecified standards for research-quality data ('up-to-standard').^{13 24}

To ascertain whether the prediction model is generalisable, we will externally validate its performance in the geographically distinct CPRD-AURUM dataset. This was launched in 2017 and encompasses only practices using EMIS Web software. It contains data for approximately 26.9 million patients and draws on data collected from practices in England only.²⁵ Any practices which previously contributed to CPRD-GOLD have been removed from the CPRD-GOLD cohort to ensure that these datasets reflect different populations. CPRD undertakes various levels of validation and quality assurance on the daily GP data collection comprising over 900 checks covering the integrity, structure and format of the data.²⁵ The included patients are broadly representative of the UK general population regarding age, sex, deprivation and geographical spread.²⁵

There is the possibility that patients may transfer from a practice in GOLD to a practice in AURUM or vice versa, but the proportion of transfers is small. In the study, we will ensure that the study period starts from registration with a practice and is censored from the date of transfer out. Therefore, there is no overlapping period for the same patient in the training/testing set and the validation set.

Recorded information in both datasets includes patients' demography, clinical symptoms, signs, investigations, diagnoses, prescriptions, referrals, behavioural factors and test results entered by clinicians and other practice staff. All clinical information is coded using Read Codes in CPRD-GOLD and SNOMED clinical terms (CT) in CPRD-AURUM.^{26 27} In the proposed study, extracted patients will have patient-level data linked to Hospital Episode Statistics (HES) Admitted Patient Care (APC) and Diagnostic Imaging Dataset (DID), Office for National Statistics (ONS) Death Registration, patient-level deprivation and practice-level deprivation to provide a more comprehensive dataset. The CPRD dataset has been used to develop or validate a range of risk prediction models, including in cardiovascular disease.^{23 28}

Patient and public involvement

Patients and public were not involved in the design of this research. However, we are convening a Scientific Advisory Board to include representatives from patients and public involvement groups, clinical experts, national health system leaders and EHR software providers to provide context advice on the research, dissemination of results and translation of the findings into clinical practice.

Inclusion and exclusion criteria

The study population will comprise all available patients in CPRD-GOLD and CPRD-AURUM eligible for data linkage and with at least 1-year follow-up in the period between 2 January 1998 and 28 February 2022. Patients will be excluded if they were under 16 years of age at the date of the first registration in CPRD, diagnosed with HF before 2 January 1998, registered for less than 1 year in CPRD or ineligible for data linkage.

Outcome ascertainment

The models will be developed to predict new onset HF. HF will be defined as the first presence of one or more of the clinical codes related to HF developed by consensus with clinical members of the research team. Code lists for HF will include Read codes and SNOMED CT in CPRD datasets, and the 10th revision of the International Statistical Classification of Diseases and Related Health Problems (ICD-10) codes in HES APC events and underlying cause of death variable in the ONS Death Registration data file. The first record of HF within the study period will be taken as the date of diagnosis (the index date). To that effect, in our analytical cohorts, there are about 100 000 HF cases in CPRD-GOLD and 800 000 HF cases in CPRD-AURUM. Misclassified data can lead to systematic prediction errors and accuracy of data may vary over time,²⁹ but CPRD has converted older ICD codes to the newer version, increasing confidence in their validity. Using incidence density sampling,³⁰ we will match HF cases by year of birth (± 5 years) and sex with up to five controls in the same practice on the index date without a diagnosis of HF on that date.

Predictor variables

A systematic review is being conducted to identify candidate predictors for inclusion (PROSPERO: CRD42022380892). The potential predictors will include age, sex, ethnicity and all disease conditions during follow-up. Candidate disease conditions will comprise hospitalised diseases, such as other cardiovascular diseases, obesity, diabetes mellitus, thyroid disorders, iron deficiency and anaemia, kidney dysfunction, electrolyte disorders, chronic lung disease, sleep-disordered breathing, hyperlipidaemia, gout, erectile dysfunction, depression, cancer and infection.⁷ Code lists for predictors will be used from publications if available, otherwise, the CPRD code browser will be used and codes checked by at least two clinicians. The code lists for predictors in GOLD and AURUM will be adapted from CALIBER and HDR UK repositories or publications. If none are available from these sources then new code lists will be developed using the OpenCodelists and checked by at least two clinicians.

For diagnoses, if medical codes are absent in a patient record, we will assume that the patient does not have that diagnosis, or that the diagnosis was not considered sufficiently important to have been recorded by the GP in case of symptoms.³¹ Ethnicity information is routinely

collected in the UK NHS and so has increasingly high completeness,³² and we will include an 'ethnicity unrecorded' category where it is unavailable because missingness is considered to be informative.³³ Accordingly, we do not expect any missing data for any of the predictor variables in the analytical cohort.

Sample size

To develop a prognostic prediction model, the required sample size may be determined by three criteria suggested by Riley *et al.*³⁴ For example, suppose a maximum of 200 parameters will be included in the prediction model and the Cox-Snell generalised R^2 is assumed to be 0.01. A total of 377 996 patients will be required to meet Riley's criterion (1) with global shrinkage factor of 0.95; this sample size also ensures a small absolute difference ($\Delta < 0.05$) in the apparent and adjusted Nagelkerke R^2 (Riley's criterion (2)) and ensures precise estimate of overall risk with a margin of error < 0.001 (Riley's criterion (3)). According to the Quality and Outcomes Framework, the prevalence of HF in England is 0.91%. Given an HF prevalence of 0.91%, only 3439 patients will be expected to develop HF from 377 996 patients. Therefore, the number of patients in the CPRD dataset with HF will provide sufficient statistical power to develop and validate a prediction model with the predefined precision and accuracy.

Data analysis plan

Data preprocessing

The CPRD-GOLD and CPRD-AURUM data will be cleaned and preprocessed for model development and validation, respectively. For categorical variables, we will address data quality issues such as inconsistent formatting and encoding errors, ensure categories are properly defined and resolve any inconsistencies in their representation to maintain data integrity. For patient features with binary values (sex and disease conditions), 0 and 1 will be mapped to the binary values. Continuous variable (age) will be kept as continuous and we will employ statistical techniques to identify potential outliers (including the use of z-scores and inspection of the distribution of the variables). Preprocessed patient-level data in CPRD-GOLD will be randomly split into an 80:20 ratio to create development and internal validation samples using the Mersenne twister pseudorandom number generator.

Descriptive analysis

We will perform descriptive analyses of all variables and test the statistical difference between cases and controls using the t-test for normally distributed continuous variables, Wilcoxon rank sum test for non-normally distributed a continuous variable (age) and Pearson's chi-squared test for categorical variables, using a $p \leq 0.05$ to represent significance.

Prediction model development

Our focus is on using the logistic regression model because it offers a more manageable approach for implementation, interpretation and training compared with machine

learning (ML) algorithms. However, we will compare the performance of the logistic regression model to a broad range of supervised ML techniques, including random forest, neural network, support vector machine, discriminants analysis and naïve Bayes classifier. We will check the assumptions of each ML method to assess its quality and whether it is appropriate for the data. To examine the comparative performance of the ML algorithms, we will apply Cochran's Q test, which allows for the evaluation of multiple MLs. The primary prediction window will set at 1 year.³⁵ We will also explore prediction models with the length of the prediction window set at 5 and 10 years.

Internal validation

We will evaluate the model performance using a validation cohort with internal bootstrap validation with 200 samples. The AUROC will be used to evaluate predictive ability (concordance index) with 95% CIs calculated using the DeLong method.³⁶ Youden's index will be established for the outcome measure as a method of empirically identifying the optimal dichotomous cut-off to assess sensitivity, specificity, positive predictive value and negative predictive value. We will calculate the Brier score, a measure of both discrimination and calibration, by taking the mean squared difference between predicted probabilities and the observed outcome. Calibration will be assessed graphically by plotting predicted HF risk against observed HF incidence at 1, 5 and 10 years. Overall ML performance, including distance between the predicted outcome and actual outcome, will be measured. Decision Curve Analysis will be used to assess whether the predictive model would do more benefit than harm.

Clinical utility will be examined by using net benefit analysis, where the harms and benefits of using a model to guide treatment decisions will be offset to assess the overall consequences of using the FIND-HF model for clinical decision making.³⁶ The model will be compared at 1 year, 5 years and 10 years with model blind methods of performing echocardiography for all patients, or not performing echo for all patients, regardless of HF risk. We will assess the net benefit across the full range of possible threshold probabilities with an HF risk. A priori we will set an HF risk at 1, 5 and 10 years as being the threshold of clinical interest, to align with the incidence of HF at these time points in routine practice.

The same methods will be employed in subgroups by age (< 65 years, ≥ 65 years; < 75 years, ≥ 75 years), sex (women, men), ethnicity (White, Black, Asian, others and unspecified) and HF phenotype (HF with preserved ejection fraction, HF with reduced ejection fraction) to assess the model's predictive performance in these clinically relevant groups.

External validation of model

The CPRD-AURUM dataset will then be used to externally validate the model performance to assess generalisability. A lack of external validation has hampered the implementation of previous prediction models for HF

in routine clinical practice.³⁷ The prediction model will be applied to each individual in the external validation cohort to give the predicted probabilities of experiencing HF at 1, 5 and 10 years. Prediction performance will be quantified by calculating the AUROC, Brier score, the observed to expected ratio, and by using calibration plots, and the same aforementioned clinical utility and subgroup analysis will be conducted. We will compare the performance against previously published models for incident HF that have been externally validated and are scalable in EHRs.³⁸

Software

All analysis will be conducted through Stata and R.

ETHICS AND DISSEMINATION

The study has been approved by CPRD (ref no: 21_000324). Those handling data have completed University of Leeds information security training. All analyses will be conducted in concordance with the CPRD study dataset agreement between the Secretary of State for Health and Social Care and the University of Leeds.

The study is informed by the Prognosis Research Strategy (PROGRESS) framework and recommendations.³⁹ The subsequent research paper will be submitted for publication in a peer-reviewed journal and will be written following Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD) reporting guidelines and the CODE-EHR best-practice framework.^{40 41}

If the model demonstrates evidence of clinical utility, it could be made readily available through EHR system providers. As such, each time the model is called within an EHR system, the risk score should be updated with new information so that prediction of an individual's HF risk is updated contemporaneously. The model could be a built-in tool for use in GPs for the targeted identification of individuals at high risk of developing new-onset HF. Future rigorous prospective study will be needed to assess the clinical impact and cost-effectiveness of this risk model.¹⁴ At the point when utilisation in clinical practice is possible, the applicable regulation on medicine devices will be adhered to.⁴¹ When in clinical use, the model itself could also be reviewed and updated by a prespecified expert consensus group on an annual basis after incorporating evidence from postservice utilisation and the curation of more data. The model will have to be updated as population characteristics change, data quality of EHRs improves and new or additional risk factors emerge.

Author affiliations

¹Leeds Institute of Cardiovascular and Metabolic Medicine, University of Leeds, Leeds, UK

²Leeds Institute for Data Analytics, University of Leeds, Leeds, UK

³Department of Pharmacoepidemiology, Graduate School of Medicine and Public Health, Kyoto University, Kyoto, Japan

⁴Department of Cardiology, Leeds Teaching Hospital NHS Trust, Leeds, UK

⁵Department of Cardiovascular Medicine, National Cerebral and Cardiovascular Center, Suita, Japan

⁶Carmel Medical Practice, Darlington & School of Medicine, Pharmacy and Health, Durham University, Durham, UK

⁷Stroke Road Surgery, Bishop's Cleeve, Cheltenham, UK

⁸The Phoenix Partnership Leeds Ltd, Horsforth, UK

⁹Department of Biostatistics and Health Data Science, Queen Mary University of London, London, UK

Twitter Yoko M Nakao @YokoMNakao and Ramesh Nadarajah @Dr_R_Nadarajah

Contributors CG conceived the concept and JW, YMN, FS and RN planned the analysis. YMN wrote the first draft, with contributions from all authors. RN amended the draft after comments from all coauthors. All authors approved the final version and jointly take responsibility for the decision to submit the manuscript to be considered for publication.

Funding This work was supported by the Japan Research Foundation for Healthy Aging. The funder of the study has no role in study design, data collection, data analysis, data interpretation, or writing the report.

Competing interests YMN reports a study grant from Bayer, outside the submitted work. JM reports personal fees from Bayer. He is the President of the Primary Care Cardiovascular Society. CG reports personal fees from AstraZeneca, Amgen, Bayer, Boehringer-Ingelheim, Daiichi Sankyo, Vifor, Pharma, Menarini, Wondr Medical, Raisio Group and Oxford University Press. He has received educational and research grants from BMS, Abbott inc., the British Heart Foundation, National Institute of Health Research, Horizon 2020, and from the European Society of Cardiology, outside the submitted work. All other authors declare no competing interests.

Patient and public involvement Patients and/or the public were not involved in the design, or conduct, or reporting, or dissemination plans of this research.

Patient consent for publication Not applicable.

Provenance and peer review Not commissioned; externally peer reviewed.

Open access This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: <http://creativecommons.org/licenses/by-nc/4.0/>.

ORCID iDs

Yoko M Nakao <http://orcid.org/0000-0002-3627-5626>

Ramesh Nadarajah <http://orcid.org/0000-0001-9895-9356>

REFERENCES

- James SL, Abate D, Abate KH. Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990-2017: a systematic analysis for the global burden of disease study 2017. *Lancet* 2018;392:1789-858.
- Conrad N, Judge A, Tran J, *et al*. Temporal trends and patterns in heart failure incidence: a population-based study of 4 million individuals. *Lancet* 2018;391:572-80.
- Simmonds R, Glogowska M, McLachlan S, *et al*. Unplanned admissions and the organisational management of heart failure: a multicentre ethnographic, qualitative study. *BMJ Open* 2015;5:e007522.
- Mohd Ghazi A, Teoh CK, Abdul Rahim AA. Patient profiles on outcomes in patients hospitalized for heart failure: a 10-year history of the Malaysian population. *ESC Heart Fail* 2022;9:2664-75.
- Conrad N, Judge A, Canoy D, *et al*. Temporal trends and patterns in mortality after incident heart failure: a longitudinal analysis of 86 000 individuals. *JAMA Cardiol* 2019;4:1102-11.
- Taylor CJ, Ordóñez-Mena JM, Roalfe AK, *et al*. Trends in survival after a diagnosis of heart failure in the United Kingdom 2000-2017: population based cohort study. *BMJ* 2019;364:i223.
- McDonagh TA, Metra M, Adamo M, *et al*. ESC guidelines for the diagnosis and treatment of acute and chronic heart failure: developed by the task force for the diagnosis and treatment of acute and chronic heart failure of the European society of cardiology (ESC) with the special contribution of the heart failure Association (HFA) of the ESC. *Eur Heart J* 2021;42:3599-726.



- 8 Mebazaa A, Davison B, Chioncel O, *et al.* Safety, tolerability and efficacy of up-titration of guideline-directed medical therapies for acute heart failure (STRONG-HF): a multinational, open-label, randomised, trial. *The Lancet* 2022;400:1938–52.
- 9 Tromp J, Ouwerkerk W, van Veldhuisen DJ, *et al.* A systematic review and network meta-analysis of pharmacological treatment of heart failure with reduced ejection fraction. *JACC Heart Fail* 2022;10:73–84.
- 10 Kwok CS, Burke H, McDermott S, *et al.* Missed opportunities in the diagnosis of heart failure: evaluation of pathways to determine sources of delay to specialist evaluation. *Curr Heart Fail Rep* 2022;19:247–53.
- 11 Heidenreich PA, Bozkurt B, Aguilar D, *et al.* AHA/ACC/HFSA guideline for the management of heart failure: A report of the American college of cardiology/American heart Association joint committee on clinical practice guidelines. *Circulation* 2022;145:e895–1032.
- 12 Jafari LA, Suen RM, Khan SS. Refocusing on the primary prevention of heart failure. *Curr Treat Options Cardiovasc Med* 2020;22:13.
- 13 Herrett E, Gallagher AM, Bhaskaran K, *et al.* Data resource profile: clinical practice research Datalink (CPRD). *Int J Epidemiol* 2015;44:827–36.
- 14 Olsen CR, Mentz RJ, Anstrom KJ, *et al.* Clinical applications of machine learning in the diagnosis, classification, and prediction of heart failure. *Am Heart J* 2020;229:1–17.
- 15 Goyal A, Norton CR, Thomas TN, *et al.* Predictors of incident heart failure in a large insured population: a one million person-year follow-up study. *Circ Heart Fail* 2010;3:698–705.
- 16 Agarwal SK, Chambless LE, Ballantyne CM, *et al.* Prediction of incident heart failure in general practice: the Atherosclerosis risk in communities (ARIC) study. *Circ Heart Fail* 2012;5:422–9.
- 17 Chahal H, Bluemke DA, Wu CO, *et al.* Heart failure risk prediction in the multi-ethnic study of Atherosclerosis. *Heart* 2015;101:58–64.
- 18 Arshi B, van den Berge JC, van Dijk B, *et al.* Implications of the ACC/AHA risk score for prediction of heart failure: the Rotterdam study. *BMC Med* 2021;19:43.
- 19 Bradley J, Schelbert EB, Bonnett LJ, *et al.* Predicting Hospitalisation for heart failure and death in patients with, or at risk of, heart failure before first Hospitalisation: a retrospective model development and external validation study. *Lancet Digit Health* 2022;4:e445–54.
- 20 Brouwers FP, van Gilst WH, Damman K, *et al.* Clinical risk stratification Optimizes value of biomarkers to predict new-onset heart failure in a community-based cohort. *Circ Heart Fail* 2014;7:723–31.
- 21 Nadarajah R, Wu J, Hogg D, *et al.* Prediction of short-term atrial fibrillation risk using primary care electronic health records. *Heart* 2023;109:1072–9.
- 22 Rao S, Li Y, Ramakrishnan R, *et al.* An Explainable transformer-based deep learning model for the prediction of incident heart failure. *IEEE J Biomed Health Inform* 2022;26:3362–72.
- 23 Nadarajah R, Wu J, Frangi AF, *et al.* Predicting patient-level new-onset atrial fibrillation from population-based nationwide electronic health records: protocol of FIND-AF for developing a precision medicine prediction model using artificial intelligence. *BMJ Open* 2021;11:e052887.
- 24 Herrett E, Thomas SL, Schoonen WM, *et al.* Validation and validity of diagnoses in the general practice research database: a systematic review. *Br J Clin Pharmacol* 2010;69:4–14.
- 25 Wolf A, Dedman D, Campbell J, *et al.* Data resource profile: clinical practice research datalink (CPRD) aurum. *Int J Epidemiol* 2019;48:1740–1740g.
- 26 Chisholm J. The read clinical classification. *BMJ* 1990;300:1092.
- 27 SNOMED clinical terms: overview of the development process and project status. Proceedings of the AMIA Symposium; American Medical Informatics Association, 2001
- 28 Hippisley-Cox J, Coupland C, Brindle P. Development and validation of Qrisk3 risk prediction algorithms to estimate future risk of cardiovascular disease: prospective cohort study. *BMJ* 2017;357:j2099.
- 29 Ehrenstein V, Nielsen H, Pedersen AB, *et al.* Clinical epidemiology in the era of big data: new opportunities, familiar challenges. *Clin Epidemiol* 2017;9:245–50.
- 30 Etminan M. Pharmacoepidemiology II: the nested case-control study—a novel approach in Pharmacoepidemiologic research. *Pharmacotherapy* 2004;24:1105–9.
- 31 Elwenspoek MMC, O'Donnell R, Jackson J, *et al.* Development and external validation of a clinical prediction model to aid coeliac disease diagnosis in primary care: an observational study. *EClinicalMedicine* 2022;46.
- 32 Routen A, Akbari A, Banerjee A, *et al.* Strategies to record and use Ethnicity information in routine health data. *Nat Med* 2022;28:1338–42.
- 33 Groenwold RHH. Informative Missingness in electronic health record systems: the curse of knowing. *Diagn Progn Res* 2020;4:8.
- 34 Riley RD, Snell KI, Ensor J, *et al.* Minimum sample size for developing a multivariable prediction model: PART II-Binary and Time-To-Event outcomes. *Stat Med* 2019;38:1276–96.
- 35 Chen R, Stewart WF, Sun J, *et al.* Recurrent neural networks for early detection of heart failure from longitudinal electronic health record data: implications for temporal modeling with respect to time before diagnosis, data density, data quantity, and data type. *Circ Cardiovasc Qual Outcomes* 2019;12:10.
- 36 DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a Nonparametric approach. *Biometrics* 1988;44:837–45.
- 37 Banerjee A, Chen S, Fatemifar G, *et al.* Machine learning for subtype definition and risk prediction in heart failure, acute coronary syndromes and atrial fibrillation: systematic review of validity and clinical utility. *BMC Med* 2021;19:85.
- 38 Bavishi A, Bruce M, Ning H, *et al.* Predictive accuracy of heart failure-specific risk equations in an electronic health record-based cohort. *Circ Heart Fail* 2020;13:11.
- 39 Steyerberg EW, Moons KGM, van der Windt DA, *et al.* Prognosis research strategy (PROGRESS) 3: prognostic model research. *PLoS Med* 2013;10.
- 40 Collins GS, Reitsma JB, Altman DG, *et al.* Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *Circulation* 2015;131:211–9.
- 41 Kotecha D, Asselbergs FW, Achenbach S, *et al.* CODE-EHR best-practice framework for the use of structured electronic health-care records in clinical research. *Lancet Digit Health* 2022;4:e757–64.