BMJ Open

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

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2023-073455
Article Type:	Protocol
Date Submitted by the Author:	06-Mar-2023

Complete List of Authors:	Nakao, Yoko; University of Leeds, Nadarajah, Ramesh; University of Leeds, Leeds Institute for Data Analytics; University of Leeds, Leeds Institute of Cardiovascular and Metabolic Medicine Shuweihdi, Farag; University of Leeds, Leeds Institute of Health Sciences Nakao, Kazuhiro; University of Leeds Fuat, Ahmet; Carmel Medical Practice, Moore, Jim; Bishop's Cleeve, Stroke Road Surgery Bates, Christopher; The Phoenix Partnership Leeds Ltd Wu, Jianhua; Queen Mary University of London Gale, Chris; University of Leeds

Keywords:	Heart failure < CARDIOLOGY, Primary Care < Primary Health Care, Primary Prevention

Title

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

Authors

Yoko M Nakao[1,2*], Ramesh Nadarajah[1,2,3*], Farag Shuweihdi[1], Kazuhiro Nakao[1,2,4], Ahmet Fuat[5], Jim Moore[6], Chris Bates[7], Jianhua Wu[1,8], Chris P Gale[1,2,3]

Affiliations:

[1] Leeds Institute for Cardiovascular and Metabolic Medicine, University of Leeds, UK

[2] Leeds Institute of Data Analytics, University of Leeds, UK

[3] Department of Cardiology, Leeds Teaching Hospitals NHS Trust, Leeds, UK

[4] Department of Cardiovascular Medicine, National Cerebral and Cardiovascular Center, Suita, Japan

[5] Carmel Medical Practice, Darlington & School of Medicine, Pharmacy and Health, Durham University, Durham, UK

[6] Stoke Road Surgery, Bishop's Cleeve, Cheltenham, UK

[7] The Phoenix Partnership, Leeds, UK

[8] Department of Biostatistics and Health Data Science, Queen Mary University of London, London, UK

*The first two authors are joint first authors.

Corresponding author

Yoko M Nakao

Leeds Institute for Cardiovascular and Metabolic Medicine

University of Leeds

6 Clarendon Way

Leeds, UK

LS2 9DA

Tel +44 (0) 113 343 3241

Email Y.Nakao@leeds.ac.uk

Twitter @YokoMNakao

Word count

2861 words

Keywords

Heart failure; Prediction; Electronic health records; Primary care; Screening; Prevention

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. Earlier identification and treatment of HF could reduce downstream healthcare impact, but predicting incident HF is challenging and statistical models are limited by performance and scalability in routine clinical practice. A HF prediction model developed in nationwide electronic health records (EHRs) could provide a scalable solution.

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 a large, representative population of England 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 (NCT05756127). A systematic review for the project was registered on PROSPERO (registration number: CRD42022380892).

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 heart failure prediction model for short and long term prediction horizons.
- 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.

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 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.[10 11] Furthermore reported differences in the way heart failure is diagnosed and managed have changed little in the past decade. Variable access to diagnostic tests, modes of care delivery and non-uniform management approaches persist.[12]

Screening and primary prevention for HF is advocated in international guidelines. In the 2022 AHA/ACC/HFSA Guideline for the Management of Heart Failure, natriuretic peptide-based screening followed by team-based care is included as a Class IIa recommendation.[13] However, the discrimination performance of natriuretic peptides for HF is only moderate (area under the receiver operating characteristic curve [AUROC] 0.60-0.70),[14] and would require an enormous burden of patients to undergo an additional blood test. Moreover, unselected screening for HF is unlikely to return a high yield of cases or be cost effective.[15] Alternative approaches that identify an enriched target population suitable for screening and prevention initiatives are warranted.

In the UK, 98% of the populace are registered in primary care and have electronic health records (EHRs).[16] 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 targeting diagnostics for HF.[17] Previous models applicable to community-based EHRs to predict HF risk have been limited. Models have seldom been externally validated,[18 19] which prohibits an understanding of their generalizability. Many have been developed in curated prospective cohort studies, and their performance may not translate to EHR data.[19 20] Others include lab results (e.g. natriuretic peptide measurement),[21] specialist investigations (e.g. cardiac MRI)[22] or observations (e.g. blood pressure and body mass index)[20 23] that are missing in the majority of primary care EHRs and which may limit their scalability and applicability across the population.[24] Predictive models developed using deep learning have yet to report calibration performance, and may be limited in clinical application by explainability.[25] Furthermore, models have either provided risk prediction over short (6 months) or long prediction horizons (10 years),[19 25] 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.[25] 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 health care as well as the development of clinical prediction models.[26 27]

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 (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 heart failure from national primary care EHRs. 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 general practices using Vision software.[26] It contains data for approximately 17.5 million patients, with 30% of contributing practices in England.[26] The included patients are broadly representative of the UK general population regarding age, sex and ethnicity.[26] In order to contribute to the database, general practices and other health centres must meet pre-specified standards for research-quality data ('up-to-standard').[16,28]

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.[29] 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 general practice data collection comprising over 900 checks covering the integrity, structure and format of the data.[29] The included patients are broadly representative of the UK general population regarding age, sex, deprivation and geographical spread.[29]

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.[30,31] 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.[27 32]

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,[32] but CPRD has converted older ICD codes to the newer version, increasing confidence in their validity. Using incidence density sampling,[34] 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).[35] 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 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.[36]

Sample size

To develop a prognostic prediction model, the required sample size may be determined by three criteria suggested by Riley et al.[37] 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 R2 (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 pre-processing

The CPRD-GOLD and CPRD-AURUM data will be cleaned and pre-processed for model development and validation, respectively. Specifically, 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.

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.

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 to 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.[38] 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% confidence intervals calculated using the DeLong method.[39] 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 assesses 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.[40] 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 a HF risk. A priori we will set a 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 heart failure in routine clinical practice.[41] 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.[42]

Software

All analysis will be conducted through STATA (version 17) 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.[43] 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.[44 45]

If the model demonstrates evidence of clinical utility, it could be made readily available through free-to-use software. The model will be designed to be amenable to in situ updating 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 general practices 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.[16] At the point when utilisation in clinical practice is possible, the applicable regulation on medicine devices will be adhered to.[46] When in clinical use, the model itself could also be reviewed and updated by a pre-specified expert consensus group on an annual basis after incorporating evidence from post-service utilisation and the curation of more data.

Conclusions

Heart failure is a common clinical problem with high healthcare burden. A prediction model that may identify in a community setting individuals at higher risk of incident HF could enable targeted investigation and primary prevention to reduce downstream morbidity, mortality and healthcare costs. This study has been designed to develop a widely-applicable and scalable HF-risk prediction model within existing healthcare structures to maximise the opportunity to translate this research for patient benefit.

Data availability statement

Patient-level data will not be made available.

Ethics Approval

Permissions for CPRD-GOLD and CPRD-AURUM datasets were obtained from CPRD (ref no: 21_000324). The CPRD ethical approval committee approved the study.

Patient consent for publication

Not required.

Contributions

CPG 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 co-authors. 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.

Conflict of 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. CPG reports personal fees from AstraZeneca, Amgen, Bayer, Boehrinher-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.

References

1. 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(10159):1789-858. doi: 10.1016/s0140-6736(18)32279-7 [published Online First: 2018/11/30]

2. Conrad N, Judge A, Tran J, et al. Temporal trends and patterns in heart failure incidence: a population-based study of 4 million individuals. The Lancet 2018;391(10120):572-80.

3. Simmonds R, Glogowska M, McLachlan S, et al. Unplanned admissions and the organisation-al management of heart failure: a multicentre ethnographic, qualitative study. BMJ open 2015;5(10):e007522.

4. Mohd Ghazi A, Teoh CK, Abdul Rahim AA. Patient profiles on outcomes in patients hospital-ized for heart failure: a 10-year history of the Malaysian population. ESC Heart Fail 2022;9(4):2664-75. doi: 10.1002/ehf2.13992 [published Online First: 20220602]

5. 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(11):1102-11. doi: 10.1001/jamacardio.2019.3593 [published Online First: 2019/09/04]

6. 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:l223. doi: 10.1136/bmj.l223

7. McDonagh TA, Metra M, Adamo M, et al. 2021 ESC Guidelines for the diagnosis and treat-ment 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. European Heart Journal 2021;42(36):3599-726. doi: 10.1093/eurheartj/ehab368

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 multination-al, open-label, randomised, trial. The Lancet 2022;400(10367):1938-52. doi: 10.1016/S0140-6736(22)02076-1

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(2):73-84. doi: 10.1016/j.jchf.2021.09.004 [published Online First: 20211208]

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(4):247-53. doi: 10.1007/s11897-022-00551-4 [published Online First: 20220606]

11. 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(10120):572-80. doi: 10.1016/s0140-6736(17)32520-5 [published Online First: 2017/11/28]

12. Hancock HC, Close H, Fuat A, et al Barriers to accurate diagnosis and effective management of heart failure have not changed in the past 10 years: a qualitative study and national survey BMJ Open 2014;4:e003866. doi: 10.1136/bmjopen-2013-003866

13. Heidenreich PA, Bozkurt B, Aguilar D, et al. 2022 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(18):e895-e1032. doi: 10.1161/cir.0000000000001063 [published Online First: 20220401]

14. Bhalla V, Isakson S, Bhalla MA, et al. Diagnostic ability of B-type natriuretic peptide and impedance cardiography: testing to identify left ventricular dysfunction in hypertensive patients. Am J Hypertens 2005;18(S2):73S-81S.

15. Don-Wauchope AC, Santaguida PL, McKelvie R, et al. Prediction of clinical outcomes using B-type natriuretic peptides in the general population: a systematic review. Heart Fail Rev 2014;19(4):541-51. doi: 10.1007/s10741-014-9446-7

16. Herrett E, Gallagher AM, Bhaskaran K, et al. Data Resource Profile: Clinical Practice Re-search Datalink (CPRD). Int J Epidemiol 2015;44(3):827-36. doi: 10.1093/ije/dyv098 [published Online First: 2015/06/08]

17. Olsen CR, Mentz RJ, Anstrom KJ, et al. Clinical applications of machine learning in the diag-nosis, classification, and prediction of heart failure. Am Heart J 2020;229:1-17. doi: 10.1016/j.ahj.2020.07.009 [published Online First: 20200716]

18. 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(6):698-705.

19. Agarwal SK, Chambless LE, Ballantyne CM, et al. Prediction of incident heart failure in gen-eral practice: the Atherosclerosis Risk in Communities (ARIC) Study. Circ Heart Fail 2012;5(4):422-29.

20. Chahal H, Bluemke DA, Wu CO, et al. Heart failure risk prediction in the Multi-Ethnic Study of Atherosclerosis. Heart 2015;101(1):58-64.

21. 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(1):1-10.

22. 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 retrospec-tive model development and external validation study. The Lancet Digital Health 2022;4(6):e445-e54.

23. 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(5):723-31.

24. Nadarajah R, Wu J, Hogg D, et al. Prediction of short-term atrial fibrillation risk using prima-ry care electronic health records. Heart 2023

25. Rao S, Li Y, Ramakrishnan R, et al. An explainable Transformer-based deep learning model for the prediction of incident heart failure. IEEE Journal of Biomedical and Health In-formatics 2022;26(7):3362-72.

26. Herrett E, Gallagher AM, Bhaskaran K, et al. Data resource profile: clinical practice research datalink (CPRD). Int J Epidemiol 2015;44(3):827-36.

27. 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 devel-oping a precision medicine prediction model using artificial intelligence. BMJ Open 2021;11(11):e052887. doi: 10.1136/bmjopen-2021-052887 [published Online First: 20211102]

28. Herrett E, Thomas SL, Schoonen WM, et al. Validation and validity of diagnoses in the Gen-eral Practice Research Database: a systematic review. British journal of clinical pharma-cology 2010;69(1):4-14.

29. Wolf A, Dedman D, Campbell J, et al. Data resource profile: Clinical Practice Research Data-link (CPRD) Aurum. International Journal of Epidemiology 2019;48(6):1740-40g. doi: 10.1093/ije/dyz034

30. Chisholm J. The Read clinical classification. BMJ: British Medical Journal 1990;300(6732):1092.

31. SNOMED clinical terms: overview of the development process and project status. Proceed-ings of the AMIA Symposium; 2001. American Medical Informatics Association.

32. Hippisley-Cox J, Coupland C, Brindle P. Development and validation of QRISK3 risk pre-diction algorithms to estimate future risk of cardiovascular disease: prospective cohort study. bmj 2017;357

33. Ehrenstein V, Nielsen H, Pedersen AB, et al. Clinical epidemiology in the era of big data: new opportunities, familiar challenges. Clinical epidemiology 2017;9:245.

34. Etminan M. Pharmacoepidemiology II: the nested case-control study--a novel approach in pharmacoepidemiologic research. Pharmacotherapy 2004;24(9):1105-9. doi: 10.1592/phco.24.13.1105.38083 [published Online First: 2004/10/06]

35. McDonagh TA, Metra M, Adamo M, et al. 2021 ESC Guidelines for the diagnosis and treat-ment 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(36):3599-726.

36. Elwenspoek MM, 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 observa-tional study. EClinicalMedicine 2022;46:101376.

37. Riley RD, Snell KI, Ensor J, et al. Minimum sample size for developing a multivariable pre-diction model: PART II-binary and time-to-event outcomes. Statistics in medicine 2019;38(7):1276-96.

38. 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):e005114. doi: 10.1161/circoutcomes.118.005114 [published Online First: 20191015]

39. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more cor-related receiver operating characteristic curves: a nonparametric approach. Biometrics 1988;44(3):837-45.

40. Vickers AJ, Van Calster B, Steyerberg EW. Net benefit approaches to the evaluation of pre-diction models, molecular markers, and diagnostic tests. BMJ 2016;352

41. 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 re-view of validity and clinical utility. BMC Medicine 2021;19(1):85. doi: 10.1186/s12916-021-01940-7

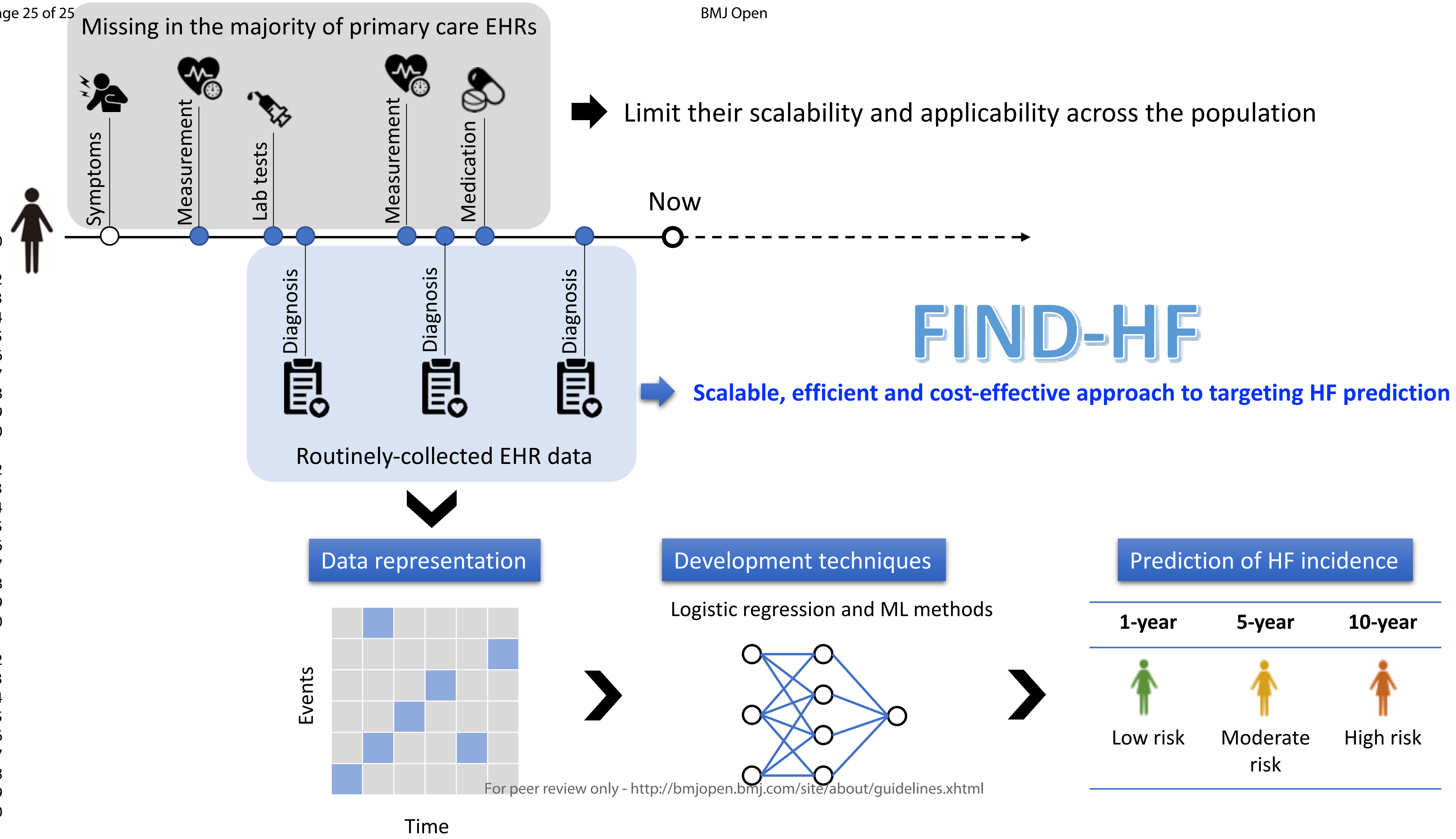
42. Bavishi A, Bruce M, Ning H, et al. Predictive accuracy of heart failure-specific risk equations in an electronic health record-based cohort. Circ Heart Fail 2020;13(11):e007462.

43. Steyerberg EW, Moons KG, van der Windt DA, et al. Prognosis Research Strategy (PRO-GRESS) 3: prognostic model research. PLoS medicine 2013;10(2):e1001381.

44. Collins GS, Reitsma JB, Altman DG, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. The TRIPOD Group. Circulation 2015;131(2):211-9. doi: 10.1161/circulationaha.114.014508 [published Online First: 20150105]

45. Kotecha D, Asselbergs FW, Achenbach S, et al. CODE-EHR best practice framework for the use of structured electronic healthcare records in clinical research. BMJ 2022;378

46. Tarricone R, Ciani O, Torbica A, et al. Lifecycle evidence requirements for high-risk im-plantable medical devices: a European perspective. Expert Review of Medical Devices 2020;17(10):993-1006.


Missing in the majority of primary care EHRs
Symptoms
Measurement
Lab tests
Measurement
Medication
Limit their scalability and applicability across the population
Now
Diagnosis
Diagnosis
Diagnosis
Routinely-collected EHR data
FIND-HF
Scalable, efficient and cost-effective approach to targeting HF prediction
Data representation
Events
Time
Development techniques
Logistic regression and ML methods
Prediction of HF incidence
1-year
5-year
10-year
Low risk
Moderate risk
High risk

BMJ Open

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

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2023-073455.R1
Article Type:	Protocol
Date Submitted by the Author:	17-Jun-2023
Complete List of Authors:	Nakao, Yoko; University of Leeds, Nadarajah, Ramesh; University of Leeds, Leeds Institute for Data Analytics; University of Leeds, Leeds Institute of Cardiovascular and Metabolic Medicine Shuweihdi, Farag; University of Leeds, Leeds Institute of Health Sciences Nakao, Kazuhiro; University of Leeds Fuat, Ahmet; Carmel Medical Practice, Moore, Jim; Bishop's Cleeve, Stroke Road Surgery Bates, Christopher; The Phoenix Partnership Leeds Ltd Wu, Jianhua; Queen Mary University of London Gale, Chris; University of Leeds
Primary Subject Heading:	Cardiovascular medicine
Secondary Subject Heading:	Epidemiology
Keywords:	Heart failure < CARDIOLOGY, Primary Care < Primary Health Care, Primary Prevention

Title

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

Authors

Yoko M Nakao[1,2,3*], Ramesh Nadarajah[1,2,4*], Farag Shuweihdi[1], Kazuhiro Nakao[1,2,5], Ahmet Fuat[5], Jim Moore[6], Chris Bates[7], Jianhua Wu[1,8], Chris P Gale[1,2,4]

Affiliations:

[1] Leeds Institute for Cardiovascular and Metabolic Medicine, University of Leeds, UK

[2] Leeds Institute of Data Analytics, University of Leeds, UK

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

[4] Department of Cardiology, Leeds Teaching Hospitals NHS Trust, Leeds, UK

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

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

[7] Stoke Road Surgery, Bishop's Cleeve, Cheltenham, UK

[8] The Phoenix Partnership, Leeds, UK

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

*The first two authors are joint first authors.

Corresponding author

Yoko M Nakao

Leeds Institute for Cardiovascular and Metabolic Medicine

University of Leeds

6 Clarendon Way

Leeds, UK

LS2 9DA

Tel +44 (0) 113 343 3241

Email Y.Nakao@leeds.ac.uk

Twitter @YokoMNakao

Word count

2867 words

Keywords

Heart failure; Prediction; Electronic health records; Primary care; Screening; Prevention

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. A 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).

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.

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 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 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 generalizability. Many have been developed in curated prospective cohort studies, and their performance may not translate to EHR data.[16 17] Others include laboratory results (e.g. natriuretic peptide measurement),[18] specialist investigations (e.g. cardiac magnetic resonance [CMR])[19] or observations (e.g. 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 health care 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 (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 heart failure from national primary care EHRs. 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 general practices using Vision software.[13] It contains data for 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, general practices and other health centres must meet pre-specified 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 general practice 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 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 un-available 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 R2 (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.

<u>Data analysis plan</u>

Data pre-processing

The CPRD-GOLD and CPRD-AURUM data will be cleaned and pre-processed 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-value ≤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 to 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% confidence intervals 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 assesses 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 a HF risk. A priori we will set a 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 heart failure 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]

<u>Software</u>

All analysis will be conducted through Stata (version 17) 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 general practices 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 pre-specified expert consensus group on an annual basis after incorporating evidence from post-service 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.

Data availability statement

Patient-level data will not be made available.

Ethics Approval

Permissions for CPRD-GOLD and CPRD-AURUM datasets were obtained from CPRD (ref no: 21_000324). The CPRD ethical approval committee approved the study.

Patient consent for publication

Not required.

Contributions

CPG 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 co-authors. 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.

Conflict of 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. CPG reports personal fees from AstraZeneca, Amgen, Bayer, Boehrinher-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.

References

1. 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(10159):1789-858. doi: 10.1016/s0140-6736(18)32279-7 [published Online First: 2018/11/30]

2. 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(10120):572-80. doi: 10.1016/s0140-6736(17)32520-5 [published Online First: 2017/11/28]

3. 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(10):e007522. doi: 10.1136/bmjopen-2014-007522 [published Online First: 20151019]

4. 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(4):2664-75. doi: 10.1002/ehf2.13992 [published Online First: 20220602]

5. 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(11):1102-11. doi: 10.1001/jamacardio.2019.3593 [published Online First: 2019/09/04]

6. 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:l223. doi: 10.1136/bmj.l223

7. McDonagh TA, Metra M, Adamo M, et al. 2021 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. *European Heart Journal* 2021;42(36):3599-726. doi: 10.1093/eurheartj/ehab368

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(10367):1938-52. doi: 10.1016/S0140-6736(22)02076-1

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(2):73-84. doi: 10.1016/j.jchf.2021.09.004 [published Online First: 20211208]

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(4):247-53. doi: 10.1007/s11897-022-00551-4 [published Online First: 20220606]

11. Heidenreich PA, Bozkurt B, Aguilar D, et al. 2022 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(18):e895-e1032. doi: 10.1161/cir.0000000000001063 [published Online First: 20220401]

12. Jafari LA, Suen RM, Khan SS. Refocusing on the Primary Prevention of Heart Failure. *Curr Treat Options Cardiovasc Med* 2020;22(7) doi: 10.1007/s11936-020-00811-3 [published Online First: 20200529]

13. Herrett E, Gallagher AM, Bhaskaran K, et al. Data Resource Profile: Clinical Practice Research Datalink (CPRD). *Int J Epidemiol* 2015;44(3):827-36. doi: 10.1093/ije/dyv098 [published Online First: 2015/06/08]

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. doi: 10.1016/j.ahj.2020.07.009 [published Online First: 20200716]

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(6):698-705. doi: 10.1161/circheartfailure.110.938175 [published Online First: 20100826]

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(4):422-9. doi: 10.1161/circheartfailure.111.964841 [published Online First: 20120515]

17. Chahal H, Bluemke DA, Wu CO, et al. Heart failure risk prediction in the Multi-Ethnic Study of Atherosclerosis. *Heart* 2015;101(1):58-64. doi: 10.1136/heartjnl-2014-305697 [published Online First: 20141107]

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(1):43. doi: 10.1186/s12916-021-01916-7 [published Online First: 20210216]

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(6):e445-e54. doi: 10.1016/s2589-7500(22)00045-0 [published Online First: 20220510]

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(5):723-31. doi: 10.1161/circheartfailure.114.001185 [published Online First: 20140723]

21. Nadarajah R, Wu J, Hogg D, et al. Prediction of short-term atrial fibrillation risk using primary care electronic health records. *Heart* 2023 doi: 10.1136/heartjnl-2022-322076 [published Online First: 20230209]

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(7):3362-72. doi: 10.1109/jbhi.2022.3148820 [published Online First: 20220701]

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(11):e052887. doi: 10.1136/bmjopen-2021-052887 [published Online First: 20211102]

24. Herrett E, Thomas SL, Schoonen WM, et al. Validation and validity of diagnoses in the General Practice Research Database: a systematic review. *British journal of clinical pharmacology* 2010;69(1):4-14.

25. Wolf A, Dedman D, Campbell J, et al. Data resource profile: Clinical Practice Research Datalink (CPRD) Aurum. *International Journal of Epidemiology* 2019;48(6):1740-40g. doi: 10.1093/ije/dyz034

26. Chisholm J. The Read clinical classification. *BMJ: British Medical Journal* 1990;300(6732):1092.

27. SNOMED clinical terms: overview of the development process and project status. Proceedings of the AMIA Symposium; 2001. American Medical Informatics Association.

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

29. Ehrenstein V, Nielsen H, Pedersen AB, et al. Clinical epidemiology in the era of big data: new opportunities, familiar challenges. *Clinical epidemiology* 2017;9:245.

30. Etminan M. Pharmacoepidemiology II: the nested case-control study--a novel approach in pharmacoepidemiologic research. *Pharmacotherapy* 2004;24(9):1105-9. doi: 10.1592/phco.24.13.1105.38083 [published Online First: 2004/10/06]

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 doi: 10.1016/j.eclinm.2022.101376

32. Routen A, Akbari A, Banerjee A, et al. Strategies to record and use ethnicity information in routine health data. *Nat Med* 2022;28(7):1338-42. doi: 10.1038/s41591-022-01842-y

33. Groenwold RHH. Informative missingness in electronic health record systems: the curse of knowing. *Diagn Progn Res* 2020;4:8. doi: 10.1186/s41512-020-00077-0 [published Online First: 20200702]

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. *Statistics in medicine* 2019;38(7):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):e005114. doi: 10.1161/circoutcomes.118.005114 [published Online First: 20191015]

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(3):837-45.

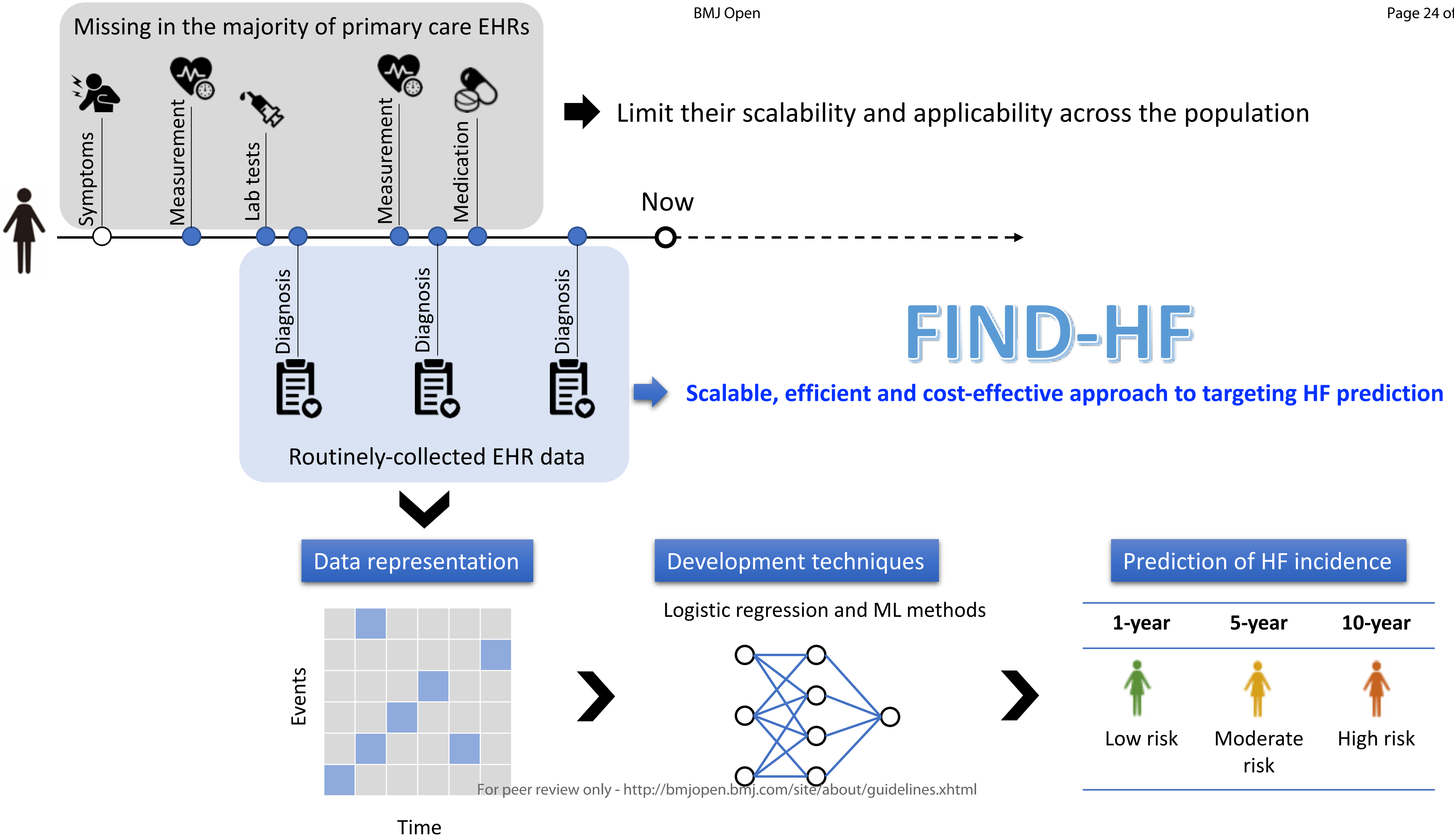
37. Banerjee A, Chen S, Fatemifar G, et al. Machine learning for subtype definition and risk prediction in heart failure, acute coronary syndromes and atrial fibrillation: systematic review of validity and clinical utility. *BMC Medicine* 2021;19(1):85. doi: 10.1186/s12916-021-01940-7

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):e007462. doi: 10.1161/circheartfailure.120.007462 [published Online First: 20201023]

39. Steyerberg EW, Moons KG, van der Windt DA, et al. Prognosis Research Strategy (PROGRESS) 3: prognostic model research. *PLoS medicine* 2013;10(2):e1001381.

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. The TRIPOD Group. *Circulation* 2015;131(2):211-9. doi: 10.1161/circulationaha.114.014508 [published Online First: 20150105]

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(10):e757-e64. doi: 10.1016/s2589-7500(22)00151-0 [published Online First: 20220829]


Missing in the majority of primary care EHRs
Symptoms
Measurement
Lab tests
Measurement
Medication
Limit their scalability and applicability across the population
Now
Diagnosis
Diagnosis
Diagnosis
Routinely-collected EHR data
FIND-HF
Scalable, efficient and cost-effective approach to targeting HF prediction
Data representation
Events
Time
Development techniques
Logistic regression and ML methods
Prediction of HF incidence
1-year
5-year
10-year
Low risk
Moderate risk
High risk

TRIPOD Checklist: Prediction Model Development and Validation

Section/Topic	Item		Checklist Item	Page
Title and abstract				
Title	1	D;V	Identify the study as developing and/or validating a multivariable prediction model, the target population, and the outcome to be predicted.	Title
Abstract	2	D;V	Provide a summary of objectives, study design, setting, participants, sample size, predictors, outcome, statistical analysis, results, and conclusions.	Abstract
Introduction				
Background and objectives	3a	D;V	Explain the medical context (including whether diagnostic or prognostic) and rationale for developing or validating the multivariable prediction model, including references to existing models.	Introduction
	3b	D;V	Specify the objectives, including whether the study describes the development or validation of the model or both.	Research Aim
Methods				
Source of data	4a	D;V	Describe the study design or source of data (e.g., randomized trial, cohort, or registry data), separately for the development and validation data sets, if applicable.	Methods and analysis – Data sources and permissions
	4b	D;V	Specify the key study dates, including start of accrual; end of accrual; and, if applicable, end of follow-up.	Methods and analysis – Inclusion and exclusion criteria
Participants	5a	D;V	Specify key elements of the study setting (e.g., primary care, secondary care, general population) including number and location of centres.	Methods and analysis – Data sources and permissions
	5b	D;V	Describe eligibility criteria for participants.	Methods and analysis – Inclusion and exclusion criteria
	5c	D;V	Give details of treatments received, if relevant.	N/A
Outcome	6a	D;V	Clearly define the outcome that is predicted by the prediction model, including how and when assessed.	Methods and analysis – Outcome ascertainment
	6b	D;V	Report any actions to blind assessment of the outcome to be predicted.	N/A
Predictors	7a	D;V	Clearly define all predictors used in developing or validating the multivariable prediction model, including how and when they were measured.	Methods and analysis – Predictor variables
	7b	D;V	Report any actions to blind assessment of predictors for the outcome and other predictors.	N/A
Sample size	8	D;V	Explain how the study size was arrived at.	Methods and analysis – Sample size
Missing data	9	D;V	Describe how missing data were handled (e.g., complete-case analysis, single imputation, multiple imputation) with details of any imputation method.	Methods and analysis – Data analysis plan Data pre-processing
Statistical analysis methods	10a	D	Describe how predictors were handled in the analyses.	Methods and analysis – Predictor variables
	10b	D	Specify type of model, all model-building procedures (including any predictor selection), and method for internal validation.	Methods and analysis – Data analysis plan Prediction model development
	10c	V	For validation, describe how the predictions were calculated.	Methods and analysis – Internal validation; External validation of the model
	10d	D;V	Specify all measures used to assess model performance and, if relevant, to compare multiple models.	Methods and analysis – Internal validation; External validation of the model
	10e	V	Describe any model updating (e.g., recalibration) arising from the validation, if done.	N/A
Risk groups	11	D;V	Provide details on how risk groups were created, if done.	Methods and analysis –

TRIPOD Checklist: Prediction Model Development and Validation

				Internal validation
Development vs. validation	12	V	For validation, identify any differences from the development data in setting, eligibility criteria, outcome, and predictors.	Methods and analysis – Data sources and permissions; Inclusion and exclusion criteria; Outcome ascertainment; Predictor variables
Results				
Participants	13a	D;V	Describe the flow of participants through the study, including the number of participants with and without the outcome and, if applicable, a summary of the follow-up time. A diagram may be helpful.	N/A
	13b	D;V	Describe the characteristics of the participants (basic demographics, clinical features, available predictors), including the number of participants with missing data for predictors and outcome.	N/A
	13c	V	For validation, show a comparison with the development data of the distribution of important variables (demographics, predictors and outcome).	N/A
Model development	14a	D	Specify the number of participants and outcome events in each analysis.	N/A
	14b	D	If done, report the unadjusted association between each candidate predictor and outcome.	N/A
Model specification	15a	D	Present the full prediction model to allow predictions for individuals (i.e., all regression coefficients, and model intercept or baseline survival at a given time point).	N/A
	15b	D	Explain how to the use the prediction model.	N/A
Model performance	16	D;V	Report performance measures (with CIs) for the prediction model.	N/A
Model-updating	17	V	If done, report the results from any model updating (i.e., model specification, model performance).	N/A
Discussion				
Limitations	18	D;V	Discuss any limitations of the study (such as nonrepresentative sample, few events per predictor, missing data).	N/A
Interpretation	19a	V	For validation, discuss the results with reference to performance in the development data, and any other validation data.	N/A
	19b	D;V	Give an overall interpretation of the results, considering objectives, limitations, results from similar studies, and other relevant evidence.	N/A
Implications	20	D;V	Discuss the potential clinical use of the model and implications for future research.	N/A
Other information				
Supplementary information	21	D;V	Provide information about the availability of supplementary resources, such as study protocol, Web calculator, and data sets.	N/A
Funding	22	D;V	Give the source of funding and the role of the funders for the present study.	Funding

*Items relevant only to the development of a prediction model are denoted by D, items relating solely to a validation of a prediction model are denoted by V, and items relating to both are denoted D;V. We recommend using the TRIPOD Checklist in conjunction with the TRIPOD Explanation and Elaboration document.