Identifying subtypes of heart failure from three electronic health record sources with machine learning: an external, prognostic, and genetic validation study

Amitava Banerjee, Ashkan Dashtban, Suliang Chen*, Laura Pasea, Johan H Thygesen, Ghazaleh Fatemifar, Benoit Tyl, Tomasz Dyszynski, Folkert W Asselbergs, Lars H Lund, Tom Lumbers, Spiros Denaxas, Harry Hemingway*

Summary

Background Machine learning has been used to analyse heart failure subtypes, but not across large, distinct, population-based datasets, across the whole spectrum of causes and presentations, or with clinical and non-clinical validation by different machine learning methods. Using our published framework, we aimed to discover heart failure subtypes and validate them upon population representative data.

Methods In this external, prognostic, and genetic validation study we analysed individuals aged 30 years or older with incident heart failure from two population-based databases in the UK (Clinical Practice Research Datalink [CPRD] and The Health Improvement Network [THIN]) from 1998 to 2018. Pre-heart failure and post-heart failure factors (n=645) included demographic information, history, examination, blood laboratory values, and medications. We identified subtypes using four unsupervised machine learning methods (K-means, hierarchical, K-Medoids, and mixture model clustering) with 87 of 645 factors in each dataset. We evaluated subtypes for (1) external validity (across datasets); (2) prognostic validity (predictive accuracy for 1-year mortality); and (3) genetic validity (UK Biobank), association with polygenic risk score (PRS) for heart failure-related traits (n=11), and single nucleotide polymorphisms (n=12).

Findings We included 188 800, 124 262, and 9573 individuals with incident heart failure from CPRD, THIN, and UK Biobank, respectively, between Jan 1, 1998, and Jan 1, 2018. After identifying five clusters, we labelled heart failure subtypes as (1) early onset, (2) late onset, (3) atrial fibrillation related, (4) metabolic, and (5) cardiometabolic. In the external validity analysis, subtypes were similar across datasets (c-statistics: THIN model in CPRD ranged from 0·79 [subtype 3] to 0·94 [subtype 1], and CPRD model in THIN ranged from 0·79 [subtype 1] to 0·92 [subtypes 2 and 5]). In the prognostic validity analysis, 1-year all-cause mortality after heart failure diagnosis (subtype 1 0·20 [95% CI 0·14–0·25], subtype 2 0·46 [0·43–0·49], subtype 3 0·61 [0·57–0·64], subtype 4 0·11 [0·07–0·16], and subtype 5 0·37 [0·32–0·41]) differed across subtypes in CPRD and THIN data, as did risk of non-fatal cardiovascular diseases and all-cause hospitalisation. In the genetic validity analysis the atrial fibrillation-related subtype showed associations with the related PRS. Late onset and cardiometabolic subtypes were the most similar and strongly associated with PRS for hypertension, myocardial infarction, and obesity ($p<0·0009$). We developed a prototype app for routine clinical use, which could enable evaluation of effectiveness and cost-effectiveness.

Interpretation Across four methods and three datasets, including genetic data, in the largest study of incident heart failure to date, we identified five machine learning-informed subtypes, which might inform aetiological research, clinical risk prediction, and the design of heart failure trials.

Funding European Union Innovative Medicines Initiative-2.

Lancet Digit Health 2023;
5: e370–79

See **Comment** page e331

*Joint second author

Institute of Health Informatics, University College London, London, UK (Prof A Banerjee DPhil, A Dashtban PhD, S Chen PhD, L Pasea PhD, Johan H Thygesen PhD, G Fatemifar PhD, Prof F W Asselbergs PhD, T Lumbers PhD, Prof S Denaxas PhD, Prof H Hemingway FMedSci); **Health Data Research UK, London, UK** (Prof A Banerjee, Prof F W Asselbergs, T Lumbers, Prof S Denaxas, Prof H Hemingway); **Barts Health NHS Trust, London, UK** (Prof A Banerjee, T Lumbers); **Department of Cardiology (Prof A Banerjee, T Lumbers) and NIHR Biomedical Research Centre** (Prof A Banerjee, T Lumbers, Prof F W Asselbergs, Prof H Hemingway), **University College London Hospitals NHS Trust, London, UK; Medical Affairs, Pharmaceuticals, Bayer HealthCare, Paris, France** (B Tyl MD); **Medical Affairs & Pharmacovigilance, Bayer AG, Berlin, Germany** (T Dyszynski); **Amsterdam University Medical Centers, Department of Cardiology, University of Amsterdam, Amsterdam, Netherlands** (Prof F W Asselbergs); **Division of Cardiology, Department of Medicine, Karolinska Institutet, Stockholm, Sweden** (Prof L H Lund PhD); **Heart and Vascular Theme, Karolinska University Hospital, Stockholm, Sweden** (Prof L H Lund)

Correspondence to:
Prof Amitava Banerjee, Institute of Health Informatics, University College London, London NW1 2DA, UK
ami.banerjee@ucl.ac.uk

Introduction

Heart failure is a heterogeneous syndrome reflecting multiple underlying causes (European Society of Cardiology [ESC]: 13 categories and 89 individual causes).[1] Disease subtypes might be relevant, whereby single causal factors in isolation (eg, diabetes, myocardial infarction) have not necessarily improved characterisation of heart failure diagnosis[2] or prognosis, discovery of new treatments, trial design, or clinical decision-making,[3] despite causal associations for those individual risk factors.[4] Current subtype classifications, including by cause (eg, ischaemic *vs* non-ischaemic), pathophysiology (eg, primary myocardial disease *vs* secondary neurohormonal activation), anatomy (eg, left-sided *vs* right-sided), haemodynamics (hypoperfusion *vs* congestion), presentation (eg, acute *vs* chronic),[1] setting (eg, outpatient *vs* inpatient), left ventricular ejection fraction (LVEF; eg, reduced *vs* mid-range, or mildly reduced *vs* preserved),[1] symptoms (eg, New York Heart Association classes 1–4 or American Heart Association heart failure stages A–D[5]), comorbidities (eg, end-stage renal disease),[6] or biomarkers (eg, N-terminus-pro-brain natriuretic peptide

Research in context

Evidence before this study

In a systematic review from January, 2000, until February, 2022, we showed that studies of machine learning in subtyping and risk prediction in cardiovascular diseases are limited by small population size, having relatively few factors, and poor generalisability of findings due to a scarcity of external validation. We further searched PubMed, medRxiv, bioRxiv, and arXiv, using the terms “machine learning”, “heart failure”, ”subtype”, and “prediction” for relevant peer-reviewed articles and preprints in English, focusing on machine learning studies in heart failure published from Jan 1, 2015, up until Dec 31, 2022. Studies remain focused on single diseases, limited risk factors, and often a single method of machine learning. Studies rarely use subtyping and risk prediction together, and have not been externally validated across datasets. For heart failure, all subtype discovery studies have identified subtypes based on clustering, but so far with no application to clinical practice.

Added value of this study

Across two distinct, population-based datasets, we used four machine learning methods for subtyping and risk prediction with 89 causal factors as well as 556 further factors for heart failure. 87 of these 645 variables were used to identify and validate five subtypes in incident heart failure, which differentially predicted outcomes. In addition, we externally validated clinical cluster differences by exploring corresponding genetic differences in a large-scale genetic cohort. Our methods and results highlight the potential value of electronic health records and machine learning in understanding disease subtypes. Moreover, our approach to external, prognostic, and genetic validity provides a framework for validation of machine learning approaches for disease subtype discovery. The clinical utility of the research methods and the potential utility of the identified subtypes in routine care (via a prototype app) were evaluated by five clinicians.

Implications of all the available evidence

Our analyses support coordinated use of large-scale, linked electronic health records to identify and validate disease subtypes with relevance for clinical risk prediction, patient selection for trials, and future genetic research.

[NT-proBNP]),[5] have not led to precision medicine, personalised care, or targeted therapies. Incomplete knowledge of subtypes across the whole range of causal factors and population has also limited primary prevention and screening guidelines for heart failure.[7,8]

In clinical practice and research, subtypes are commonly classified by LVEF for diagnosis and prognosis, but in a 2018 machine learning study in the Swedish national heart failure registry, LVEF did not predict survival.[9] Machine learning is rarely used to identify subtypes in large, nationally representative datasets linked across health-care settings (ie, primary and secondary care), whereby so-called agnostic (ie, unrestricted by particular subgroups, variables, or stages in care pathways), unsupervised subtype discovery across risk factors might inform heart failure treatment and prevention. Moreover, studies to improve heart failure subtype classification and risk prediction have neither compared different machine learning methods in one study nor validated machine learning-based subtypes in a separate population, with few studies of risk prediction or underlying biology. Our six-stage 2021 framework (clinical relevance, patients, algorithm, internal validation, external validation, clinical utility, and effectiveness) for practical machine learning implementation might yield more clinically relevant results.[10]

Therefore, in a large population of individuals with incident heart failure and 645 factors across three population-based datasets, we aimed to use four unsupervised machine learning methods to (1) identify subtypes with clinical relevance throughout the course of heart failure, and low risk of bias for patient selection and algorithms (development); (2) demonstrate internal validility (across methods), external validility (across datasets), prognostic validility (predictive accuracy for 1-year all-cause mortality), and genetic validility (using known single nucleotide polymorphisms [SNPs] associated with heart failure; validation); and (3) develop potential clinical pathways to improve impact (clinical use and effectiveness; impact).

Methods

In this external, prognostic and genetic validation study, we used our published framework for machine learning implementation to inform our methods.[10] Ethical approval was given by the Medicines and Healthcare products Regulatory Agency Independent Scientific Advisory Committee (18_217R) Section 251 (NHS Social Care Act 2006), the Scientific Review Committee (17THIN038-A1), and the UK Biobank (15422).

Generating subtypes (development)

Clinical relevance

To improve diagnostic and prognostic prediction of heart failure, our research question was relevant to potential patient benefit by focusing on definition and risk prediction in heart failure. We used two distinct population-based primary care electronic health records (EHRs) with validity for heart failure and cardiovascular disease research (target condition applicability: whether the disease defined in data matches research questions). Primary care EHR (The Health Improvement Network [THIN][11] and Clinical Practice Research Datalink [CPRD]-GOLD) were linked by CPRD and NHS Digital (using unique national health-care identifiers) with hospital admissions (Hospital Episodes Statistics) and death registry (Office for National Statistics).[12] Both datasets are representative of the UK population, with prospective recording and follow-up (data suitability). For genetic validation, we used UK Biobank data,[13] comprising of the initial release of genotyping for a random sample of 150 000 of 502 641 participants, aged 40–69 years (recruited 2006–10), linked to primary care (approximately 50%) and secondary care (100%).

Patients

Individuals aged 30 years or older with incident heart failure between Jan 1, 1998, and Jan 1, 2018, with 1 year or more of follow-up in CPRD and THIN, were included in our study. Given overlap between THIN and CPRD practices, we avoided double counting individuals using validated methods (appendix 1 p 2).[14,15] We defined incident heart failure as first record of fatal or non-fatal, hospitalised or non-hospitalised heart failure in primary care (Read coding) or secondary care (International Statistical Classification of Diseases—10th version) based on Health Data Research UK CALIBER phenotypes (appendix 1 pp 3–4).[16] Patient informed consent was not required or provided.

Algorithm

645 types of factors were used from the EHR datasets: (1) demography (eg, age; n=16); (2) cause based on ESC classification (n=258);[1,8] (3) comorbidities (eg, depression; n=114); (4) symptoms (eg, dyspnoea; n=39); (5) medication use and persistence (by 90-day prescription gap over 1 year; heart failure and non-heart failure; n=84); (6) examination (eg, blood pressure; n=11); (7) investigations (eg, kidney function; n=24); and (8) non-cardiovascular disease factors, based on a previous machine learning study (n=99; appendix 2).[17] Existing phenotypes were used, if possible, and new disease phenotypes were developed using a standardised, rule-based approach (n=23).[12] Factors were classified as before (in the 5 years before or at time of heart failure diagnosis—eg, previous ACE-inhibitor treatment), after (post 2-year follow-up—eg, ACE-inhibitor treatment after heart failure diagnosis), or ever (before or after heart failure diagnosis). Like previous studies,[17] use of pre-heart failure factors and post-heart failure factors maximised available data use and disease trajectory at an individual level, and clinical and research use over the course of heart failure (ie, not just baseline). To reduce risk of immortal time bias we limited length of the post-diagnosis window to only 2 years and we used only variables recorded before diagnosis of heart failure to predict subtypes (with associated outcomes). In a sensitivity analysis, we also compared similarity indices using only pre-diagnosis variables (appendix 1 p 13). To describe subtypes, we used all variables before and after diagnosis of heart failure. Factors with more than 30% missing data were excluded from clustering analyses (n=10). In remaining continuous factors we imputed missing data[18] by principal component analysis (timed scores regression) and for categorical factors we imputed missing data by multiple correspondence analysis. For dimensionality reduction, we used Random Forest supervised classification for 1-year mortality, ranking variable importance by prediction accuracy and Gini coefficient (figure 1; appendix 1 p 5).[19] We reduced the risk of algorithmic bias by applying and comparing four machine learning methods: K-means (partitioning and non-parametric), hierarchical (agglomerative), K-medoids (partitioning and dissimilarity matrix, and non-parametric), and mixture modelling (parametric).

Demonstrating validity (validation)

Internal validation (within dataset and across methods)

See **Online** for appendix 1

After training and validation (CPRD 15-fold; THIN 10-fold, based on size of dataset and computing capacity), one of the folds (groups into which the data were divided) was selected on the basis of similarity indices to represent the whole population in each dataset. We determined the optimal number of clusters (silhouette width and prediction strength)[20] and machine learning method (similarity indices and matrices [Rand and Jaccard means indices] and cluster stability [Fowlkes–Mallows index]; appendix 1 pp 14–15).[21]

External validation (across datasets)

Clusters were compared by accuracy between datasets (eg, CPRD clusters predicting clusters in THIN) by use of the c-statistic, and baseline continuous (means: ANOVA) and categorical (proportions: Pearson's χ^2 test) factors (appendix 1 p 17).

Prognostic validation (predictive accuracy for 1-year all-cause mortality)

See **Online** for appendix 2

We analysed prevalence and incidence for risk factors diseases and drugs in each CPRD and THIN cluster before and after incident heart failure, comparing Kaplan-Meier 1-year survival (log-rank for differences; p<0·01).

Genetic validation (polygenic risk scores [PRS] and SNPs)

Using identified cluster labels, we built a supervised learning model to predict clusters in patients with heart failure, to identify potential underlying biological mechanisms for heart failure subtypes. We assessed cross-cluster genetic differences via curated PRS[22] for 11 heart failure risk factors (atrial arrhythmias, diabetes, heavy alcohol intake, hypertension, myocardial infarction, obesity, severe anaemia, smoking, stable angina, thyroid disorders, and unstable angina; appendix 2), calculated for all UK Biobank individuals with heart failure using PLINK 2.00 alpha.

For more on **PLINK 2.00 alpha** see https://www.cog-genomics.org/plink/2.0/

We also assessed association with 12 heart failure SNPs[23] by extracting allelic dosages, inverted before

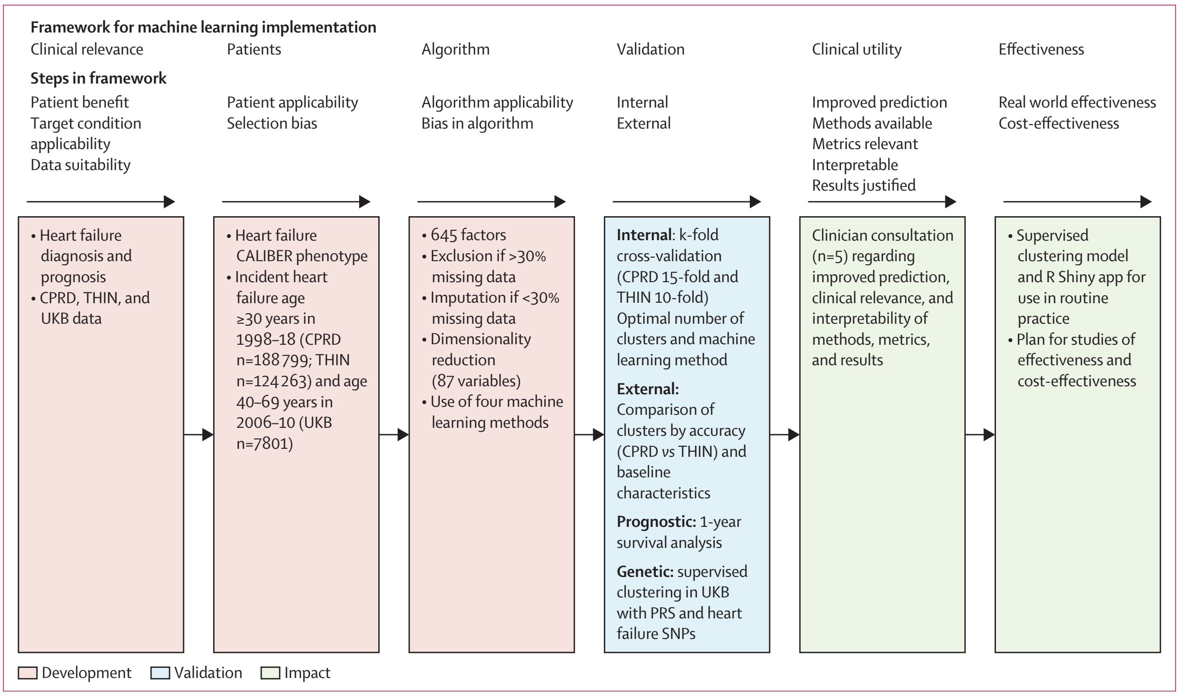

Framework for machine learning implementation
Clinical relevance
Patients
Algorithm
Validation
Clinical utility
Effectiveness
Steps in framework
Patient benefit
Target condition applicability
Data suitability
Patient applicability
Selection bias
Algorithm applicability
Bias in algorithm
Internal
External
Improved prediction
Methods available
Metrics relevant
Interpretable
Results justified
Real world effectiveness
Cost-effectiveness
• Heart failure diagnosis and prognosis
• CPRD, THIN, and UKB data
• Heart failure CALIBER phenotype
• Incident heart failure age ≥30 years in 1998–18 (CPRD n=188 799; THIN n=124 263) and age 40–69 years in 2006–10 (UKB n=7801)
• 645 factors
• Exclusion if >30% missing data
• Imputation if <30% missing data
• Dimensionality reduction (87 variables)
• Use of four machine learning methods
Internal: k-fold cross-validation (CPRD 15-fold and THIN 10-fold) Optimal number of clusters and machine learning method
External: Comparison of clusters by accuracy (CPRD vs THIN) and baseline characteristics
Prognostic: 1-year survival analysis
Genetic: supervised clustering in UKB with PRS and heart failure SNPs
Clinician consultation (n=5) regarding improved prediction, clinical relevance, and interpretability of methods, metrics, and results
• Supervised clustering model and R Shiny app for use in routine practice
• Plan for studies of effectiveness and cost-effectiveness
□ Development
□ Validation
□ Impact

Figure 1: **Study design for development, validation, and evaluation of impact of machine learning-led subtyping in incident heart failure**
CPRD=Clinical Practice Research Datalink. CALIBER=Cardiovascular disease research using linked bespoke studies and electronic health records. PRS=polygenic risk scores. SNP=single-nucleotide polymorphisms. THIN=The Health Innovation Network. UKB=UK Biobank.

analysis to reflect heart failure risk-increasing alleles. To test associations between PRS, heart failure SNPs and predicted clusters, we transformed predicted heart failure subtypes into five binary outcomes (cases [within cluster] and controls [all other participants]). By multiple logistic regression, we determined associations between heart failure related PRS, SNPs, and subtypes, visualising by heatmaps of p values.

Developing pathways to improve impact (impact)

Clinical utility

Evidence of improved outcome prediction and open methods were explored by asking five heart failure clinicians (recruited at UCL Hospitals or Barts Health NHS Trusts) about clinical relevance, justification, and interpretability of results.

Effectiveness

Based on clinician input, we developed (1) a model predicting cluster and survival using labels for identified clusters and 22 routinely available factors, and (2) a heart failure cluster app, which can be used by clinicians to identify the cluster which a particular patient falls within, and their predicted survival. Five heart failure clinicians (the same clinicians questioned about clinical utility) were asked to report whether the model and app could be effective and cost-effective. Analyses and visualisations were done in R (version 3.4.3), Microsoft Excel, Python (version 3), and R Shiny.

For the **failure cluster app** see https://pasea.shinyapps.io/hf_cluster_app/

Role of the funding source

The funder of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report.

Results

There were 264 366 and 182 570 individuals with incident heart failure in CPRD and THIN, respectively, but 41 092 were excluded from THIN due to overlapping patients. A further 75 567 (CPRD) and 17 215 (THIN) individuals did not have necessary demographic information and were excluded. Therefore, we included 188 800, 124 262, and 9573 individuals with incident heart failure from CPRD, THIN, and UK Biobank, respectively, between Jan 1, 1998, and Jan 1, 2018 (appendix 1 pp 3–4). For the algorithm we selected 87 of the 645 available factors after dimensionality reduction (appendix 1 pp 5, 14–15).

In the internal validation, the optimal number of clusters was five. Identified clusters were stable (Rand index, Jaccard means, and Fowlkes–Mallows indices >0·8 for all subtypes, using all machine learning methods except hierarchical clustering). Across datasets, we used similarity matrices to find the most representative algorithm, which was K-medoids (appendix 1 p 6). A sensitivity analysis confirmed similarity when only pre-diagnosis risk factors were included (appendix 1 p 13), suggesting low risk of immortal time bias.

In the external validation, subtypes were similar across datasets, especially when using the THIN model in

CPRD (c-statistic 0·94 [subtype 1], 0·80 [subtype 2], 0·79 [subtype 3], 0·83 [subtype 4], and 0·92 [subtype 5]); being less similar for the CPRD model in THIN (0·79 [subtype 1], 0·92 [subtype 2], 0·90 [subtype 3], 0·89 [subtype 4], and 0·92 [subtype 5]; appendix 1 p 17).

Five clusters were identified based on demography, cardiovascular disease risk factor burden, atrial fibrillation, cardiovascular disease (particularly atherosclerotic disease), medications, and laboratory factors. In CPRD, THIN, and the UK Biobank, we labelled the clusters as subtypes after studying each cluster's characteristics: (1) early onset, (2) late onset, (3) atrial fibrillation-related, (4) metabolic, and (5) cardiometabolic. Distribution of subtypes was similar across THIN and CPRD, with late onset (38 397 [30·9%] *vs* 67 213 [35·6%]) and cardiometabolic (36 906 [29·7%] *vs* 54 186 [28·7%]) being the most common subtypes and atrial fibrillation-related (11 059 [8·9%] *vs* 17 558 [9·3%]) being the least common (figure 2).

Age and gender varied by subtype (oldest participants in the late-onset subtype; youngest participants in the early-onset subtype, most females in the metabolic subtype, and fewest females in the cardiometabolic subtype). In THIN the cardiometabolic subtype had highest prevalence of cardiovascular risk factors and diseases—eg, hypertension 72·9% (n=26 904), obesity 4·7% (n=1735), diabetes 34·5% (n=12 735), and atherosclerotic cardiovascular disease 59·2% (n=21 848; table; figure 2). Age, blood investigations, BMI, and blood pressure did not discriminate well between subtypes or mortality by subtype (table; appendix 1 pp 7–8).

In the prognostic validity analysis in CPRD using the THIN model, 1-year mortality was 0·20 (95% CI 0·14–0·25; subtype 1), 0·46 (0·43–0·49; subtype 2), 0·61 (0·57–0·64; subtype 3), 0·11 (0·07–0·16; subtype 4), and 0·37 (0·32–0·41; subtype 5), with c-statistics of 0·68 (95% CI 0·65–0·71; subtype 1), 0·62 (0·59–0·65; subtype 2), 0·57 (0·55–0·59; subtype 3), 0·71 (0·70–0·73; subtype 4), and 0·68 (0·65–0·70; subtype 5; appendix 1 p 11). There were differences in mortality for clusters 1 (early onset) and 5 (cardiometabolic), but not other clusters, between THIN and CPRD (figure 3; appendix 1 p 18). Atrial fibrillation occurred after heart failure diagnosis in the atrial fibrillation-related subtype (proportion=0·96 [95% CI 0·94–0·99]) and was more likely to be before heart failure diagnosis for other subtypes. Hypertension (0·61 [0·59–0·63]), myocardial infarction (0·27 [0·24–0·29]), stroke (0·17 [0·13–0·22]), and peripheral vascular disease (0·16 [0·12–0·20]) occurred predominantly before heart failure diagnosis in the cardiometabolic subtype, and after heart failure diagnosis in atrial fibrillation-related subtype and early onset subtype (figure 4; appendix 1 p 9). After heart failure diagnosis, the proportion of use of the following drugs was highest in patients with atrial fibrillation-related and early-onset subtypes: β blockers (0·38 [95% CI 0·35–0·40] and 0·19 [0·16–0·23]), angiotensin converting enzyme inhibitor (0·44 [0·41–0·46] and 0·21 [0·18–0·24]), and aldosterone antagonists (0·42 [0·40–0·44] and 0·11 [0·07–0·14]; appendix 1 p 10). All-cause hospitalisation after heart failure diagnosis (overall rate 0·013 [95% CI 0·009–0·017]) also varied by heart failure subtype.

A

Cluster	Cluster label	Cluster-specific characteristics
1	Early onset	Young, low rate of risk factors
2	Late onset	Old, female, low rate of prescribed medications and cardiovascular disease
3	Atrial fibrillation-related	Atrial fibrillation, valve disease
4	Metabolic	Overweight, medium rate of risk factors, low rate of cardiovascular disease
5	Cardiometabolic	Overweight, high rate of prescribed medications, risk factors, and cardiovascular disease

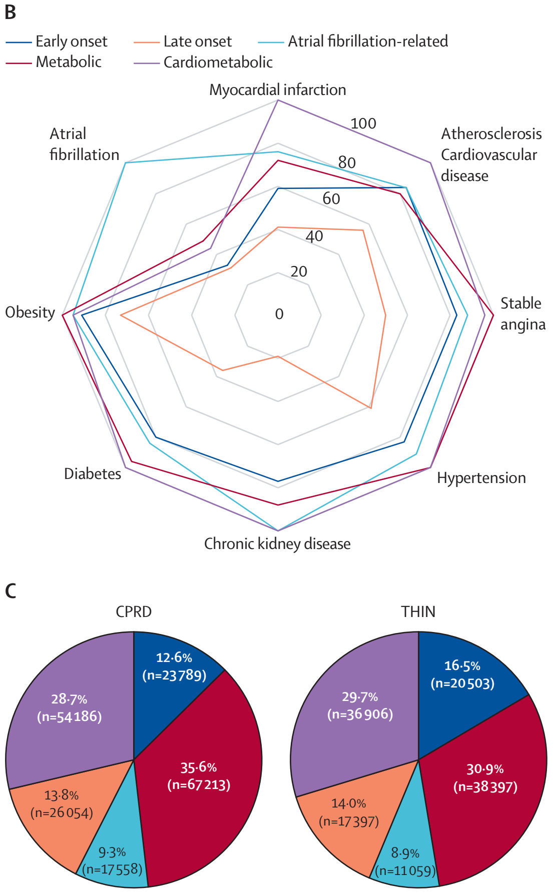


***Figure 2:* Externally validated clusters in incident heart failure in two UK primary care populations (THIN and CPRD; n=313 062)**
(A) Cluster characteristics. (B) Relative prevalence of common risk factors across clusters. For each risk factor, the highest prevalence was designated as 100% and the prevalence in each of the other clusters was relative to that prevalence (0–100). (C) Proportion of each cluster in overall population. CPRD=Clinical Practice Research Datalink. THIN=The Health Innovation Network.

In the genetic validity analysis 7801 of 9573 patients with heart failure in the UK Biobank had necessary

	Subtype 1: early onset (n=44 292)		Subtype 2: late onset (n=105 610)		Subtype 3: atrial fibrillation-related (n=28 617)		Subtype 4: metabolic (n=43 451)		Subtype 5: cardiometabolic (n=91 092)	
	THIN n=20 503 (16·5%)	CPRD n=23 789 (12·6%)	THIN n=38 397 (30·9%)	CPRD n=67 213 (35·6%)	THIN n=11 059 (8·9%)	CPRD n=17 558 (9·3%)	THIN n=17 397 (14·0%)	CPRD n=26 054 (13·8%)	THIN n=36 906 (29·7%)	CPRD n=54 186 (28·7%)
Demographics										
Age, years	69·1 (13·3)	70·4 (12·4)	79·9 (11·0)	79·6 (12·3)	73·6 (10·7)	77·1 (10·5)	74·0 (12·0)	77·0 (11·3)	75·3 (10·7)	78·4 (10·5)
Sex										
Female	9062 (44·2%)	10 515 (44·2%)	22 078 (57·5%)	36 766 (54·7%)	4943 (44·7%)	9095 (51·8%)	7707 (44·3%)	12 714 (48·8%)	16 054 (43·5%)	25 955 (47·9%)
Male	11 441 (55·8%)	13 274 (55·8%)	16 319 (42·5%)	30 447 (45·3%)	6116 (55·3%)	8463 (48·2%)	9690 (55·7%)	13 340 (51·2%)	20 852 (56·5%)	28 231 (52·1%)
Ethnicity										
White	10 795 (52·7%)	10 940 (46%)	21 339 (55·6%)	32 726 (48·7%)	5500 (49·7%)	8264 (37·7%)	9973 (57·3%)	1409 (54·1%)	21 158 (63·1%)	27 263 (55·3%)
Black	165 (0·8%)	768 (3·2%)	322 (0·8%)	2299 (3·4%)	13 (0·0%)	85 (0·4%)	135 (0·8%)	99 (3·8%)	302 (0·9%)	1791 (3·6%)
Asian	135 (0·7%)	723 (3%)	226 (0·6%)	1936 (2·9%)	11 (0·0%)	65 (0·3%)	110 (0·6%)	81 (3·1%)	250 (0·7%)	1647 (3·3%)
Unknown or other	9390 (45·8%)	12 085 (50·8%)	16 507 (43·0%)	30 279 (45·1%)	6624 (59·9%)	10 810 (61·6%)	7185 (41·3%)	10162 (39·0%)	13 028 (35·3%)	20 482 (37·8%)
Risk factors										
Obesity, BMI ≥40 kg/m²	738 (3·6%)	809 (3·4%)	576 (1·5%)	1344 (2·0%)	575 (5·2%)	615 (3·5%)	870 (5%)	1303 (5·0%)	1735 (4·7%)	2601 (4·8%)
BMI, kg/m²	27·7 (5·9)	27·9 (5·7)	25·5 (5·0)	26·2 (5·2)	28·5 (6·3)	27·7 (5·9)	28·7 (6·3)	28·5 (6·4)	28·5 (6·2)	28·1 (6·2)
Hypertension	11 912 (58·1%)	20 411 (85·8%)	16 703 (43·5%)	41 538 (61·8%)	7255 (65·6%)	14 538 (82·8%)	14 161 (81·4%)	23 527 (90·3%)	26 904 (72·9%)	48 442 (89·4%)
Blood pressure	133·5/77·1 (22·9/12·9)	136·9/78·1 (13·0/13·0)	138·4/77·3 (22·6/11·9)	139·8/78·3 (11·4/11·4)	137·7/78·2 (22·5/12·2)	139·6/77·3 (12·2/12·2)	131·6/74·9 (19·9/11·8)	132·9/74·6 (11·7/11·7)	132·1/74·3 (21·1/11·8)	133·8/73·7 (11·6/11·6)
Total cholesterol, mmol/L	4·8 (1·3)	··	5·1 (0·8)	··	4·8 (1·1)	··	4·4 (1·1)	··	4·4 (1·2)	··
Diabetes	5392 (26·3%)	8492 (35·7%)	3878 (10·1%)	11 225 (16·7%)	3063 (27·7%)	5355 (30·5%)	5410 (31·1%)	9874 (37·9%)	12 732 (34·5%)	21 729 (40·1%)
Chronic kidney disease, stage 3	6315 (30·8%)	7922 (33·3%)	2880 (7·5%)	3562 (5·3%)	3981 (36%)	4723 (26·9%)	6263 (36·0%)	9223 (35·4%)	14 541 (39·4%)	18 532 (34·2%)
Renal failure	156 (0·8%)	262 (1·1%)	138 (0·4%)	181 (0·3%)	81 (0·7%)	193 (1·1%)	75 (0·4%)	255 (1·0%)	406 (1·1%)	444 (0·8%)
Creatinine, mmol/L	104·4 (48·7)	106·6 (49·8)	115·3 (57·3)	107·0 (60·8)	107·5 (45·8)	109·5 (57·8)	103·3 (50·1)	104·3 (61·2)	112·9 (61·5)	112·5 (65·2)
Haemoglobin, g/dL	13·3 (1·8)	13·2 (1·9)	12·6 (1·6)	12·9 (1·7)	13·1 (1·8)	12·9 (1·8)	13·1 (1·8)	12·9 (1·9)	12·9 (1·8)	12·8 (1·9)
Baseline cardiovascular disease										
Atrial fibrillation	6540 (31·9%)	15 463 (65%)	10 828 (28·2%)	30 246 (45%)	10 318 (93·3%)	17 312 (98·6%)	8020 (46·1%)	17 091 (65·6%)	14 799 (40·1%)	31 590 (58·3%)
Myocardial infarction	6581 (32·1%)	11 086 (46·6%)	5798 (15·1%)	15 526 (23·1%)	3119 (28·2%)	7058 (40·2%)	4767 (27·4%)	9822 (37·7%)	13 950 (37·8%)	24 004 (44·3%)
Unstable angina	1353 (6·6%)	5138 (21·6%)	1229 (3·2%)	4839 (7·2%)	829 (7·5%)	3635 (20·7%)	1270 (7·3%)	4846 (18·6%)	4281 (11·6%)	11 108 (20·5%)
Stable angina	12 466 (60·8%)	17 414 (73·2%)	14 975 (39%)	29 238 (43·5%)	7011 (63·4%)	12 220 (69·6%)	12 787 (73·5%)	19 280 (74·0%)	26 904 (72·9%)	39 718 (73·3%)
Stroke	3239 (15·8%)	4686 (19·7%)	5529 (14·4%)	10 351 (15·4%)	2101 (19%)	3582 (20·4%)	2923 (16·8%)	4377 (16·8%)	7418 (20·1%)	11 867 (21·9%)
Peripheral vascular disease	2809 (13·7%)	5567 (23·4%)	4531 (11·8%)	10 418 (15·5%)	1780 (16·1%)	4811 (27·4%)	2575 (14·8%)	6409 (24·6%)	7418 (20·1%)	14 901 (27·5%)
Atherosclerotic cardiovascular disease	10 046 (49·0%)	15 415 (64·8%)	13 247 (34·5%)	29 103 (43·3%)	5430 (49·1%)	11 202 (63·8%)	8107 (46·6%)	15 059 (57·8%)	21 848 (59·2%)	35 925 (66·3%)
Drugs at baseline										
Antiplatelet	15 357 (74·9%)	19 103 (80·3%)	20 811 (54·2%)	30 111 (44·8%)	8847 (80%)	13 432 (76·5%)	13 535 (77·8%)	20 270 (77·8%)	32 736 (88·7%)	46 600 (86·0%)
Statin	15 336 (74·8%)	18 888 (79·4%)	4876 (12·7%)	5511 (8·2%)	7310 (66·1%)	9394 (53·5%)	13 309 (76·5%)	19 462 (74·7%)	32 736 (88·7%)	47 955 (88·5%)
Angiotensin converting enzyme inhibitor	17 284 (84·3%)	21 244 (89·3%)	20 197 (52·6%)	33 875 (50·4%)	9566 (86·5%)	14 714 (83·8%)	15 118 (86·9%)	23 110 (88·7%)	33 843 (91·7%)	46 221 (85·3%)
β blocker	15 090 (73·6%)	16 985 (71·4%)	11 097 (28·9%)	19 693 (29·3%)	8150 (73·7%)	10 113 (57·6%)	15 153 (87·1%)	20 661 (79·3%)	29 968 (81·2%)	37 984 (70·1%)
Diuretics	18 063 (88·1%)	··	33 981 (88·5%)	··	10 617 (96·0%)	15 978 (91·0%)	15 675 (90·1%)	22 771 (87·4%)	34 544 (93·6%)	47 359 (87·4%)
Aldosterone antagonist	7853 (38·3%)	··	6681 (17·4%)	··	5253 (47·5%)	··	7081 (40·7%)	··	14 246 (38·6%)	··
Warfarin	6130 (29·9%)	12 465 (52·4%)	6911 (18·0%)	18 416 (27·4%)	6735 (60·9%)	8867 (50·5%)	6872 (39·5%)	12 063 (46·3%)	13 212 (35·8%)	19 344 (35·7%)

Data are n (%) or mean (SD). CPRD= Clinical Practice Research Datalink. THIN=The Health Innovation Network.

Table: **Baseline characteristics by subtype of incident heart failure in two UK primary care populations (n=313 062)**

genetic data for analyses of PRS and SNPs. The number of patients in the metabolic cluster was low (n=49), but other clusters were well-represented (1586 in late onset, 1981 in atrial fibrillation, 1553 in early onset, and 2633 in cardiometabolic).

The associations between heart failure subtypes and both PRS and heart failure-related SNPs are shown in figure 5 (see also appendix 1 pp 19–20). PRS for atrial arrhythmias, diabetes, hypertension, myocardial infarction, obesity, stable angina, and unstable angina were all associated with one or more heart failure subtypes after correction for multiple testing ($p<0{\cdot}0009$). The late onset and cardiometabolic subtypes broadly associated with similar PRS. No PRS was associated with the metabolic subtype, potentially due to small numbers in this group (n=49). Eight SNPs were nominally associated ($p=0{\cdot}049$) with predicted heart failure subtypes. Four of these associated SNPs were confined to the atrial fibrillation-related subtype: rs11745324, rs17042102, rs4746140, and rs4766578 (figure 5B), corresponding to the PITX2 and FAM241A, SYNPO2L, AGAP5, and ATXN2 genes, respectively. Associations between rs17042102 and the atrial fibrillation-related subtype persisted even after correcting for multiple testing ($p=0{\cdot}05/60=8{\cdot}3\times10^{-4}$), suggesting importance of chromosome 4 in atrial fibrillation-related heart failure.

To determine impact, the clinical utility was assessed. Sample clinicians (n=5) reported that the included factors and the identified clusters had clinical relevance as per our 2021 framework, in terms of a research question related to patient benefit, target condition applicability, and data suitable for the clinical question. Differences between clusters by baseline characteristics and survival were distinguishable and interpretable. The framework or methods proposed were felt to be transparent and generalisable. To assess the effectiveness we developed an open access heart failure cluster app which can be used by clinicians to identify the cluster which a particular patient falls within, and their predicted survival. The interviewed clinicians felt that this app was a feasible use of the identified clusters to identify which cluster a patient belonged to during consultations in routine care and that it could enable testing of effectiveness and cost-effectiveness in appropriately designed, prospective studies (as elaborated in the discussion section).

Based on clinician and researcher input, effectiveness could be tested in six ways: (1) prospective validation of the clusters in routine care; (2) comparison of treatment and care pathways with app for cluster identification versus usual care (possibly in a trial); (3) predictive accuracy for survival compared with existing risk prediction tools; (4) patient-reported outcomes; (5) patient satisfaction; and (6) clinician satisfaction and ease of use in clinical practice. Cost-effectiveness could be estimated by modelling the impact on care and outcomes based on the above analysis of effectiveness, the time required to estimate and communicate subtype in clinical settings, and the potential effect on health-care use and outcomes.

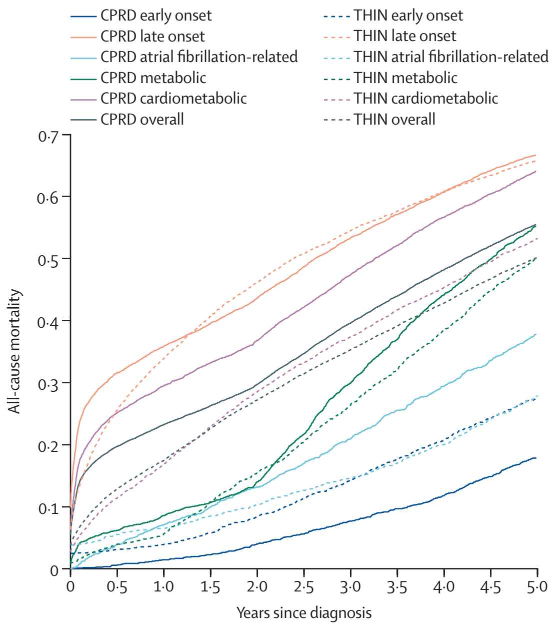


Figure 3: **Prognostic validation by all-cause mortality using clusters in incident heart failure in two UK primary care populations (THIN and CPRD; n=313 062)**

Pairwise comparisons using log-rank test (appendix 1 p 18) revealed statistically significant differences between all survival curves ($p<0{\cdot}0001$) except for THIN early-onset and CPRD early-onset ($p=0{\cdot}036$), THIN atrial fibrillation-related and CPRD atrial fibrillation-related ($p=0{\cdot}82$), and THIN metabolic and CPRD metabolic ($p=0{\cdot}40$). See appendix 1 (p 21) for patients at risk. CPRD=Clinical Practice Research Datalink. THIN=The Health Innovation Network.

Discussion

To our knowledge, this is the first study to define and validate data-driven heart failure disease clusters across multiple machine learning methods, nationally representative datasets, and validation methods, with three distinct advances. First, we identified five incident heart failure subtypes: (1) early onset, (2) late onset, (3) atrial fibrillation-related, (4) metabolic, and (5) cardiometabolic. Second, we confirmed internal, external, prognostic, and genetic validity. Third, we developed a means of using identified subtypes in routine practice and suggest ways of evaluating effectiveness.

For the **failure cluster app** see https://pasea.shinyapps.io/hf_cluster_app/

Our five subtypes are compatible with findings from two major clustering studies, previously describing four heart failure subtypes,[9,10] differing by age and prevalence of myocardial infarction, hypertension, and diabetes. Our data and identified subtypes are likely to be more representative than machine learning studies in heart failure to date, which have seldom used EHR data, nationwide data, or population-based data.[10] For cause, guidelines, and research, studies have predominantly focused on ischaemic versus non-ischaemic heart failure, and heart failure defined by cut-offs based on measures of

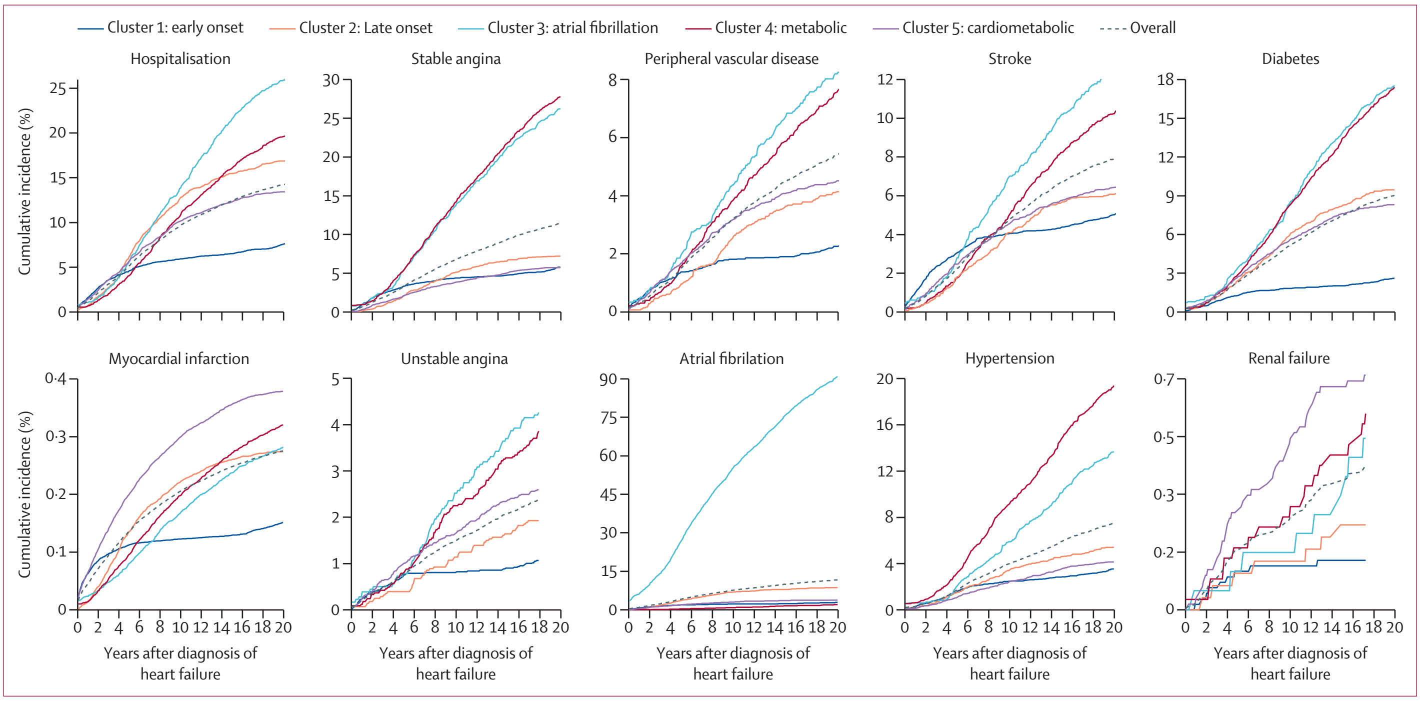
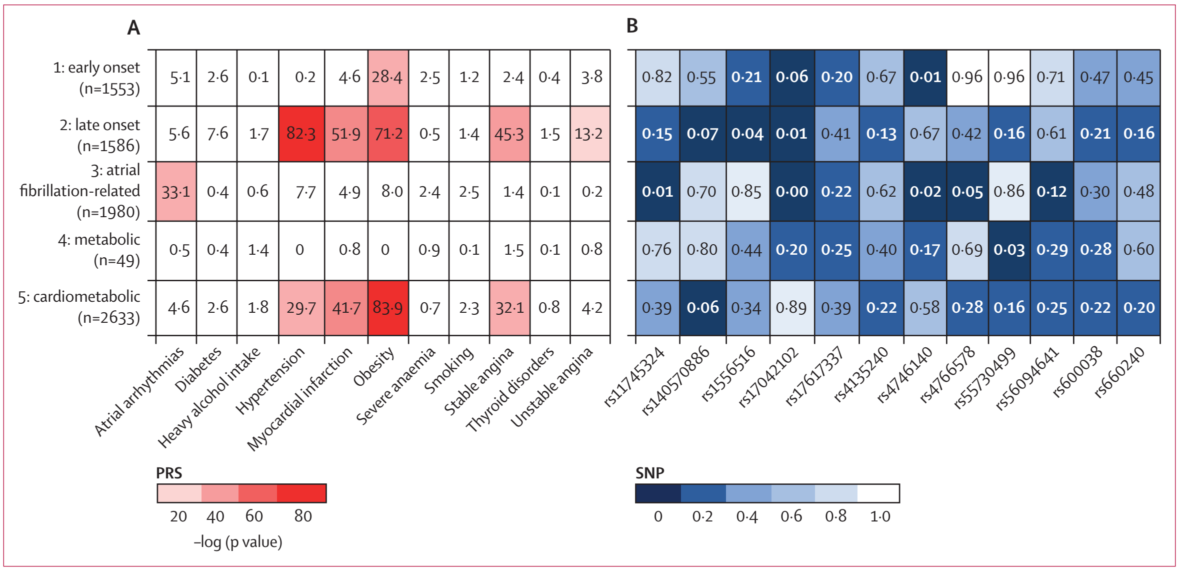


Figure 4: **Risk of non-fatal cardiovascular diseases and all-cause hospitalisation in five heart failure subtypes after diagnosis**



A

	Atrial arrhythmias	Diabetes	Heavy alcohol intake	Hypertension	Myocardial infarction	Obesity	Severe anaemia	Smoking	Stable angina	Thyroid disorders	Unstable angina
1: early onset (n=1553)	5·1	2·6	0·1	0·2	4·6	28·4	2·5	1·2	2·4	0·4	3·8
2: late onset (n=1586)	5·6	7·6	1·7	82·3	51·9	71·2	0·5	1·4	45·3	1·5	13·2
3: atrial fibrillation-related (n=1980)	33·1	0·4	0·6	7·7	4·9	8·0	2·4	2·5	1·4	0·1	0·2
4: metabolic (n=49)	0·5	0·4	1·4	0	0·8	0	0·9	0·1	1·5	0·1	0·8
5: cardiometabolic (n=2633)	4·6	2·6	1·8	29·7	41·7	83·9	0·7	2·3	32·1	0·8	4·2

B

	rs11745324	rs140570886	rs1556516	rs17042102	rs17617337	rs4135240	rs4746140	rs4766578	rs55730499	rs56094641	rs600038	rs660240
1: early onset (n=1553)	0·82	0·55	0·21	0·06	0·20	0·67	0·01	0·96	0·96	0·71	0·47	0·45
2: late onset (n=1586)	0·15	0·07	0·04	0·01	0·41	0·13	0·67	0·42	0·16	0·61	0·21	0·16
3: atrial fibrillation-related (n=1980)	0·01	0·70	0·85	0·00	0·22	0·62	0·02	0·05	0·86	0·12	0·30	0·48
4: metabolic (n=49)	0·76	0·80	0·44	0·20	0·25	0·40	0·17	0·69	0·03	0·29	0·28	0·60
5: cardiometabolic (n=2633)	0·39	0·06	0·34	0·89	0·39	0·22	0·58	0·28	0·16	0·25	0·22	0·20

Figure 5: **Heart failure subtypes and genotype associations (n=7801)**

Numbers represent p values from the logistic regression of polygenic risk score or single nucleotide polymorphism versus cluster output. (A) association between heart failure subtypes and PRS for 11 related traits. (B) direct association between heart failure subtypes and 12 related SNPs. PRS=polygenic risk score. SNP=single-nucleotide polymorphism.

LVEF.[24] The cardiometabolic subtype captures ischaemic cause.[1,2,5] The atrial fibrillation-related subtype is consistent with doubling of risk of incident atrial fibrillation observed in prevalent heart failure, compared to no heart failure (hazard ratio 2·18 [95% CI 1·26–3·76]).[25] The high proportion of prevalent atrial fibrillation in other subtypes is consistent with atrial fibrillation causing heart failure. Further study of atrial, ventricular, and atrioventricular cardiomyopathies will inform temporal associations between atrial fibrillation and different heart failure subtypes. Patients with the metabolic subtype were younger and with higher prevalence of atrial fibrillation and obesity,[26] but lower prevalence of atherosclerotic cardiovascular disease, than the cardiometabolic subtype, though this was not entirely distinguishable. Age was predictive of overall heart failure and particular subtypes in prior studies. Therefore, early-onset and late-onset subtypes are plausible,[9] warranting further investigation across countries and factors—eg, echocardiography.

Methodologically, we offer advances in external validation of machine learning in subtype classification and in risk prediction in heart failure, which has been rare (only four of 27 and two of 31 studies, respectively) and in small samples (sample sizes of 44 to 3203 for studies of heart failure subtypes).[10] Our robust, structured framework of internal, external, prognostic, and genetic validation could extend acceptability and generalisability of machine learning to clinical practice and is transferable to other diseases. Our subtypes showed good accuracy within and across datasets, and good predictive accuracy for early-onset, metabolic, and cardiometabolic subtypes; although, less accurate for atrial fibrillation-related and late-onset subtypes. The exact reason for these differences in predictive accuracy between subtypes cannot be ascertained from this study but might be due to more nuanced changes in risk factors and trajectories over time than we captured in our data or our models. Further research is necessary to understand these differences. The c-statistic for LVEF, the most commonly used feature to define heart failure subtype, was only 0·52 in a large Swedish national registry study using machine learning.[9] Even after inclusion of more clinical factors (eg, echocardiography and NT-proBNP) or focus on particular subgroups or clinical scenarios,[27] improved risk prediction for mortality and other outcomes remains challenging. Our findings of PRS and SNPs associated with the atrial fibrillation-related subtype are novel, signalling potential use of assessment for biological validity of cluster analyses and their linkage to EHRs.[23] The mild associations observed with related PRS for the early-onset heart failure subtype (with the exception of strong association with obesity and atrial arrhythmias), compared with late-onset and cardiometabolic subtypes, are of interest. Studies of machine learning in heart failure should focus on further validation in representative datasets from other countries, disease definition, and use of high-dimensionality proteomics and imaging data.

Recent guidelines describe the need for systematic approaches to design, evaluation, and implementation of machine learning in health care.[10,28,29] We address issues at the development and validation stages to use of machine learning for subtype classification and risk prediction in heart failure, and therefore we did not use checklists designed specifically for prediction tools, but we completed the recently developed TRIPOD-clustering checklist[29] post-hoc (appendix 1 p 22). Our approach to clinical utility (relevance, justification, and interpretability) illustrates how specialist and patient views can be assessed and incorporated in the evaluation of machine learning in health care, where there is currently little guidance to aid implementation in health care. Although we interviewed a small number of clinicians, the approach could be used at national and international level. To assess effectiveness, we offer a prototype for application of our identified subtypes in routine care which needs further investigation at the implementation stage, especially analyses of effectiveness and cost-effectiveness, which are currently lacking.

Our study is one of the largest EHR analyses to date to use machine learning in subtype classification and risk prediction of heart failure, and was the first to investigate multiple machine learning methods, multiple datasets, and multiple validation methods. By using routine EHRs, our derivation and validation cohorts are representative of real-world patients, increasing the likelihood of clinical utility and applicability. We incorporated factors before and after heart failure diagnosis, enabling insights into trajectory as well as cause. However, there are several limitations. First, we are using EHR phenotypes of heart failure, which do not have complete biochemical (eg, NT-pro-BNP) and imaging (eg, LVEF) profiles, and therefore some previous classifications are not possible (eg, heart failure with preserved ejection fraction). However, our phenotypes have been validated and used in prior large-scale studies.[12,16] Second, although we used 645 factors, risk factor phenotypes are limited by timing and accuracy of the clinician recording in the EHRs, which might affect analyses of factors before and after heart failure. Third, although we use two large, nationally representative primary care datasets, both are UK-based and might not be representative of heart failure in other countries or settings. Fourth, we performed only supervised analyses of PRS of 11 traits related to heart failure and 12 SNPs previously associated with heart failure (and numbers in the metabolic cluster were small), necessitating further genetic analysis in larger cohorts. Fifth, there is risk of immortal time bias in our models because we included variables up to 2 years post-heart failure diagnosis. Finally, further, larger scale clinician and patient input is required in the implementation stage of the subtypes.

Across three large, population-scale datasets, four machine learning methods, 645 factors, and four validation methods, we identify five heart failure subtypes with good discriminatory accuracy within and across datasets, and good predictive accuracy for 1-year mortality. These subtypes might have implications for research in terms of use of EHR and machine learning to identify heart failure subtypes in future clinical trials and observational studies, as well as clinical practice in terms of management and prognosis.

Contributors

AB and HH conceived the research question. AB, SC, AD, GF, SD, and HH designed the study and analysis plan. SC, AD, LP, JHT, and GF conducted different parts of experiments. AB, AD, SC, and HH drafted the initial and final versions of manuscript. All authors critically

reviewed early and final versions of the manuscript. All authors had access to all data, and SC and AD have verified the data. All authors had final responsibility for the decision to submit for publication.

Declaration of interests

BT is an employee of Bayer and was previously an employee of Servier. TD is an employee of Bayer. AB is supported by research funding from the National Institute for Health Research (NIHR), British Medical Association, AstraZeneca, and UK Research and Innovation. HH is supported by Health Data Research UK (grant number LOND1), which is funded by the UK Medical Research Council, Engineering and Physical Sciences Research Council, Economic and Social Research Council, Department of Health and Social Care (England), Chief Scientist Office of the Scottish Government Health and Social Care Directorates, Health and Social Care Research and Development Division (Welsh Government), Public Health Agency (Northern Ireland), British Heart Foundation, and Wellcome Trust; and is a NIHR Senior Investigator. AB, SD, FWA, and HH are funded by the NIHR University College London Hospitals Biomedical Research Centre.

Data sharing

All relevant data are included in the Article or appendices. Data used in this study cannot be made available to other researchers by the study team because access to data from CPRD, THIN, or UK Biobank is via open application and ethical review process directly with organisations responsible for these datasets.

Acknowledgments

All authors are supported by the BigData@Heart Consortium, funded by the European Union Innovative Medicines Initiative-2 joint undertaking under grant agreement number 116074. This joint undertaking receives support from the EU's Horizon 2020 research and innovation programme and the European Federation of Pharmaceutical Industries and Associations.

References

1 Ponikowski P, Voors AA, Anker SD, et al. 2016 ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure: The Task Force for the diagnosis and treatment of acute and chronic heart failure of the European Society of Cardiology (ESC)developed with the special contribution of the Heart Failure Association (HFA) of the ESC. *Eur Heart J* 2016; **37:** 2129–200.

2 Mordi IR, Pearson ER, Palmer CNA, Doney ASF, Lang CC. Differential association of genetic risk of coronary artery disease with development of heart failure with reduced versus preserved ejection fraction. *Circulation* 2019; **139:** 986–88.

3 Solomon SD, Pfeffer MA. The future of clinical trials in cardiovascular medicine. *Circulation* 2016; **133:** 2662–70.

4 Seidelmann SB, Feofanova E, Yu B, et al. Genetic variants in *SGLT1*, glucose tolerance, and cardiometabolic risk. *J Am Coll Cardiol* 2018; **72:** 1763–73.

5 Yancy CW, Jessup M, Bozkurt B, et al. 2017 ACC/AHA/HFSA focused update of the 2013 ACCF/AHA guideline for the management of heart failure: a report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines and the Heart Failure Society of America. *J Am Coll Cardiol* 2017; **70:** 776–803.

6 Chawla LS, Herzog CA, Costanzo MR, et al. Proposal for a functional classification system of heart failure in patients with end-stage renal disease: proceedings of the acute dialysis quality initiative (ADQI) XI workgroup. *J Am Coll Cardiol* 2014; **63:** 1246–52.

7 Arnett DK, Blumenthal RS, Albert MA, et al. ACC/AHA guideline on the primary prevention of cardiovascular disease. *Circulation* 2019; **2019:** CIR0000000000000678.

8 Banerjee A, Pasea L, Chung SC, et al. A population-based study of 92 clinically recognized risk factors for heart failure: co-occurrence, prognosis and preventive potential. *Eur J Heart Fail* 2022; **24:** 466–80.

9 Ahmad T, Lund LH, Rao P, et al. Machine learning methods improve prognostication, identify clinically distinct phenotypes, and detect heterogeneity in response to therapy in a large cohort of heart failure patients. *J Am Heart Assoc* 2018; **7:** e008081.

10 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.

11 Banerjee A, Benedetto V, Gichuru P, et al. Adherence and persistence to direct oral anticoagulants in atrial fibrillation: a population-based study. *Heart* 2020; **106:** 119–26.

12 Denaxas S, Gonzalez-Izquierdo A, Direk K, et al. UK phenomics platform for developing and validating electronic health record phenotypes: CALIBER. *J Am Med Inform Assoc* 2019; **26:** 1545–59.

13 Biobank UK. Detailed genetic data on half a million people. https://www.ukbiobank.ac.uk/enable-your-research/about-our-data/genetic-data (accessed Sept 26, 2022).

14 Cai B, Xu W, Bortnichak E, Watson DJ. An algorithm to identify medical practices common to both the General Practice Research Database and The Health Improvement Network database. *Pharmacoepidemiol Drug Saf* 2012; **21:** 770–74.

15 Carbonari DM, Saine ME, Newcomb CW, et al. Use of demographic and pharmacy data to identify patients included within both the Clinical Practice Research Datalink (CPRD) and The Health Improvement Network (THIN). *Pharmacoepidemiol Drug Saf* 2015; **24:** 999–1003.

16 Koudstaal S, Pujades-Rodriguez M, Denaxas S, et al. Prognostic burden of heart failure recorded in primary care, acute hospital admissions, or both: a population-based linked electronic health record cohort study in 2.1 million people. *Eur J Heart Fail* 2017; **19:** 1119–27.

17 Steele AJ, Denaxas SC, Shah AD, Hemingway H, Luscombe NM. Machine learning models in electronic health records can outperform conventional survival models for predicting patient mortality in coronary artery disease. *PLoS One* 2018; **13:** e0202344.

18 Josse J, Husson F. missMDA: a package for handling missing values in multivariate data analysis. *J Stat Softw* 2016; **70:** 1–31.

19 Saraswat M, Arya KV. Feature selection and classification of leukocytes using random forest. *Med Biol Eng Comput* 2014; **52:** 1041–52.

20 Fujita A, Takahashi DY, Patriota AG. A non-parametric method to estimate the number of clusters. *Comput Stat Data Anal* 2014; **73:** 27–39.

21 Gates AJ, Wood IB, Hetrick WP, Ahn YY. Element-centric clustering comparison unifies overlaps and hierarchy. *Sci Rep* 2019; **9:** 8574.

22 Lambert SA, Gil L, Jupp S, et al. The Polygenic Score Catalog as an open database for reproducibility and systematic evaluation. *Nat Genet* 2021; **53:** 420–25.

23 Shah S, Henry A, Roselli C, et al. Genome-wide association and Mendelian randomisation analysis provide insights into the pathogenesis of heart failure. *Nat Commun* 2020; **11:** 163.

24 Bhambhani V, Kizer JR, Lima JAC, et al. Predictors and outcomes of heart failure with mid-range ejection fraction. *Eur J Heart Fail* 2018; **20:** 651–59.

25 Santhanakrishnan R, Wang N, Larson MG, et al. Atrial fibrillation begets heart failure and vice versa: temporal associations and differences in preserved versus reduced ejection fraction. *Circulation* 2016; **133:** 484–92.

26 Savji N, Meijers WC, Bartz TM, et al. The association of obesity and cardiometabolic traits with incident HFpEF and HFrEF. *JACC Heart Fail* 2018; **6:** 701–09.

27 Yoon J, Zame WR, Banerjee A et al. Personalized survival predictions via Trees of Predictors: an application to cardiac transplantation. *PLoS One* 2018; **13:** e0194985.

28 Cruz Rivera S, Liu X, Chan A-W, et al. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. *Nat Med* 2020; **26:** 1351–63.

29 Debray TPA, Collins GS, Riley RD, et al. Transparent reporting of multivariable prediction models developed or validated using clustered data: TRIPOD-Cluster checklist. *BMJ* 2023; **380:** e071018.