AI-based derivation of atrial fibrillation phenotypes in the general and critical care populations

Ryan A. A. Bellfield,[a,b] Ivan Olier,[a,b] Robyn Lotto,[b,c] Ian Jones,[b,c] Ellen A. Dawson,[b,d] Guowei Li,[e] Anil M. Tuladhar,[f] Gregory Y. H. Lip,[b,g] and Sandra Ortega-Martorell[a,b,*]

[a]Data Science Research Centre, Liverpool John Moores University, Liverpool L3 3AF, UK
[b]Liverpool Centre for Cardiovascular Science at University of Liverpool, Liverpool John Moores University and Liverpool Heart & Chest Hospital, Liverpool, UK
[c]School of Nursing and Advanced Practice, Liverpool John Moores University, Liverpool L2 2ER, UK
[d]Research Institute for Sport and Exercise Science, Liverpool John Moores University, Liverpool L3 3AF, UK
[e]Center for Clinical Epidemiology and Methodology (CCEM), Guangdong Second Provincial General Hospital, Guangzhou 510317, China
[f]Department of Neurology, Radboud University Medical Centre, Donders Institute for Brain, Cognition and Behavior, Nijmegen, the Netherlands
[g]Danish Center for Health Services Research, Department of Clinical Medicine, Aalborg University, Aalborg, Denmark

eBioMedicine
2024;107: 105280
Published Online xxx
https://doi.org/10.1016/j.ebiom.2024.105280

Summary

Background Atrial fibrillation (AF) is the most common heart arrhythmia worldwide and is linked to a higher risk of mortality and morbidity. To predict AF and AF-related complications, clinical risk scores are commonly employed, but their predictive accuracy is generally limited, given the inherent complexity and heterogeneity of patients with AF. By classifying different presentations of AF into coherent and manageable clinical phenotypes, the development of tailored prevention and treatment strategies can be facilitated. In this study, we propose an artificial intelligence (AI)-based methodology to derive meaningful clinical phenotypes of AF in the general and critical care populations.

Methods Our approach employs generative topographic mapping, a probabilistic machine learning method, to identify micro-clusters of patients with similar characteristics. It then identifies macro-cluster regions (clinical phenotypes) in the latent space using Ward's minimum variance method. We applied it to two large cohort databases (UK-Biobank and MIMIC-IV) representing general and critical care populations.

Findings The proposed methodology showed its ability to derive meaningful clinical phenotypes of AF. Because of its probabilistic foundations, it can enhance the robustness of patient stratification. It also produced interpretable visualisation of complex high-dimensional data, enhancing understanding of the derived phenotypes and their key characteristics. Using our methodology, we identified and characterised clinical phenotypes of AF across diverse patient populations.

Interpretation Our methodology is robust to noise, can uncover hidden patterns and subgroups, and can elucidate more specific patient profiles, contributing to more robust patient stratification, which could facilitate the tailoring of prevention and treatment programs specific to each phenotype. It can also be applied to other datasets to derive clinically meaningful phenotypes of other conditions.

Funding This study was funded by the DECIPHER project (LJMU QR-PSF) and the EU project TARGET (10113624).

Keywords: Clinical phenotypes; Atrial fibrillation; Generative topographic mapping; UK-Biobank; MIMIC-IV; Probabilistic modelling; Machine learning; Clustering; Stratification

Introduction

Atrial fibrillation (AF) is the commonest heart arrhythmia worldwide,[1] affecting 2% of the European population (15 M patients). AF risk increases with age, with ~18 M patients with AF estimated by 2060.[2] AF is linked to a higher risk of mortality and morbidity from

*Corresponding author. Data Science Research Centre, Liverpool John Moores University, Liverpool L3 3AF, UK.
E-mail address: S.Ortega-Martorell@ljmu.ac.uk (S. Ortega-Martorell).

Research in context

Evidence before this study

Clinical complexity associated with atrial fibrillation (AF) patients has major implications for treatments and outcomes. To predict AF and AF-related complications, clinical risk scores are commonly employed, but their predictive accuracy is generally limited, given the inherent complexity and heterogeneity of patients with AF. Conventional classification of patients with AF based solely on disease subtypes or arrhythmia patterns (e.g., paroxysmal, persistent, or permanent) may fall short of adequately characterising this diverse population. By classifying different presentations of AF into coherent and manageable clinical phenotypes, the development of tailored prevention and treatment strategies can be facilitated. Previous studies have demonstrated the value of phenotyping, with each identifying between three and six clinically distinct AF phenogroups. However, the methodological approaches followed to derive such phenotypes may not be particularly suited to model complex relationships in the data, and they lack resiliency to data uncertainty and robustness across datasets.

Added value of this study

Our study proposes an AI-based probabilistic approach to identify clinically relevant AF phenotypes for specific patient cohorts, from the general and the critical care populations. Our approach can handle uncertainty, is robust to noise, derives more specific patient profiles, and can uncover hidden subgroups, contributing to more robust patient stratification. We tested our methodology on two large databases, and generated phenotypes using two different AF cohorts: one derived from general population data from the UK-Biobank, and the other derived from critically ill patients admitted to the intensive care unit from the MIMIC-IV database. The phenotypes in both cohorts were derived from vitals and laboratory test data (no medical history/comorbidities or demographic data was explicitly included in the modelling stage to prevent possible bias), and remarkably, the derived phenogroups were still able to identify significant differences in those variables when studied post-hoc. Link to the code: (https://zenodo.org/doi/10.5281/zenodo.12207621).

Implications of all the available evidence

Using our methodology, we identified and characterised clinical phenotypes of AF across diverse patient populations, which could facilitate the tailoring of prevention and treatment programs specific to each phenotype. The proposed approach not only can be used to extract AF phenotypes but can also be applied to other datasets to derive clinically meaningful phenotypes of other conditions.

stroke, heart failure, dementia, and hospitalisations. Patients with AF are often associated with various cardiovascular and non-cardiovascular risk factors,[2] and these often do not occur in isolation, co-existing in clusters of comorbidities, leading to multimorbidity, polypharmacy and frailty.[3] Such clinical complexity associated with patients with AF major implications for treatments and outcomes.[4] To predict AF and AF-related complications, clinical risk scores are commonly employed, but their predictive accuracy is generally limited, given the inherent complexity and heterogeneity of patients with AF.

Artificial Intelligence (AI), and more specifically machine learning (ML), is increasingly used in clinical practice for disease prediction and detection, as well as events and treatment optimisation.[5] Most ML applications in AF leverage supervised ML learning (requiring labelled data), however in recent years, there has been a rise in the application of unsupervised ML approaches as they can be used for exploring and understanding the inherent structure and characteristics of the data without requiring labelled outcomes or targets.

Conventional classification of patients with AF based solely on disease subtypes or arrhythmia patterns (e.g., paroxysmal, persistent, or permanent) may fall short of adequately characterising this diverse population.[1] The task of categorising patients into meaningful subgroups/phenotypes is inherently challenging and susceptible to misclassification. These phenotypes, in the context of medical research, are constructs based on clinical and physiological measurements that enable the characterisation of patient subgroups within a specific disease.[6] They comprise either individual disease attributes or combinations thereof, offering a comprehensive description of distinctions among affected individuals, including clinically significant outcomes such as symptoms, exacerbations, treatment responses, disease progression rate, or mortality. By classifying different presentations of AF into coherent and manageable clinical phenotypes, the development of tailored prevention and treatment strategies can be facilitated. This is aligned with the current holistic approach to AF management,[7] as recommended in guidelines.[8]

Different approaches have been followed previously to identify AF phenotypes such as hierarchal clustering (namely Ward's minimum variance method[9–11] and complete linkage using Gowers distance[12]) and k-prototype.[1] These methods are not particularly suited to model complex relationships in the data, they assume clusters are generally homogeneous, they tend to be less interpretable,[13] they may be sensitive to initialisation,[14,15] they may not handle cluster membership uncertainty, and they lack robustness across datasets.[14] However, these studies all demonstrate the potential value of phenotyping, with each identifying between three and

six clinically distinct AF phenogroups. The population groups studied also vary, including Japanese,[1,10,16] European,[9,11,17] and North American[9] populations.

This study proposes a methodological approach for generating clinically relevant AF phenotypes for specific patient cohorts, from the general and the critical care populations. To test the proposed approach, we generated phenotypes using two different AF cohorts: one derived from general population data from the UK-Biobank, and the other derived from critically ill patients admitted to the intensive care unit (ICU) from the MIMIC-IV database. These databases were chosen as they are both large and offer a rich pool of variables.

Our approach employs generative topographic mapping (GTM),[18,19] a probabilistic ML method chosen for its ability to elucidate meaningful data representations from large datasets. AF phenotypes were derived from the GTM model, and the inherent clinical characteristics associated with each of them were explored for both cohorts.

Methods

Proposed AI-based methodology to generate reliable phenotypes

Micro-cluster segmentation using GTM

Our approach (Fig. 1) first uses GTM, an unsupervised ML methodology grounded in probability theory[18] that offers a principled alternative to the widely used Self Organising Map algorithm.[20] GTM addresses several known issues associated with SOM, such as non-guaranteed convergence, limited neighbourhood preservation, lack of an objective function, and the absence of an explicitly defined probability density function.[21] Alternative algorithms such as t-SNE[22] and UMAP[23] have become popular for reducing dimensionality and visualising data. Whilst they have different mathematical underpinnings, both methods aim to reflect the underlying structure of the data. However, as opposed to GTM, they are not probabilistic methods; t-SNE and UMAP are deterministic techniques that focus on preserving local and global structures without explicitly modelling probability distributions. This is a limitation of the latter two methods since we are interested in generating probabilistic representations and explicit cluster modelling for the AF phenotypes. A probabilistic approach would offer advantages such as uncertainty quantification, robustness to noise, more specific patient profiles, and the ability to uncover hidden subgroups, ultimately contributing to a more robust stratification of patients.

GTM operates by assuming first that the observed data are generated through a nonlinear, topology-preserving mapping from a low-dimensional latent space to a high-dimensional data space. Let the data in the original data space D be represented as $\boldsymbol{x} = (x_1, x_2, ..., x_D)$ and the latent variables be represented as $\boldsymbol{u} = (u_1, u_2, ..., u_L)$. The projection of points from the latent space to the data space is carried out using the non-linear function $y(\boldsymbol{u}; \boldsymbol{W})$ where $\boldsymbol{W}$ represents a set of parameters that maps points $\boldsymbol{u}$ in the latent space into the points $y(\boldsymbol{u}; \boldsymbol{W})$ that lie in the data space. The probability density function of the latent space, $p(\boldsymbol{u})$, is set to the sum of delta functions, as described in eq. (1), constraining the latent points to a uniform discrete grid of centres.

$$p(u)=\frac{1}{K}\sum_{l=1}^{K}\delta(u-u_l)$$

Each centre in the latent space, $\boldsymbol{x}_l$, is responsible for generating a spherical Gaussian density function in the data space centred on $y(\boldsymbol{x}_l; \boldsymbol{W})$, with variance β for a given $\boldsymbol{x}_l$ and $\boldsymbol{W}$. The distribution in the dataspace can

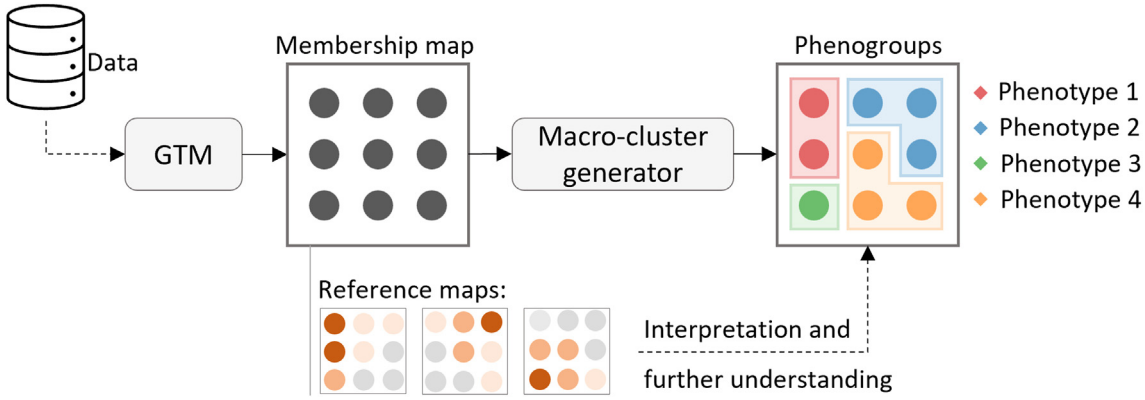


***Fig. 1:* Proposed AI-based methodology to generate reliable phenotypes.** Data is modelled by the GTM algorithm, which projects the data into a 2-dimensional latent space, visualised in the membership map. The GTM also produces reference maps, which are used to indicate the influence of a variable over a micro-cluster. Hierarchical clustering is then applied to the reference vectors to group similar micro-clusters together into larger macro-clusters, which in turn are used to derive the phenotypes.

therefore be understood as a Gaussian mixture model defined by eq. (2).

$$p(x|W,\beta)=\frac{1}{K}\sum_{l=1}^{K}p(x|u_l,W,\beta)$$

Where the parameters $\boldsymbol{W}$ and β can be determined by using maximum likelihood, whereby the log-likelihood is defined as

$$L(W,\beta)=\sum_{n=1}^{N}ln\left\{\frac{1}{K}\sum_{l=1}^{K}p(x_n|u_l,W,\beta)\right\}$$

The optimisation of this log-likelihood is carried out using a variant of the expectation-maximisation (EM) algorithm. For the full details on the calculations, please refer to the original publications.[18,21]

In practice, GTM calculates the probability of an observed data point, represented in here by a patient/participant, belonging to each cluster. The cluster with the highest probability determines the final cluster assignment, resulting in a fine-grained, micro-segmentation of the original data space. This means that GTM performs soft assignments of patients to clusters. This soft assignment strategy yields data clusters within the latent space, where all participants within a given cluster exhibit similar characteristics. This robust approach minimises the likelihood of data clusters comprising dissimilar participants. Since we have chosen a 2-dimensional latent space, these data clusters can be visually represented on a 2-dimensional map, which we will refer to as the "membership map".

To perform the GTM modelling, we used the "ugtm" Python package. As with any ML modelling, a crucial step in the development of ML models is the careful selection of appropriate hyperparameters. This is to ensure the model can learn the key relationships within the data whilst minimising the risk of overfitting and ensuring the model can generalise to unseen data. Although there are scenarios where hyperparameter tuning may be less critical with the GTM method, in this context, where the intended use of phenotypes is not purely prescriptive, the paramount objective was to ensure that the model could generalise effectively, and accurately project new, unseen patients into the most fitting phenotype.

Consequently, we conducted a comprehensive search of a predefined parameter space to identify the most suitable hyperparameters for our model. The specific hyperparameters subjected to tuning included the number of latent clusters (and by extension, the number of Gaussian centres in the data space), the number of radial basis functions (RBFs) (denoted as "W") employed for projecting data from the latent space to the data space, and the penalisation term used to regulate the mapping process.

Each combination of hyperparameters underwent rigorous evaluation through 10-fold cross-validation. The primary performance metric for each test involved assessing the negative log-likelihood of the test data fold projections. The optimal hyperparameters were selected based on their ability to perform exceptionally well on the test data while also exhibiting minimal standard deviation across all results from each cross-validation fold. The results of the hyperparameter tuning showed that the parameter set of a latent space grid size of 15 × 15, 196 RBFs arranged in a 14 × 14 grid with a regularisation term of 1 was optimal and was therefore used when training the GTM models for both the UK Biobank and MIMIC-IV cohorts.

After obtaining a trained GTM model, each cluster centre can be seen as a composite representation of the data residing in the observed data space, hereafter referred to as the "reference vector". The components of these reference vectors, derived from the data used to train the model, serve as the basis for creating reference maps for the variables (Fig. 1), which help to show their influence on each patient cluster through heatmap visualisations, i.e., the intensity of high and low values represents the extent to which each variable influences different areas of the membership map. An additional approach to interpreting the clusters involves superimposing other variables not seen by the model during the training, presented in the form of a heatmap onto the membership map visualisations. This provides users with an alternative method for comprehending the clusters through post-hoc analysis.

A crucial property of GTM is the preservation of data topology, meaning that similar clusters should be positioned closer together in the latent space. Even if the most probable cluster assigned to a participant does not precisely correspond to the actual one, it is expected to be closer to the correct one. This makes GTM representations valuable for visualising complex high-dimensional data in a more interpretable lower-dimensional space. In contrast, common clustering techniques such as k-means, lacking probabilistic foundations, are not specifically designed to handle such levels of uncertainty.

Macro-cluster analysis to generate AF phenotypes

Defining macro-clusters within the array of micro-clusters generated by GTM is crucial for the identification of AF phenotypes. The outcome of such analysis would shed light on regions in the latent space where micro-clusters with similar characteristics are concentrated, representing natural groupings and inherent common patterns in the data space. As defined in eq. (2), the centres in the latent space are projected into the data space to create a non-linear manifold using GTM.

Our approach (Fig. 1) was inspired by an algorithm introduced by Vellido et al.[24] Instead of identifying macro-cluster regions in the latent space, we used

agglomerative hierarchal clustering using Ward's minimum variance method[25] on the reference vectors, and the distances between the vectors were computed using the Euclidean metric. The reference vectors corresponded to the Gaussian centres projected from the centres in the latent space, each residing in the data space. Subsequently, the cluster assignment of each reference vector was mapped to their respective centres in the latent space, effectively generating the desired macro-clusters comprising the latent space's micro-cluster centres. The full code implementing this approach can be found on Zenodo[26] at the following link (https://zenodo.org/doi/10.5281/zenodo.12207621).

Data used for deriving AF phenotypes

Modelling variables extracted from the UK-Biobank database

The first dataset used for this analysis was a subset extracted from the UK-Biobank, a large, population-based database[27] encompassing over 500,000 participants aged 40–69 from across the UK. To identify eligible participants with AF, we searched ICD-10 codes related to AF diagnosis recorded in the participants' conditions and causes of death variables. Eligible participants would have at least one of these codes recorded. See list of codes in SM, Supplementary Table S1.

In total, 67 variables from the UK-Biobank were used for modelling, 40 genomic variables and 27 biological sample variables. We only included these variables to ensure that participants were clustered based on the similarity of their biological and genetic profiles, rather than being influenced by external demographic factors. The genomic variables are a set of 40 principal components generated using >100,000 single nucleotide polymorphisms (SNPs).[28] The 27 biological sample variables selected aim to represent key risk markers associated with AF: clotting, inflammation, renal function, liver function, cholesterol, diabetes, and sex-related markers.[29]

Modelling variables extracted from the MIMIC-IV database

Data was extracted from the Medical Information Mart for Intensive Care IV (MIMIC-IV[30]), a freely available database of de-identified electronic health records linked to patients admitted to the Beth Israel Deaconess Medical Centre in Boston, Massachusetts. We used version 2.2 (January/2023), which includes 73,181 ICU stays.

Patients were included in this study if they had at least one episode of AF during the ICU admission. The latter was extracted from the *chartevent* table, using the code for heart rhythm: 220048, and identifying from those the ones that have value "AF (Atrial Fibrillation)". Therefore, this would include patients with pre-existing AF, and those with new-onset AF, although the first AF episode recorded occurred after the first 24 h of the ICU admission. Patients <18 years old, patient admissions with short ICU stays (<24 h), and patients with multiple ICU stays were excluded from the study.

In total, 21 variables from the MIMIC-IV database were used for modelling. These variables were extracted from sequences of vitals (e.g., temperature, and heart rate) and lab test results (e.g., glucose and haemoglobin) used to monitor the condition of the patient in the ICU. The variables used for modelling were selected as they represent key risk markers associated with AF in ICU.[31,32]

Selection of variables associated with AF

AF in the general population: UK-Biobank data

AF is associated with ageing and comorbidities, as reflected in our phenotypic data. Indeed, multiple studies have shown how comorbid risk factors do not occur in isolation, but cluster together contributing to clinical complexity phenotypes.[3,4] There are well-recognised associations of common comorbidities such as hypertension, heart failure and diabetes, as well as renal and liver dysfunction.[33] The choice of biological sample variables selected for our modelling aims to represent key risk markers associated with AF since they are essential for a comprehensive understanding of the factors contributing to AF. For example, inflammatory processes play a role in the development and progression of AF.[34] Certain genetic variants have also shown significant association with silent AF.[35]

Various risk prediction tools have been proposed for the prediction of incident AF,[36] e.g., CHARGE-AF (The Cohorts for Heart and Ageing Research in Genomic Epidemiology AF) score, developed for the general population, which uses variables such as age, ethnicity, height, weight, blood pressure, medication use, and comorbidities.[37] Simpler clinical risk factor scores such as C_2HEST have also been investigated to predict incident AF in population and post-stroke cohorts.[38]

AF in the critical care population: UK-Biobank data

AF stands as the most prevalent arrhythmia among critically ill patients, occurring at an incidence rate of 10–15%[39] within the critical care population. Patients in the ICU that have AF suffer with a worse prognosis, longer ICU stays and higher mortality.[40] Treatments for managing AF that are used for patients in the general population may not be appropriate for critically ill patients,[41] therefore having ICU focused results is crucial for optimising patient outcomes. The risk factors for AF can significantly differ between the general and the critical care populations. Common risk factors for AF in the community involve structural and valvular heart disease, but these factors may not be distinctly associated with AF in critical illness.[42] In addition, acute factors are thought to be associated with increased risk for newly diagnosed AF during critical illness.[37] For example, invasive ventilation is associated with AF episodes in critically ill patients.[42] Monitoring oxygenation is crucial in these patients to assess respiratory function and optimise oxygen delivery, as compromised

oxygenation can exacerbate cardiovascular stress and contribute to complications.[43] Electrolyte imbalances, such as phosphate abnormalities, observed in medical conditions like kidney dysfunction, may indirectly contribute to AF development.

Additional investigative variables

Additional investigative variables were extracted for further exploration, and they were not used during model development. Instead, they are employed post-hoc to provide further context/insights related to the composition of individual or group of clusters and to help identify potential meaningful AF phenotypes.

Investigative variables extracted from the UK-Biobank database

We used a set of 18 UK-Biobank variables for visualisation purposes. This selection consisted of 15 assessment centre variables, and two population characteristic variables, with the remaining variable belonging to the health-related outcomes category. Several of these variables were previously identified in prior AF studies[29] and includes sex (determined either from the NHS central registry or by what was self-reported by the participant), BMI, activity levels and alcohol consumption.

We consider that incorporating comorbidity data is fundamental for understanding how various medical conditions can be differentiated among clusters of AF participants in the general population. To effectively convey information on thousands of diverse comorbidities in a clear, meaningful manner, we integrated the use of *phecodes*.[44] Each phecode is composed of several individual diagnoses, defined using ICD-10 codes, which are subsequently grouped into various phecode categories.

In our analysis of AF participants from the general population using UK-Biobank data, we included several phecode categories that encompassed diagnoses from a predefined set of comorbidities commonly associated with individuals suffering from AF. To assign a phecode, and subsequently associate it with a phecode category, a patient's record was examined for a match with the ICD-10 code of either primary or secondary diagnoses to one within a phecode. The list of all phecodes, and their respective phecode categories, that were considered in this study can be found in Supplementary Table S2. For the full details regarding which ICD10 codes make up each phecode, please refer to the original publication.[44]

Investigative variables extracted from the MIMIC-IV database

A selection of 27 variables from the MIMIC-IV database were extracted for further investigation. They include demographic data such as sex (reported in the dataset as gender however we used this term as it is more appropriate as it refers to the biological sex of the patient), age, and ethnicity. They also include the Glasgow Coma Scale (GCS), a neurological assessment tool commonly employed in critical care settings, which is used to evaluate a patient's level of consciousness based on their eye, verbal, and motor responses. Ventilation status (invasive and non-invasive), acute kidney injury (AKI) and acute respiratory distress syndrome (ARDS) are also investigated as variables of interest, as well as a series of variables related to length of stay and mortality.

Data pre-processing

To ensure the development of a robust and representative dataset for modelling, we undertook several pre-processing steps. First, we implemented a set of missingness criteria (defining appropriate levels/thresholds of data completion) to determine which variables and participants to include. The thresholds were set at 25% and 30% for data that could be missing for a variable or a participant, respectively. We also identified certain variables that exhibited positive skewness in their value distributions. To address this, we applied a log transformation to these variables, rendering their distributions more Gaussian in nature.

Subsequently, any remaining missing data were addressed through imputation, employing a multivariate imputer. This imputer estimated missing values by considering known values from other variables. To accomplish this, we utilised the "IterativeImputer" function, which is part of the Scikit-Learn Python package and draws inspiration from the R MICE package 6. Invalid values of the variables (e.g., heart rate < 0) were marked as not available. Variables recorded with different units were harmonised, e.g., in MIMIC-IV, height was present in inches and centimetres (cm), and they were all converted to cm.

Ethics

The UK Biobank is approved from the North West Multi-centre Research Ethics Committee as a Research Tissue Bank and researchers do not require separate ethical clearance. The use of MIMIC-IV data did not require ethical approval as the analysis is based on secondary data which is publicly available, and no permission is required to access the data.

Statistics

Medians and interquartile ranges were calculated for continuous variables, and frequencies and proportions (percentages) were used for categorical variables. There were several ordinal variables used for the exploratory analysis of the GTM output. These were one-hot encoded and then treated as a categorical variable and represented in the data as such.

To study the characteristics of the generated phenotype groups, differences between continuous variables were analysed using the Kruskal–Wallis test and differences between categorical variables were analysed using the Chi-squared test. In both cases, a p-value < 0.05 was the threshold for statistical significance.

Variable name	Value
Modelling variables:	
Inflammation markers:	
Neutrophil count [x10^9 cells/L]	4.3 (3.49, 5.24)
Lymphocyte percentage [%]	27.03 (22.3, 31.93)
Monocyte percentage [%]	7.24 (5.91, 8.68)
C-reactive protein [mg/L]	1.77 (0.86, 3.57)
Clotting markers:	
Haematocrit percentage [%]	41.78 (39.31, 44.1)
Mean corpuscular volume [Femtolitres]	91.7 (88.95, 94.5)
Red blood cell (erythrocyte) distribution width [%]	13.5 (13.07, 14.09)
Platelet count [x10^9 cells/L]	235 (201, 274)
Mean platelet (thrombocyte) volume [Femtolitres]	9.3 (8.64, 10.06)
Platelet distribution width [%]	16.5 (16.2, 16.86)
Mean reticulocyte volume [Femtolitres]	106.99 (102.5, 111.83)
Mean sphered cell volume [Femtolitres]	83.1 (79.8, 86.66)
Diabetes risk markers:	
Glucose [mmol/L]	5.04 (4.68, 5.49)
Glycated haemoglobin (HbA1c) [mmol/mol]	36.4 (33.8, 39.5)
Liver function:	
Albumin [g/L]	44.65 (43.13, 46.1)
Alanine aminotransferase [U/L]	21.56 (16.68, 28.19)
Direct bilirubin [umol/L]	1.74 (1.39, 2.24)
Gamma glutamyltransferase [U/L]	32.4 (22.2, 50.3)
Renal function:	
Creatinine [umol/L]	75.6 (65.6, 86.1)
Sodium in urine [millimole/L]	69.3 (44.0, 100.5)
Urea [mmol/L]	5.69 (4.85, 6.63)
Urate [umol/L]	338.01 (284, 393.7)
Cholesterol markers:	
Cholesterol [mmol/L]	5.31 (4.53, 6.09)
HDL cholesterol [mmol/L]	1.32 (1.11, 1.57)
Triglycerides [mmol/L]	1.6 (1.14, 2.23)
Sex-related markers:	
SHBG [nmol/L]	44.98 (33.62, 58.9)
Testosterone [nmol/L]	8.73 (1.62, 12.2)
Additional investigative variables:	
Demographics:	
Age at recruitment [years]	63 (59, 67)
Sex	
Male	23,284 (63.5%)
Female	13,396 (36.5%)
Waist circumference [cm]	96 (87, 106)
Hip circumference [cm]	105 (99, 111)
Standing height [cm]	172 (164, 178)
Weight [kg]	83.3 (72.9, 95)
BMI [kg/m^2]	28.16 (27.1, 29.98)
Activity level:	
Summed minutes activity [mins]	95 (50, 180)
MET minutes per week for vigorous activity [mins/week]	120 (0, 720)
Blood pressure:	
Diastolic blood pressure, automated reading [mmHg]	82 (75, 90)
Systolic blood pressure, automated reading [mmHg]	143 (130, 157)
Pulse rate, automated reading [bpm]	68 (60, 77)
Respiratory measures:	
Forced expiratory volume in 1 second (FEV1) [L]	2.68 (2.15, 3.27)
Peak expiratory flow (PEF) [L/min]	383 (295, 484)
Forced expiratory volume in 1 second (FEV1) Z-score	0.62 (-0.12, 1.37)
FEV1/ FVC ratio Z-score	0.4 (-0.13, 1)
Alcohol intake frequency:	
Daily or almost daily [yes]	7,170 (19.6%)
Three or four times a week [yes]	6,417 (17.5%)
Once or twice a week [yes]	6,869 (18.7%)
One to three times a month [yes]	2,734 (7.5%)
Special occasions only [yes]	3,354 (9.1%)
Never [yes]	2,734 (7.5%)
Ethnic background:	
White [yes]	35,536 (96.9%)
Asian or Asian British [yes]	406 (1.1%)
Black or Black British [yes]	247 (0.7%)
Mixed [yes]	111 (0.3%)
Other ethnic group [yes]	160 (0.4%)
Chinese [yes]	36 (0.1%)
AF and flutter diagnosis (main/secondary):	
ICD10 - AF and flutter [yes]	20,966 (57.2%)
ICD10 - Paroxysmal AF [yes]	6,558 (17.9%)
ICD10 - Persistent AF [yes]	1,274 (3.5%)
ICD10 - Chronic AF [yes]	570 (1.6%)
ICD10 - Typical AF [yes]	216 (0.6%)
ICD10 - Atypical atrial flutter [yes]	86 (0.2%)
ICD10 - AF and atrial flutter, unspecified [yes]	21,767 (59.3%)
Systems (phecode categories):	
Endocrine/metabolic [yes]	10,119 (27.6%)
Circulatory system [yes]	26,628 (72.6%)
Respiratory [yes]	6,097 (16.6%)
Diabetes:	
Type 1 diabetes [yes]	839 (2.3%)
Type 1 diabetes with ketoacidosis [yes]	81 (0.2%)
Type 1 diabetes with renal manifestations [yes]	60 (0.2%)
Type 1 diabetes with ophthalmic manifestations [yes]	175 (0.5%)
Type 1 diabetes with neurological manifestations [yes]	96 (0.3%)
Diabetes type 1 with peripheral circulatory disorders [yes]	52 (0.1%)
Type 2 diabetes [yes]	7,130 (19.4%)
Type 2 diabetes with ketoacidosis [yes]	96 (0.3%)
Type 2 diabetes with renal manifestations [yes]	233 (0.6%)
Type 2 diabetes with ophthalmic manifestations [yes]	852 (2.3%)
Type 2 diabetes with neurological manifestations [yes]	427 (1.2%)
Diabetes type 2 with peripheral circulatory disorders [yes]	351 (1%)
Hypertension:	
Essential hypertension [yes]	24,442 (66.6%)
Other hypertensive complications [yes]	86 (0.2%)
Cardiovascular disease:	
Myocardial infarction [yes]	6,544 (17.8%)
Other forms of chronic heart disease [yes]	2 (0%)
Congestive heart failure (CHF) NOS [yes]	3,760 (10.3%)
Chronic pulmonary heart disease [yes]	1,105 (3%)
Heart failure NOS [yes]	4,680 (12.8%)
Coronary atherosclerosis [yes]	163 (0.4%)
Peripheral vascular disease:	
Peripheral vascular disease, unspecified [yes]	1,911 (5.2%)
Other specified peripheral vascular diseases [yes]	23 (0.1%)
Pulmonary hypertension:	
Primary pulmonary hypertension [yes]	403 (1.1%)

(Table 1 continues on next column)

(Continued from previous column)

Stroke:	
Hemiplegia [yes]	1,214 (3.3%)
Liver disease:	
Liver abscess and sequelae of chronic liver disease [yes]	373 (1%)
Alcoholic liver damage [yes]	379 (1%)
Other chronic non-alcoholic liver disease [yes]	1,441 (3.9%)
Other disorders of the liver [yes]	808 (2.2%)
Kidney disease:	
End-stage renal disease [yes]	484 (1.3%)

All data presented below was taken from the first data instance available. Medians and interquartile ranges were calculated for continuous variables, and frequencies and proportions (as percentages) were calculated for the categorical variables. Red shades were used for the modelling variables, whilst blue was used for the additional investigative variables.

Table 1: **Characteristics of the participant subset extracted from the UK-Biobank database.**

Role of funders

The funders did not participate in the study's design and implementation, data collection, management, analysis, or interpretation. They were also not involved in the preparation, review, or approval of the manuscript, nor in the decision to submit it for publication.

Results

Characteristics of the participants/patients cohorts

From the UK-Biobank we extracted 36,680 participants with AF from this general population cohort using the criteria set out in 2.2.1 (median age 63 years (IQR 59–67), range 40–72 years; 63.5% male). Table 1 contains the summary of the biological variables used for modelling, and the investigative variables used in the post-hoc analysis. A second dataset of 2695 critically ill patients with AF using the criteria set out in section 2.2.2 (median age 73 years (IQR 65–81), range 21–89 years; 60.4% male) was extracted from the MIMIC-IV (Table 2).

Visualisation of reference vectors for the modelling variables

Reference vectors of the modelling variables – used to derive AF phenotypes

Fig. 2 contains the reference vectors extracted from the trained GTM models for the UK-Biobank and MIMIC-IV AF cohorts. For the UK-Biobank data, it contains the reference vectors for the biological sample variables, with plots grouped by the different risk factors they relate to, whilst for the MIMIC-IV, it displays all modelling variables used for modelling. Each point in every plot within Fig. 2 corresponds exactly to the same point in their respective membership map (SM, Supplementary Fig. S1). A light grey–red colour scheme was used for the reference vectors plot such that areas of the plots that are redder indicate that participants in that cluster had a higher value of that variable. Likewise, if the point in the reference vector is greyer, the lower the value is for participants in this cluster. All plots using the light grey–red colour scheme indicate variables used in the GTM model development, whereas plots using a light grey–teal

Variable name	Value
Modelling variables:	
Diabetes risk marker:	
Glucose [mg/dL]	131.88 (118.17, 155.5)
Bone profile:	
Phosphate [mg/dL]	3.58 (3.05, 4.22)
Oxygenation:	
Oxygen saturation [%]	96.33 (94.38, 97.83)
Respiratory rate [breaths per min]	18.51 (16.5, 21.27)
Fraction inspired oxygen, FiO2 [%]	56.47 (50, 63.24)
Positive end-expiratory pressure (PEEP) [cmH2O]	5.6 (5, 7.11)
Partial pressure of oxygen [mmHg]	135.08 (99.15, 168.5)
Haemoglobin [g/dL]	10.16 (9.11, 11.48)
Respiratory/metabolic markers:	
pH	7.35 (7.21, 7.39)
Anion Gap [mEq/L]	13.42 (11.33, 16.0)
Lactate [mmol/L]	2.0 (1.49, 2.75)
Cardiac markers:	
Heart rate [beats per min]	81.33 (74.24, 90.42)
Capillary refill rate	0.0 (0.0, 0.02)
Diastolic BP [mmHg]	57.25 (51.5, 63.38)
Systolic BP [mmHg]	111.93 (104.73, 121.34)
Clotting markers:	
Prothrombin time [sec]	14.47 (13.07, 16.45)
Platelet count [K/uL]	165.0 (125.12, 223.0)
Renal function:	
Creatinine [mg/dL]	1.03 (0.8, 1.56)
Electrolytes:	
Magnesium [mg/dL]	2.15 (1.91, 2.44)
Potassium [mEq/L]	4.29 (3.95, 4.61)
Other:	
Temperature [°C]	36.74 (36.55, 37.0)
Additional investigative variables:	
Demographics:	
Age [years]	73 (65, 81)
Sex	
Male	1,627 (60.4%)
Female	1,068 (39.6%)
Height [cm]	170.09 (162.78, 177.9)
Weight [kg]	82.43 (68.39, 97.37)
Ethnicity:	
White [yes]	1971 (73.1%)
Other ethnic group [yes]	453 (16.8%)
Black [yes]	138 (5.1%)
Hispanic [yes]	68 (2.5%)
Asian [yes]	65 (2.4%)
Glasgow Coma Scale (GCS):	
GCS eye-opening	2.88 (1.92, 3.75)
GCS motor response	4.83 (3.5, 6)
GCS verbal response	2.54 (1, 4.33)
Ventilation:	
Non-Invasive ventilation [yes]	209 (7.8%)
Invasive ventilation [yes]	2116 (78.5%)
Outcomes:	
Time to AF diagnosis [hours]	53 (38, 83)
In-hospital length of stay [hours]	256.78 (166.48, 407.12)
In-ICU length of stay [hours]	109.18 (72.9, 200.43)
Death after ICU [hours]	167.07 (17.64, 2700.04)
Death after hospital discharge [hours]	20.44 (10.55, 2551.06)
Death after hospital discharge [days]	0.85 (0.44, 106.29)
In-hospital mortality [yes]	567 (21.0%)
In-ICU length of stay of 3+ days [yes]	2040 (75.7%)
In-ICU length of stay of 7+ days [yes]	840 (31.2%)
Mortality after hospital discharge within 30 days [yes]	711 (26.4%)
Mortality after hospital discharge within 365 days [yes]	936 (34.7%)
Mortality after hospital discharge after 365 days [yes]	152 (5.6%)
Acute Kidney Injury (AKI) [yes]	545 (20.2%)
Acute Respiratory Distress Syndrome (ARDS) [yes]	174 (6.5%)

Summary statistics and colours as in Table 1. The data represented for each variable is the average of all data recorded during the ICU stay.

Table 2: **Characteristics of the ICU patient subset extracted from the MIMIC-IV database.**

represent variables that were not used in the modelling and have no direct impact on the clusters themselves.

Visualisation of the additional investigative variables

Fig. 3 contains a selection of visualisations showing how data from different investigative variables are distributed within the membership maps for the UK-Biobank and MIMIC-IV cohorts. The visualisations representing the investigative variables all use a light grey-teal colour scheme as they were not used in model development. The value assigned to each micro-cluster is the average of the variable for all participants assigned to each cluster, the more teal a micro-cluster is, the higher the value. In SM, section 5, visualisations for all investigated variables are displayed.

Description of AF phenotypes

For the UK-Biobank cohort, we identified five clusters within the reference vectors residing in the data space, as demonstrated by the dendrogram in Fig. 4(a). Transferring these reference vector cluster assignments to their corresponding latent centres gave five macro-cluster regions, which in turn were used to define the five AF phenotypes. These macro-cluster regions are visualised in Fig. 4(b) and (c).

When applied to the MIMIC-IV cohort, the analysis identified four clusters within the reference vectors, as presented in Fig. 5(a). The macro-cluster regions generated by transferring these clusters to their respective latent centres are presented in Fig. 5(b) and (c). The baseline data for each of the two databases were split according to the number of phenotypes and compared, in Tables 3 and 4 for the UK-Biobank and MIMIC-IV data, respectively. A description of the headline features that characterise both sets of phenotypes can be found in Fig. 4(d) and 5(d).

Interpreting the visualisations

The membership maps show us which participants share the same cluster indicating that they share similar features. The probabilistic foundations of GTM allows us to calculate, for each data point, the probability that it was generated from the ith latent node. By calculating the probability for each latent node and overlaying the result onto the membership it allows the user to visualise the probability distribution for each data point. Some examples of this are displayed in Fig. 6. Fig. 6(a) and (b) are the probability distributions for participants from the UK Biobank dataset and Fig. 6(c) and (d) are the probability distributions for patients from the MIMIC-IV dataset. As discussed in section 2.1.1, the latent node that has the highest distribution of generating the data point determines its final cluster assignment. These plots illustrate the soft cluster assignments GTM performs, whilst also demonstrating the robustness of the approach in so far as the next highest probability surround the node the data point was assigned to.

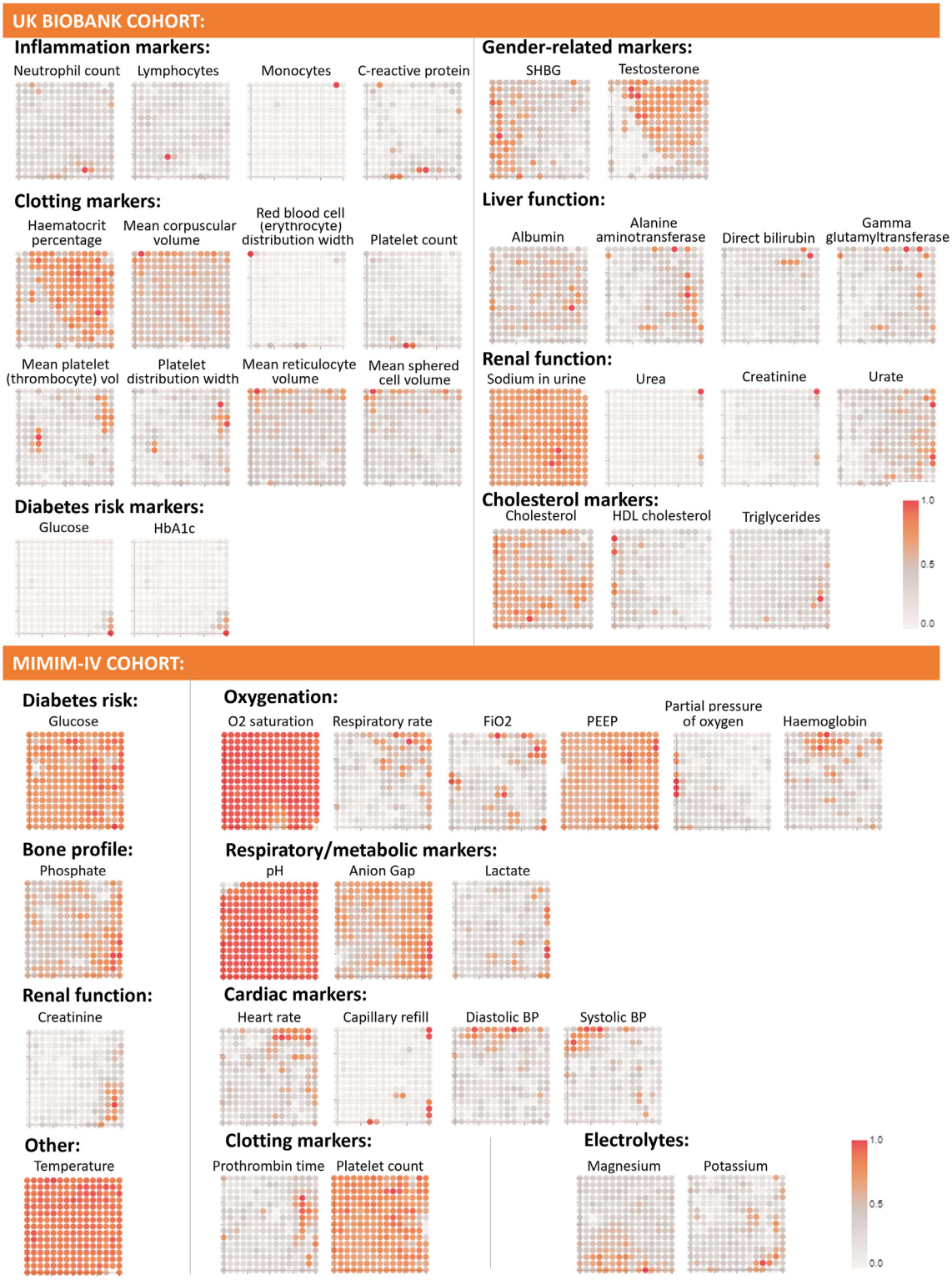


Fig. 2: Reference vector visualisations demonstrating how each biological sample variable affects the cluster distribution in the latent space for both, the UK-Biobank and the MIMIC-IV AF cohorts.

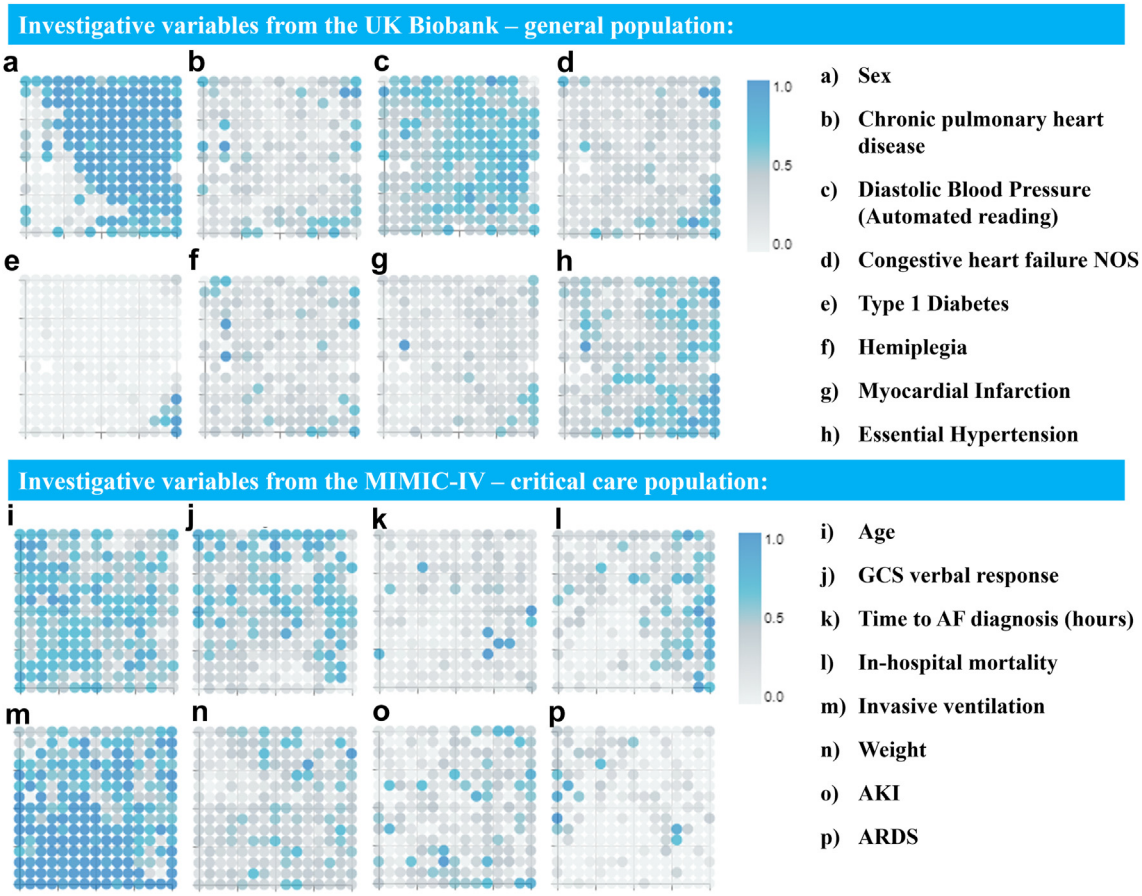


***Fig.* 3: Membership maps showing how a selection of investigative variables data are distributed within the latent space for the UK-Biobank and the MIMIC-IV cohorts.** AF: Atrial Fibrillation. AKI: Acute Kidney Injury. ARDS: Acute Respiratory Distress Syndrome. GCS: Glasgow Coma Scale.

To unlock deeper insights, superimposing modelling data onto the membership maps provides a better understanding of why patients were clustered in such a way (Figs. 2 and 3). Extra insights can be learnt by superimposing post-hoc data, unseen during modelling. One example from the UK-Biobank cohort relates to sex-related markers, specifically testosterone and SHBG levels. By assessing their respective reference vectors, individuals with higher testosterone and lower SHBG tended to be in the middle and top-right sections of the membership map. In contrast, those with heightened SHBG and lower testosterone were clustered towards the bottom left. Given that testosterone levels are generally higher in males,[45] and SHBG levels are typically elevated in females,[46] we can deduce that the membership map effectively delineated male and female participants during clustering. This can be seen in Fig. 3(A), where we visually represent the participants' sex (the bluer area in Fig. 3(A) predominantly corresponds to males), and in Supplementary Fig. S2 (in SM), which shows the membership map stratified by sex.

Discussion

Using our AI methodology, we have identified and characterised clinical phenotypes of AF across diverse patient populations, which could facilitate the tailoring of prevention and treatment programs specific to each phenotype. The principal findings of this study are: (i) The proposed AI-based methodology showed its ability to derive meaningful clinical phenotypes of AF in the general and critical care populations. (ii) Our approach is probabilistic, offering advantages such as the ability to handle uncertainty, robustness to noise, more specific patient profiles, and the ability to uncover hidden subgroups, contributing to more robust patient stratification and visualising complex high-dimensional data in a more interpretable lower-dimensional space, enhancing understanding.

Identifying clinical phenotypes of diseases using methods like hierarchical clustering (specifically Ward's minimum variance method and complete linkage with Gowers distance) and k-prototype used in previous phenotyping studies,[9–12] may not always be the best

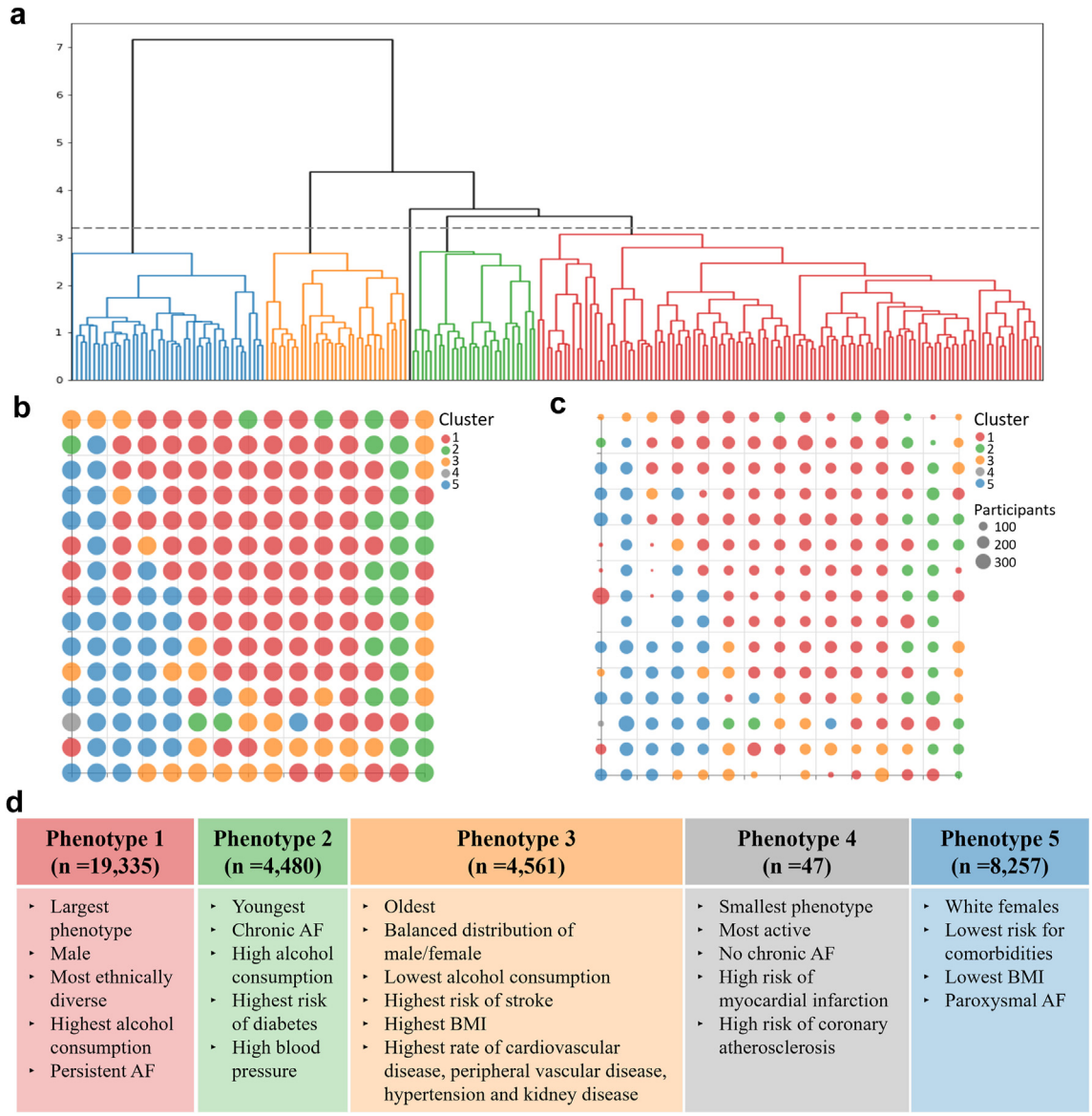
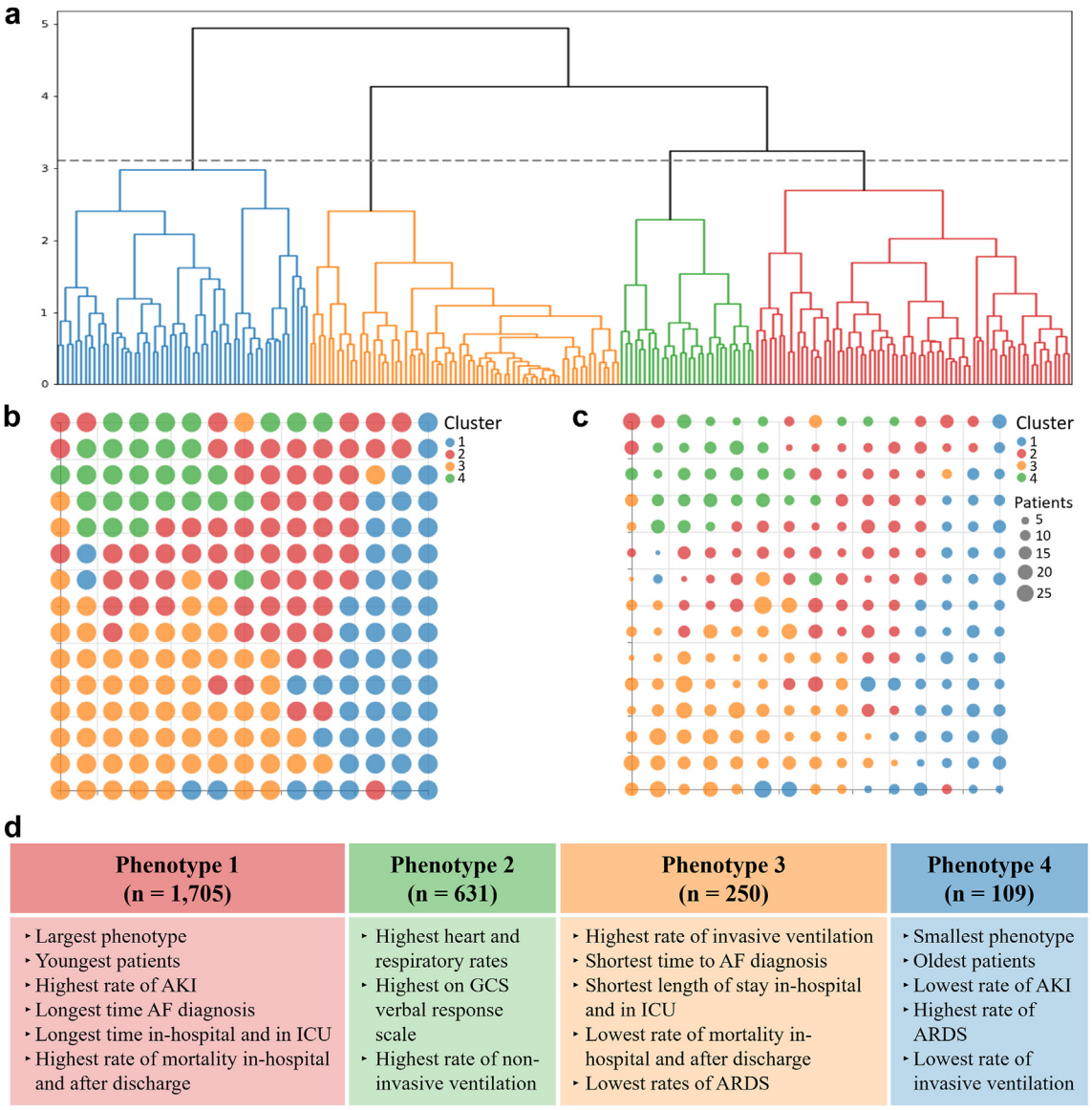


Fig. 4: **Derived phenotypes of AF in the general population using UK-Biobank data.** a) Dendrogram produced using Ward's minimum variance method. The graph shows the 5 clusters that are used to define the 5 AF phenotypes for the general population. b) Membership map with a uniform size for the micro-clusters to show the distribution of the macro-cluster regions. c) The size of the micro-clusters in the membership map dictated by the number of participants assigned to it. d) Main characterising features for each of the phenotypes.

option for several reasons: (i) Clinical data often contains diverse information, and these methods may not effectively capture the complexity of relationships within the data, and they may also be influenced by outliers or noise.[47] (ii) In clinical phenotyping, diseases may exhibit considerable heterogeneity,[9–12] however hierarchical clustering assumes that data points within a cluster are homogeneous. (iii) High-dimensional clinical data may pose challenges for hierarchical clustering and k-prototype methods for interpreting results, which in the context of clinical phenotypes may render unintuitive.[13,47] (iv) In the case of k-prototype, it can be sensitive to the choice of initial cluster centroids and may converge to local minima.[13] (v) Clinical data often includes a mix of continuous and categorical variables. Some clustering methods, like k-prototype, handle both types, but the integration of different variable types can be challenging and may not fully capture the information. (vi) Results obtained from these methods may not generalise well across different datasets or populations



d

Phenotype 1 (n = 1,705)	Phenotype 2 (n = 631)	Phenotype 3 (n = 250)	Phenotype 4 (n = 109)
‣ Largest phenotype ‣ Youngest patients ‣ Highest rate of AKI ‣ Longest time AF diagnosis ‣ Longest time in-hospital and in ICU ‣ Highest rate of mortality in-hospital and after discharge	‣ Highest heart and respiratory rates ‣ Highest on GCS verbal response scale ‣ Highest rate of non-invasive ventilation	‣ Highest rate of invasive ventilation ‣ Shortest time to AF diagnosis ‣ Shortest length of stay in-hospital and in ICU ‣ Lowest rate of mortality in-hospital and after discharge ‣ Lowest rates of ARDS	‣ Smallest phenotype ‣ Oldest patients ‣ Lowest rate of AKI ‣ Highest rate of ARDS ‣ Lowest rate of invasive ventilation

Fig. 5: **Derived phenotypes of AF in the general population using MIMIC-IV data.** a) Dendrogram produced using Ward's minimum variance method. The graph shows the 4 clusters that are used to define the 4 AF phenotypes for ICU patients. b) Membership map with a uniform size for the micro-clusters to show the distribution of the macro-cluster regions. c) The size of the micro-clusters in the membership map dictated by the number of participants assigned to it. d) Main characterising features for each of the phenotypes.

due to variations in data characteristics.[11] (vii) They lack probabilistic foundations and hence are not specifically designed to handle such levels of uncertainty.[18,19]

Alternative approaches, such as probabilistic or ensemble methods, may provide more robust and interpretable clinical phenotypes. Our approach involves deriving micro-clusters using a probabilistic method (i.e., GTM), followed by hierarchical clustering to identify macro-clusters, i.e., the phenotypes. The latter differs from previous studies as the hierarchical methods were applied to the reference vectors from a probabilistic model rather than the original data space, which makes the clusters more stable and resilient to data uncertainty. Our use of GTM often provides highly interpretable representations as it explicitly models clusters and prototypes, offering insights into the underlying structure of the data. The membership map produced by GTM captures the underlying relationships and clusters within the data by mapping data points to these prototypes. This enables comprehensible and interpretable representations of complex data, aiding in knowledge extraction and facilitating insights that might otherwise remain hidden

Variable name	Phenotype 1 (n =19,335)	Phenotype 2 (n =4,480)	Phenotype 3 (n =4,561)	Phenotype 4 (n =47)	Phenotype 5 (n =8,257)	p-value
MODELLING VARIABLES:						
Inflammation markers:						
Neutrophil count	4.27 (3.46, 5.18)	4.37 (3.57, 5.3)	5.1 (4.1, 6.35)	4.46 (3.95, 5.22)	3.99 (3.25, 4.77)	<0.05
Lymphocyte percentage	26.53 (22.1, 31.4)	27.27 (22.63, 32.1)	24.5 (19.01, 29.74)	26.15 (23.53, 29.4)	29.2 (24.67, 33.9)	<0.05
Monocyte percentage	7.6 (6.24, 9.04)	7.4 (6.11, 8.8)	6.7 (5.4, 8.15)	7.45 (6.16, 8.76)	6.7 (5.53, 7.91)	<0.05
C-reactive protein	1.54 (0.79, 2.94)	2.15 (1.07, 4.11)	4.75 (2.08, 10.82)	2.05 (0.9, 3.33)	1.44 (0.72, 2.86)	<0.05
Clotting markers:						
Haematocrit percentage	43 (40.98, 44.93)	42.92 (40.6, 45.13)	39.82 (37.39, 42.18)	42.3 (38.6, 45.16)	39.3 (37.55, 41.07)	<0.05
Mean corpuscular volume	92.06 (89.46, 94.73)	91.82 (89.03, 94.9)	90.1 (86.8, 93.28)	91.6 (89.03, 93.55)	91.53 (88.9, 94.12)	<0.05
Red blood cell distribution width	13.5 (13.06, 14)	13.43 (13, 13.99)	13.95 (13.34, 14.89)	13.6 (13.1, 13.94)	13.47 (13, 14)	<0.05
Platelet count	228 (198, 261.45)	209 (174, 248.53)	262 (223.6, 308)	242 (197.45, 275.5)	253.4 (218.6, 292.8)	<0.05
Mean platelet volume	9.27 (8.6, 9.91)	9.9 (9, 10.95)	9.19 (8.53, 9.8)	9.17 (8.65, 10.01)	9.3 (8.61, 10.04)	<0.05
Platelet distribution width	16.5 (16.2, 16.8)	16.9 (16.5, 17.36)	16.49 (16.2, 16.8)	16.5 (16.17, 16.9)	16.37 (16.08, 16.7)	<0.05
Mean reticulocyte volume	107.37 (102.93, 112.11)	106.47 (101.9, 111.62)	106.39 (101.8, 111.82)	105.6 (101.82, 108.46)	106.6 (102.28, 111.3)	<0.05
Mean sphered cell volume	83.27 (80, 86.7)	82.55 (79.36, 86.5)	81.9 (78.5, 85.56)	81.71 (79.19, 85.15)	83.7 (80.4, 87.13)	<0.05
Diabetes risk markers:						
Glucose	5.02 (4.66, 5.44)	5.28 (4.8, 6.36)	5.13 (4.72, 5.76)	5.09 (4.73, 5.42)	4.97 (4.67, 5.31)	<0.05
HbA1c	36.2 (33.6, 39.1)	37.6 (34.2, 44.63)	38.5 (35.6, 42.6)	37.2 (33.35, 40.95)	35.6 (33.4, 37.9)	<0.05
Liver function:						
Albumin	44.81 (43.38, 46.2)	45.09 (43.42, 46.9)	43.72 (41.96, 45.22)	44.42 (43.41, 46.41)	44.5 (43.08, 45.86)	<0.05
Alanine aminotransferase	22.72 (17.88, 28.67)	30.56 (22.69, 42.64)	20.25 (15.68, 26.3)	21.54 (16.14, 28.11)	17.33 (14.13, 21.39)	<0.05
Direct bilirubin	1.91 (1.52, 2.41)	1.88 (1.47, 2.46)	1.57 (1.25, 1.99)	1.66 (1.31, 2.11)	1.48 (1.22, 1.81)	<0.05
Gamma glutamyltransferase	34.1 (24.3, 50.6)	53.9 (34.5, 96.3)	34.1 (24.2, 52.3)	34.9 (22.1, 51.55)	22 (16.9, 31.3)	<0.05
Renal function:						
Creatinine	79.8 (71.7, 88.8)	77.1 (67.2, 87.4)	76.1 (64, 95)	81.8 (63.3, 90.25)	63.8 (57.3, 71.5)	<0.05
Sodium in urine	76.4 (49.5, 108.6)	74.9 (48.9, 106)	69 (43.5, 96.3)	57.4 (35.65, 86.15)	53.2 (34.3, 77.7)	<0.05
Urea	5.73 (4.94, 6.63)	5.69 (4.83, 6.66)	6.08 (5, 7.79)	5.94 (5.05, 6.52)	5.41 (4.61, 6.23)	<0.05
Urate	354.8 (310.5, 402.6)	370.1 (312.3, 428.42)	354.44 (297.4, 429)	358.9 (312.05, 402.7)	269.3 (230.3, 311.2)	<0.05
Cholesterol markers:						
Cholesterol	5.16 (4.4, 5.88)	5.3 (4.44, 6.18)	5.09 (4.29, 5.95)	5.13 (4.41, 6.07)	5.8 (5.1, 6.53)	<0.05
HDL cholesterol	1.26 (1.08, 1.46)	1.17 (0.99, 1.43)	1.24 (1.04, 1.45)	1.27 (1.16, 1.56)	1.6 (1.4, 1.84)	<0.05
Triglycerides	1.6 (1.14, 2.18)	2.24 (1.44, 3.4)	1.82 (1.32, 2.5)	1.78 (1.27, 2.44)	1.32 (0.99, 1.76)	<0.05
Sex-related markers:						
SHBG	42.75 (33.16, 53.73)	37.27 (26.76, 50.01)	38.3 (28.58, 50.66)	46.57 (37.98, 59.25)	61.46 (48.69, 78.23)	<0.05
Testosterone	11.03 (8.4, 13.72)	9.28 (6.05, 12.16)	5.03 (1.09, 9.64)	9.8 (1.34, 13.18)	1.17 (0.76, 2.48)	<0.05

(Table 3 continues on next page)

(Continued from previous page)

ADDITIONAL INVESTIGATIVE VARIABLES:						
Demographics:						
Age at recruitment	63 (59,67)	62 (58,66)	64 (60,67)	63 (60.5,67)	63 (60,67)	<0.05
Sex [Male]	16,842 (87.1%)	3,535 (78.9%)	2,216 (48.6%)	30 (63.8%)	661 (8%)	<0.05
Waist circumference	98 (91,106)	102 (94,111)	100 (91,110)	100 (91.75,105.5)	85 (77,93)	<0.05
Hip circumference	104 (100,110)	107 (101,113)	107 (101,116)	106 (101.75,113)	103 (97,109)	<0.05
Standing height	175 (169,180)	174 (168,180)	168 (161,175)	173 (163.25,180)	164 (159,169)	<0.05
Weight	86.2 (77.2,96.7)	90.2 (80.3,102.5)	85.4 (74.4,98.8)	86.5 (73.2,95.3)	70.9 (63.4,80.33)	<0.05
BMI	28.15 (27.03,29.85)	29.79 (28.45,31.64)	30.26 (28.7,32.26)	28.9 (27.47,29.41)	26.36 (25.08,28.12)	<0.05
Activity level:						
Summed minutes activity	100 (50,180)	90 (40,160)	80 (30,150)	120 (62.5,180)	105 (55,180)	<0.05
MET minutes/week for vigorous activity	160 (0,960)	0 (0,720)	0 (0,480)	320 (0,960)	120 (0,720)	<0.05
Blood pressure:						
Diastolic BP	83 (76,91)	84 (77,92)	81 (74,89)	81.5 (73,87)	80 (73,88)	<0.05
Systolic BP	144 (131,157)	145 (133,160)	143 (130,157)	145 (124.75,151.75)	142 (128,156)	<0.05
Pulse rate	67 (59,76)	70 (61,80.25)	71 (63,81)	69 (63.75,76.25)	68 (61,76)	<0.05
Respiratory measures:						
(FEV1)	2.99 (2.42,3.49)	2.85 (2.26,3.39)	2.28 (1.84,2.77)	2.71 (2.19,3.18)	2.27 (1.93,2.64)	<0.05
PEF	433 (334,520)	414 (313,507.75)	332 (258,415)	366 (304.5,469.5)	318 (260,375)	<0.05
FEV1 Z-score	0.57 (-0.18,1.33)	0.77 (0.07,1.53)	0.97 (0.22,1.73)	0.72 (0.08,1.08)	0.5 (-0.22,1.2)	<0.05
FEV1/FVC ratio Z-score	0.36 (-0.17,0.98)	0.29 (-0.22,0.95)	0.43 (-0.12,1.08)	0.45 (-0.28,0.95)	0.51 (0.01,1.02)	<0.05
Alcohol intake frequency:						
Daily or almost daily	4,196 (21.7%)	1,071 (23.9%)	624 (13.7%)	15 (31.9%)	1,264 (15.3%)	<0.05
3 or 4 times a week	3,761 (19.5%)	794 (17.7%)	580 (12.7%)	5 (10.6%)	1,277 (15.5%)	<0.05
Once or twice a week	3,665 (19%)	801 (17.9%)	822 (18%)	7 (14.9%)	1,574 (19.1%)	0.363
1 to 3 times a month	1,241 (6.4%)	299 (6.7%)	409 (9%)	5 (10.6%)	780 (9.5%)	<0.05
Special occasions only	1,404 (7.3%)	329 (7.3%)	640 (14%)	4 (8.5%)	977 (11.8%)	<0.05
Never	1,172 (6.1%)	311 (6.9%)	532 (11.7%)	4 (8.5%)	715 (8.7%)	<0.05
Ethnic background:						
White	18,578 (96.1%)	4,445 (99.2%)	4,264 (93.5%)	46 (97.9%)	8,203 (99.4%)	<0.05
Asian or Asian British	157 (0.8%)	2 (0%)	244 (5.4%)	1 (2.1%)	2 (0%)	<0.05
Black or Black British	243 (1.3%)	1 (0%)	2 (0%)	0 (0%)	1 (0%)	<0.05
Mixed	72 (0.4%)	9 (0.2%)	15 (0.3%)	0 (0%)	15 (0.2%)	0.0641
Other ethnic group	135 (0.7%)	5 (0.1%)	12 (0.3%)	0 (0%)	8 (0.1%)	<0.05
Chinese	36 (0.2%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	<0.05

(Table 3 continues on next page)

(Continued from previous page)

AF and flutter diagnosis (main/secondary):						
ICD10 - AF and flutter	11,302 (58.5%)	2,684 (59.9%)	2,740 (60.1%)	27 (57.5%)	4,213 (51%)	<0.05
ICD10 - Paroxysmal AF	3,209 (16.6%)	696 (15.5%)	811 (17.8%)	7 (14.9%)	1,835 (22.2%)	<0.05
ICD10 - Persistent AF	740 (3.8%)	149 (3.3%)	106 (2.3%)	1 (2.1%)	278 (3.4%)	<0.05
ICD10 - Chronic AF	327 (1.7%)	80 (1.8%)	63 (1.4%)	0 (0%)	100 (1.2%)	<0.05
ICD10 - Typical AF	128 (0.7%)	31 (0.7%)	18 (0.4%)	0 (0%)	39 (0.5%)	0.1045
ICD10 - Atypical atrial flutter	43 (0.2%)	13 (0.3%)	13 (0.3%)	0 (0%)	17 (0.2%)	0.8072
ICD10 - AF and atrial flutter, unspecified	11,455 (59.2%)	2,723 (60.8%)	2,678 (58.7%)	24 (51.1%)	4,887 (59.2%)	0.6494
Systems (phecode categories):						
Endocrine/metabolic	4,467 (23.1%)	1,865 (41.6%)	1,947 (42.7%)	12 (25.5%)	1,828 (22.1%)	<0.05
Circulatory system	14,062 (72.7%)	3,559 (79.4%)	3,783 (82.9%)	35 (74.5%)	5,189 (62.8%)	<0.05
Respiratory	2,991 (15.5%)	804 (18%)	1,200 (26.3%)	9 (19.2%)	1,093 (13.2%)	<0.05
Diabetes:						
Type 1 diabetes	300 (1.6%)	258 (5.8%)	225 (4.9%)	0 (0%)	56 (0.7%)	<0.05
Type 1 diabetes with ketoacidosis	18 (0.1%)	40 (0.9%)	14 (0.3%)	0 (0%)	9 (0.1%)	<0.05
Type 1 diabetes with renal manifestations	16 (0.1%)	13 (0.3%)	29 (0.6%)	0 (0%)	2 (0%)	<0.05
Type 1 diabetes with ophthalmic manifestations	58 (0.3%)	61 (1.4%)	41 (0.9%)	0 (0%)	15 (0.2%)	<0.05
Type 1 diabetes with neurological manifestations	26 (0.1%)	36 (0.8%)	29 (0.6%)	0 (0%)	5 (0.1%)	<0.05
Diabetes type 1 with peripheral circulatory disorders	13 (0.1%)	13 (0.3%)	23 (0.5%)	0 (0%)	3 (0%)	<0.05
Type 2 diabetes	3,400 (17.6%)	1,620 (36.2%)	1,462 (32.1%)	9 (19.2%)	639 (7.7%)	<0.05
Type 2 diabetes with ketoacidosis	35 (0.2%)	41 (0.9%)	14 (0.3%)	0 (0%)	6 (0.1%)	<0.05
Type 2 diabetes with renal manifestations	66 (0.3%)	55 (1.2%)	103 (2.3%)	1 (2.1%)	8 (0.1%)	<0.05
Type 2 diabetes with ophthalmic manifestations	326 (1.7%)	244 (5.5%)	226 (5%)	3 (6.4%)	53 (0.6%)	<0.05
Type 2 diabetes with neurological manifestations	132 (0.7%)	139 (3.1%)	137 (3%)	2 (4.3%)	17 (0.2%)	<0.05
Diabetes type 2 with peripheral circulatory disorders	122 (0.6%)	109 (2.4%)	110 (2.4%)	0 (0%)	10 (0.1%)	<0.05
Hypertension:						
Essential hypertension	12,827 (66.3%)	3,334 (74.4%)	3,571 (78.3%)	31 (66%)	4,679 (56.7%)	<0.05
Other hypertensive complications	34 (0.2%)	5 (0.1%)	42 (0.9%)	0 (0%)	5 (0.1%)	<0.05

(Table 3 continues on next page)

(Continued from previous page)

Cardiovascular disease:						
Myocardial infarction	3,684 (19.1%)	972 (21.7%)	1,027 (22.5%)	11 (23.4%)	850 (10.3%)	<0.05
Other forms of chronic heart disease	2 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0.7735
Congestive heart failure (CHF) NOS	1,891 (9.8%)	539 (12%)	727 (15.9%)	2 (4.3%)	601 (7.3%)	<0.05
Chronic pulmonary heart disease	500 (2.6%)	168 (3.8%)	209 (4.6%)	0 (0%)	228 (2.8%)	<0.05
Heart failure NOS	2,452 (12.7%)	662 (14.8%)	834 (18.3%)	8 (17%)	724 (8.8%)	<0.05
Coronary atherosclerosis	92 (0.5%)	28 (0.6%)	30 (0.7%)	1 (2.1%)	12 (0.2%)	<0.05
Peripheral vascular disease:						
Peripheral vascular disease, unspecified	934 (4.8%)	316 (7.1%)	408 (9%)	2 (4.3%)	251 (3%)	<0.05
Other specified peripheral vascular diseases	8 (0%)	4 (0.1%)	6 (0.1%)	0 (0%)	5 (0.1%)	0.2495
Pulmonary hypertension:						
Primary pulmonary hypertension	193 (1%)	53 (1.2%)	84 (1.8%)	0 (0%)	73 (0.9%)	<0.05
Stroke:						
Hemiplegia	598 (3.1%)	167 (3.7%)	208 (4.6%)	1 (2.1%)	240 (2.9%)	<0.05
Liver disease:						
Liver abscess and sequelae of chronic liver disease	165 (0.9%)	117 (2.6%)	60 (1.3%)	0 (0%)	31 (0.4%)	<0.05
Alcoholic liver damage	155 (0.8%)	147 (3.3%)	65 (1.4%)	0 (0%)	12 (0.2%)	<0.05
Other chronic non-alcoholic liver disease	654 (3.4%)	300 (6.7%)	268 (5.9%)	1 (2.1%)	218 (2.6%)	<0.05
Other disorders of the liver	387 (2%)	135 (3%)	125 (2.7%)	0 (0%)	161 (2%)	<0.05
Kidney disease:						
End-stage renal disease	155 (0.8%)	54 (1.2%)	247 (5.4%)	0 (0%)	28 (0.3%)	<0.05

As in Table 1, medians and interquartile ranges were calculated for continuous variables, and frequencies and proportions (as percentages) were calculated for the categorical variables. Shades of red/blue were used per variable to illustrate differences between lower and higher values. Red shades were used for the modelling variables, whilst blue was used for the additional investigative variables.

Table 3: **Characteristics of the participants per phenotype of AF in the general population using UK-Biobank data.**

in the original high-dimensional space. Indeed, GTM has been applied in diverse real-world situations spanning various domains such as bioinformatics[48,49]; in the financial sector[50]; and more recently also in modelling freedom of expression.[51] To the best of our knowledge, GTM has not been used before to study AF or to generate clinical phenotypes.

The identification and characterisation of clinical phenotypes of AF across diverse patient populations show potential for personalised risk assessment and prognosis. Leveraging these phenotypes could facilitate the tailoring of prevention and treatment programs specific to each phenotype.

The proposed methodology provides several advantages to extract meaningful phenotypes. First, as opposed to previous approaches,[1,9,10,12,17] we define phenotypes based on a non-linear clustering approach which can capture more complex relationships. Furthermore, we can visualise the clusters, and by extension the phenotypes, and how each variable affects each cluster, which provides interpretability, crucial for validation and understanding. It also allows for a convenient method of looking at phenotype differences. For example, phenotype 2 in Fig. 4(b) occupies predominantly the right side of the membership map. The reference vector for glucose in Fig. 2 (top) highlights that participants in the bottom right micro-clusters have the highest glucose values when compared to the other micro-clusters. This information can be translated back to phenotype 2 to provide more context about its

Variable name	Phenotype 1 (n = 1,705)	Phenotype 2 (n = 631)	Phenotype 3 (n = 250)	Phenotype 4 (n = 109)	p-value
MODELLING VARIABLES:					
Diabetes:					
Glucose	139 (115.22, 184.71)	136.15 (118.3, 160.6)	127.94 (119.24, 137.89)	134.04 (114.84, 159.54)	< 0.05
Bone profile:					
Phosphate	4.57 (3.65, 5.65)	3.5 (3, 4.1)	3.38 (2.99, 3.79)	3.36 (2.8, 3.86)	< 0.05
Oxygenation:					
Oxygen saturation	96.08 (93.88, 97.75)	96.22 (94.67, 97.65)	96.36 (93.66, 97.85)	97.03 (95.37, 98.4)	< 0.05
Respiratory rate	19.25 (16.9, 22.32)	20.5 (17.97, 23.09)	16.98 (15.7, 18.62)	18.46 (16.5, 20.63)	< 0.05
FiO2	57.5 (50, 66.27)	56.07 (50, 62.16)	58.33 (52.08, 64.58)	53.57 (46.15, 57.54)	< 0.05
PEEP	6.45 (5.08, 8.11)	6.37 (5.1, 7.68)	5.05 (5, 5.94)	5.38 (5, 6.24)	< 0.05
Partial pressure of oxygen	109 (72.0, 150.97)	114.79 (85.64, 139.45)	168.96 (143.3, 205.26)	133.93 (111.1, 152.19)	< 0.05
Haemoglobin	9.62 (8.58, 10.79)	10.5 (9.14, 11.91)	9.92 (9.23, 10.79)	11.81 (10.4, 13.2)	< 0.05
Respiratory/metabolic markers:					
pH	7.29 (7.17, 7.36)	7.32 (7.15, 7.38)	7.37 (7.35, 7.4)	7.22 (7.08, 7.38)	< 0.05
Anion Gap	17 (14.0, 20.21)	13.83 (12, 15.97)	11.67 (10, 13.08)	14 (12.16, 15.94)	< 0.05
Lactate	2.33 (1.6, 3.39)	1.9 (1.4, 2.62)	2.14 (1.62, 2.78)	1.6 (1.16, 2.12)	< 0.05
Cardiac markers:					
Heart rate	83.39 (73.36, 93.75)	85.03 (76.83, 96.06)	80.46 (75.33, 85.81)	75.86 (68.26, 86.2)	< 0.05
Capillary refill	0.03 (0, 0.42)	0 (0, 0.02)	0 (0, 0)	0 (0, 0)	< 0.05
Diastolic BP	56 (50.34, 61.62)	58.21 (52.49, 63.98)	55.19 (50.21, 59.87)	65.62 (59, 72.69)	< 0.05
Systolic BP	109.24 (101.74, 119.08)	110.19 (103.3, 118.4)	111.01 (105.38, 117.38)	131.09 (121.45, 143.2)	< 0.05
Clotting markers:					
Prothrombin time	16.53 (13.95, 22.29)	14.65 (13.02, 16.7)	14.2 (13.2, 15.37)	13.1 (12.17, 14.3)	< 0.05
Platelet count	148.42 (102.73, 223.19)	187.79 (139.22, 254.14)	146.29 (120.05, 185.56)	197 (151.08, 245.71)	< 0.05
Renal function:					
Creatinine	2.12 (1.3, 3.7)	1 (0.75, 1.33)	0.9 (0.73, 1.16)	0.9 (0.7, 1.2)	< 0.05
Electrolytes:					
Magnesium	2.11 (1.91, 2.4)	2 (1.8, 2.25)	2.4 (2.19, 2.7)	2 (1.8, 2.13)	< 0.05
Potassium	4.49 (4.05, 4.92)	4 (3.83, 4.55)	4.33 (4.11, 4.57)	4.05 (3.74, 4.33)	< 0.05
Other:					
Temperature	57.5 (36.45, 36.97)	56.07 (36.62, 37.11)	58.33 (36.52, 36.85)	53.57 (36.67, 37.24)	< 0.05

(Table 4 continues on next page)

(Continued from previous page)

ADDITIONAL INVESTIGATIVE VARIABLES:					
Demographics:					
Age	71.0 (63.0, 81.0)	73.0 (64.0, 82.0)	74.0 (67.0, 80.0)	75.0 (65.75, 84.0)	< 0.05
Sex	405 (63.4%)	453 (57.2%)	563 (62.3%)	206 (57.2%)	0.3317
Height	172.86 (162.78, 177.9)	170.09 (162.72, 177.9)	170.09 (162.78, 177.9)	172.86 (162.78, 180.17)	0.2896
Weight	83.93 (69.84, 98.29)	81.42 (65.9, 99.36)	83.05 (70.33, 95.92)	79.79 (65.85, 95.97)	0.0768
Ethnicity:					
White	434.0 (67.9%)	589.0 (74.4%)	700.0 (77.4%)	248.0 (68.9%)	0.1263
Other ethnic group	117.0 (18.3%)	128.0 (16.2%)	136.0 (15.0%)	72.0 (20.0%)	0.1785
Black	51.0 (8.0%)	43.0 (5.4%)	23.0 (2.5%)	21.0 (5.8%)	< 0.05
Hispanic	16.0 (2.5%)	16.0 (2.0%)	28.0 (3.1%)	8.0 (2.2%)	0.5508
Asian	21.0 (3.3%)	16.0 (2.0%)	17.0 (1.9%)	11.0 (3.1%)	0.2400
Glasgow Coma Scale (GCS):					
GCS eye-opening	2.83 (1.75, 3.83)	3.29 (2.29, 4.0)	2.5 (1.67, 3.08)	3.29 (2.34, 4.0)	< 0.05
GCS motor response	5.0 (3.06, 6.0)	5.67 (4.28, 6.0)	4.12 (2.79, 4.75)	5.79 (4.67, 6.0)	< 0.05
GCS verbal response	2.04 (1.0, 4.62)	3.33 (1.0, 5.0)	2.25 (1.0, 3.5)	3.25 (1.0, 5.0)	< 0.05
Ventilation:					
Non-Invasive ventilation	56.0 (8.8%)	75.0 (9.5%)	54.0 (6.0%)	24.0 (6.7%)	< 0.05
Invasive ventilation	485.0 (75.9%)	557.0 (70.3%)	852.0 (94.2%)	222.0 (61.7%)	< 0.05
Outcomes:					
Time to AF diagnosis (hours)	59.0 (41.0, 94.0)	52.0 (36.0, 91.0)	49.0 (37.0, 70.0)	55.0 (36.75, 89.0)	< 0.05
In-hospital length of stay (hours)	296.32 (180.18, 498.3)	262.41 (169.22, 427.41)	228.08 (159.62, 340.88)	246.62 (161.07, 413.97)	< 0.05
In-ICU length of stay (hours)	143.89 (82.93, 264.94)	112.97 (70.99, 211.78)	98.33 (69.63, 148.69)	110.16 (69.28, 212.26)	< 0.05
Death after ICU (hours)	26.57 (16.46, 1021.61)	183.27 (17.17, 2350.5)	1558.35 (21.98, 10015.99)	394.49 (18.8, 3513.18)	< 0.05
Death after hospital discharge (hours)	17.5 (8.5, 849.0)	20.25 (10.1, 2106.35)	1330.07 (16.3, 9930.92)	27.81 (12.62, 3271.6)	< 0.05
Death after hospital discharge (days)	0.73 (0.35, 35.38)	0.84 (0.42, 87.76)	55.42 (0.68, 413.79)	1.16 (0.53, 136.32)	< 0.05
In-hospital mortality	245.0 (38.3%)	191.0 (24.1%)	60.0 (6.6%)	71.0 (19.7%)	< 0.05
In-ICU length of stay of 3+ days	526.0 (82.3%)	587.0 (74.1%)	665.0 (73.6%)	262.0 (72.8%)	0.1785
In-ICU length of stay of 7+ days	274.0 (42.9%)	257.0 (32.4%)	186.0 (20.6%)	123.0 (34.2%)	< 0.05
Mortality after hospital discharge within 30 days	301.0 (47.1%)	245.0 (30.9%)	77.0 (8.5%)	88.0 (24.4%)	< 0.05
Mortality after hospital discharge Within 365 days	368.0 (57.6%)	325.0 (41.0%)	121.0 (13.4%)	122.0 (33.9%)	< 0.05
Mortality after hospital discharge after 365 days	36.0 (5.6%)	49.0 (6.2%)	43.0 (4.8%)	24.0 (6.7%)	0.5042
AKI	161.0 (25.2%)	159.0 (20.1%)	184.0 (20.4%)	41.0 (11.4%)	< 0.05
ARDS	33.0 (5.2%)	58.0 (7.3%)	37.0 (4.1%)	46.0 (12.8%)	< 0.05

As in Table 2, medians and interquartile ranges were calculated for continuous variables, and frequencies and proportions (as percentages) were calculated for the categorical variables. As in Table 3, shades of red/blue were used per variable to illustrate differences between lower and higher values. Red shades were used for the modelling variables, whilst blue was used for the additional investigative variables.

Table 4: **Characteristics of the participants per phenotype of AF in an ICU population using the MIMIC-IV database.**

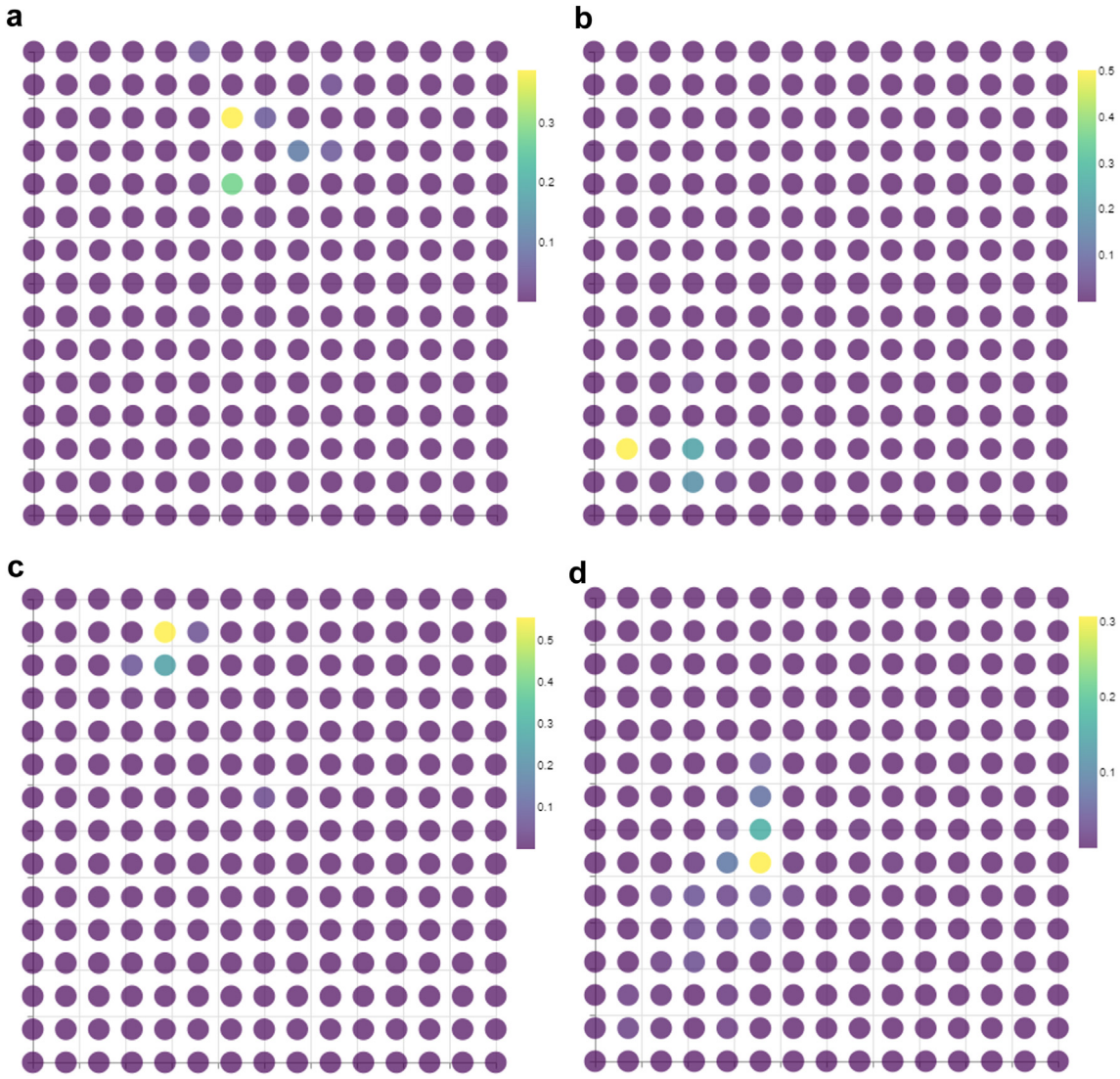


Fig. 6: **Membership map with the probability distributions for different data points superimposed.** Maps a) and b) show the probability distribution for two randomly selected participants from the general population taken from the UK Biobank database. Maps c) and d) show the probability distribution for two randomly selected patients from the critical care population taken from the MIMIC-IV database.

participants, and how risk factors may not be uniformly distributed within a given phenotype.

Comparing the phenotypes of previous studies with those derived from our proposed methodology is not straightforward. Starting with the general population phenotypes generated using the UK Biobank data, the population we analyse (UK) differs from the Japanese,[1,10,16] European,[9,11,17] and North American[9] populations previously analysed. As determinants of AF can greatly differ across geographical locations,[9] this introduces a certain level of expected difference between our results and those already stated. However, one example that stands out is that phenotype 2 (Fig. 4(d)) matches almost identically to cluster 3 identified as part the study conducted by Vitolo et al.,[9] which groups the youngest participants/patients who are likely to be male with high burden of cardiovascular comorbidities and risk factors, along with the highest rates of chronic (permanent) AF. We also see other similarities however they are not fully homogeneous, for example comparing phenotype 3 again in Fig. 4(d) with cluster 2 outlined in the study by Bisson et al.[17] They both group together the oldest patients/participants with a high prevalence of cardiac conditions, however they differ in that phenotype 3 is split between Male and Female, whereas cluster 2 defined in Bisson et al. is mostly male with almost

exclusively permanent AF. What this does indicate is that our approach is able to capture the key relationships between patients with AF and find population-specific relationships that allow the phenotypes to be more representative. The phenotypes generated for the critical care population in our study will be inherently different to the general population, which means a comparison with those developed in the literature would not be appropriate.

Another key difference lies in the selection of modelling variables. The phenotypes for both data cohorts were generated using only vitals and laboratory test data, as opposed to previous studies that also included demographics and medical history/comorbidity information in the modelling. This results in their stated phenotypes having significant differences for such variables as they were used to initially stratify the data. The phenotypes generated in our study show significant differences with these key risk factors, but without including explicit information on these variables during modelling. Additionally, as the between-phenotype differences for variables such as demographics and comorbidities are performed post-hoc, should new data become available from variables not yet examined, their distribution between and within each phenotype can be swiftly identified.

From a clinical perspective, the availability of reliable and robust phenotypes could be a major asset to their assortment of diagnostic tools. Phenotypes provide a different way of visualising a targeted population, which for context of this study is patients with AF. Many of these patients have multiple comorbidities, and management based on a single comorbidity in a binary (yes/no) matter is inappropriate, as many comorbidities tend to cluster leading to clinically complex phenotypes. While clustering can be performed using biostatistical approaches, our proposed methodology using GTM provides a more principled approach to clustering, with the capacity to elucidate more specific patient profiles. This would result in more robust patient stratification, as well as the tailoring of prevention and treatment programs specific to each phenotype.

One of the limitations of this study relates to the genomic principal components used for the UK-Biobank cohort, as their loadings were not available, limiting the ability to interpret them. Another limitation is related to the transferability of the derived phenotypic clusters to other cohorts of data, as they could vary across diverse populations due to genetic, environmental, and cultural differences. Additionally, differences in clinical settings, such as healthcare access, diagnostic criteria, and treatment approaches, may contribute to distinct phenotypic patterns among various patient groups. Since this study's main objective is to present a robust AI methodology for the derivation of AF phenotypes, this limitation can be mitigated by the derivation of specific phenotypes for different patient cohorts, as and when required. The dynamic nature of risk is also another possible limitation, as the current approach does not address how phenotypes change over time.

Our study proposed an AI-based approach for the derivation of clinically meaningful AF phenotypes. We applied it to two large cohort databases representing general and critical care populations. Our approach is probabilistic, contributing to robust patient stratification. It produces interpretable visualisation of complex high-dimensional data, enhancing understanding. It showed its ability to identify clinical phenotypes of AF, which could enable prevention and treatment programs specific to each phenotype. Our methodology can be applied to other datasets to derive clinically meaningful phenotypes of other conditions.

Contributors

S.O.M. conceptualised the methodological approach and led the study. S.O.M., I.O., G.Y.H.L., R.L., I.J., A.M.T, and E.A.D. secured the funding. R.A.A.B., I.O. and S.O.M. extracted the data. R.L., I.J. and G.Y.H.L. advised on the selection of clinically relevant variables. R.A.A.B. implemented the code and ran the experiments. S.O.M. and I.O. supervised the study. E.A.D., A.M.T. and G.L. contributed to the discussions. R.A.A.B., S.O.M. and I.O. drafted the early version of the manuscript. All authors reviewed and edited the final manuscript. All authors have read and agreed to the published version of the manuscript.

Data sharing statement

The UK Biobank database is available for approved projects only (application process detailed at https://www.ukbiobank.ac.uk/enable-your-research/apply-for-access) through the UK Biobank Access Management System (https://www.ukbiobank.ac.uk). The MIMIC-IV database is available on the PhysioNet portal (https://physionet.org/content/mimiciv/2.2/) for credentialed users only.

Declaration of interests

All authors declare no competing interests.

Acknowledgements

R.A.A.B was supported by the DECIPHER project (LJMU QR-PSF). S.O.M, I.O., G.Y.H.L, R.L., I.J., E.A.D., and A.M.T. were supported by the EU project TARGET, which has received funding from the EU HORIZON EUROPE framework programme for research and innovation under the Grant Agreement No. 101136244.

Appendix A. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.ebiom.2024.105280.

References

1 Saito Y, Omae Y, Nagashima K, et al. Phenotyping of atrial fibrillation with cluster analysis and external validation. *Heart*. 2023;109:1751–1758.
2 Zhang J, Johnsen SP, Guo Y, Lip GYH. Epidemiology of atrial fibrillation. *Card Electrophysiol Clin*. 2021;13(1):1–23.
3 Lip GYH, Genaidy A, Tran G, Marroquin P, Estes C, Sloop S. Improving stroke risk prediction in the general population: a comparative assessment of common clinical rules, a new multimorbid index, and machine-learning-based algorithms. *Thromb Haemost*. 2021;122(1):142–150.
4 Romiti GF, Proietti M, Bonini N, et al. Clinical complexity domains, anticoagulation, and outcomes in patients with atrial fibrillation: a report from the GLORIA-AF registry phase II and III. *Thromb Haemost*. 2022;122(12):2030–2041.

5 Olier I, Ortega-Martorell S, Pieroni M, Lip GYH. How machine learning is impacting research in atrial fibrillation: implications for risk prediction and future management. *Cardiovasc Res*. 2021;117(7):1700–1717.
6 Chung KF, Adcock IM. How variability in clinical phenotypes should guide research into disease mechanisms in asthma. *Ann Am Thorac Soc*. 2013;10:S109–S117.
7 Romiti GF, Pastori D, Rivera-Caravaca JM, et al. Adherence to the 'atrial fibrillation better care' pathway in patients with atrial fibrillation: impact on clinical outcomes-A systematic review and meta-analysis of 285,000 patients. *Thromb Haemost*. 2022;122(3):406–414 [cited 2022 Nov 13];Available from: https://pubmed.ncbi.nlm.nih.gov/34020488/.
8 Chao TF, Joung B, Takahashi Y, et al. 2021 focused update consensus guidelines of the asia pacific heart rhythm society on stroke prevention in atrial fibrillation: executive summary. *Thromb Haemost*. 2022;122(1):20–47 [cited 2022 Nov 13];Available from: https://pubmed.ncbi.nlm.nih.gov/34773920/.
9 Vitolo M, Proietti M, Shantsila A, Boriani G, Lip GYH. Clinical phenotype classification of atrial fibrillation patients using cluster analysis and associations with trial-adjudicated outcomes. *Biomedicines*. 2021;9(7):1–11.
10 Watanabe E, Inoue H, Atarashi H, et al. Clinical phenotypes of patients with non-valvular atrial fibrillation as defined by a cluster analysis: a report from the J-RHYTHM registry. *Int J Cardiol Heart Vasc*. 2021;37:100885.
11 Proietti M, Vitolo M, Harrison SL, et al. Impact of clinical phenotypes on management and outcomes in European atrial fibrillation patients: a report from the ESC-EHRA EURObservational Research Programme in AF (EORP-AF) General Long-Term Registry. *BMC Med*. 2021;19(1):256.
12 Inohara T, Piccini JP, Mahaffey KW, et al. A cluster analysis of the Japanese multicenter outpatient registry of patients with atrial fibrillation. *Am J Cardiol*. 2019;124(6):871–878.
13 Tobin J, Zhang M. Clustering of big data with mixed features. *arXiv*. 2020. preprint; arXiv:2011.06043.
14 Ezugwu AE, Shukla AK, Agbaje MB, Oyelade ON, José-García A, Agushaka JO. Automatic clustering algorithms: a systematic review and bibliometric analysis of relevant literature. *Neural Comput Appl*. 2021;33(11):6247–6306.
15 Rodriguez MZ, Comin CH, Casanova D, et al. Clustering algorithms: a comparative approach. *PLoS One*. 2019;14(1):e0210236.
16 Ogawa H, An Y, Nishi H, et al. Characteristics and clinical outcomes in atrial fibrillation patients classified using cluster analysis: the Fushimi AF Registry. *Europace*. 2021;23(9):1369–1379.
17 Bisson A, M Fawzy A, Romiti GF, et al. Phenotypes and outcomes in non-anticoagulated patients with atrial fibrillation: an unsupervised cluster analysis. *Arch Cardiovasc Dis*. 2023;116(6–7):342–351.
18 Bishop CM, Svensén M, Williams CKI. GTM: the generative topographic mapping. *Neural Comput*. 1998;10(1):215–234.
19 Olier I, Vellido A. Advances in clustering and visualization of time series using GTM through time. *Neural Network*. 2008;21(7):904–913.
20 Kohonen T. *Self-organizing maps*. Berlin, Heidelberg: Springer Berlin Heidelberg; 2001 (Springer Series in Information Sciences; vol. vol. 30).
21 Bishop C, Svensén M, Williams C. GTM: a principled alternative to the self-organizing map. In: Mozer MC, Jordan M, Petsche T, eds. *Advances in neural information processing systems*. MIT Press; 1996.
22 Van Der Maaten L, Hinton G. Visualizing data using t-SNE. *J Mach Learn Res*. 2008;9:2579–2625.
23 McInnes L, Healy J, Saul N, Großberger L. UMAP: uniform manifold approximation and projection. *J Open Source Softw*. 2018;3(29):861.
24 Vellido A, Lisboa PJG, Meehan K. The generative topographic mapping as a principal model for data visualization and market segmentation: an electronic commerce case study. *Int J Comput Syst Signals*. 2000;1:119–138.
25 Ward Jr JH. Hierarchical grouping to optimize an objective function. *J Am Stat Assoc*. 1963;58(301):236–244.
26 Bellfield RAA, Ortega-Martorell S, Olier I. *Code: AI-based derivation of atrial fibrillation phenotypes in the general and critical care populations*. Zenodo; 2024.
27 Sudlow C, Gallacher J, Allen N, et al. UK Biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med*. 2015;12(3):1–10.
28 Bycroft C, Freeman C, Petkova D, et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature*. 2018;562(7726):203–209.
29 Papadopoulou A, Harding D, Slabaugh G, Marouli E, Deloukas P. Prediction of atrial fibrillation and stroke using machine learning models in UK Biobank. *Heliyon*. 2024;10(7):e28034. https://doi.org/10.1016/j.heliyon.2024.e28034.
30 Johnson AEW, Bulgarelli L, Shen L, et al. MIMIC-IV, a freely accessible electronic health record dataset. *Sci Data*. 2023;10(1):1.
31 Ortega-Martorell S, Olier I, Johnston BW, Welters ID. Sepsis-induced coagulopathy is associated with new episodes of atrial fibrillation in patients admitted to critical care in sinus rhythm. *Front Med*. 2023;15:10.
32 Ortega-Martorell S, Pieroni M, Johnston BW, Olier I, Welters ID. Development of a risk prediction model for new episodes of atrial fibrillation in medical-surgical critically ill patients using the AmsterdamUMCdb. *Front Cardiovasc Med*. 2022;9:897709.
33 Allan V, Honarbakhsh S, Casas JP, et al. Are cardiovascular risk factors also associated with the incidence of atrial fibrillation?: a systematic review and field synopsis of 23 factors in 32 population-based cohorts of 20 million participants. *Thromb Haemostasis*. 2017;117:837–850.
34 Nso N, Bookani KR, Metzl M, Radparvar F. Role of inflammation in atrial fibrillation: a comprehensive review of current knowledge. *J Arrhythm*. 2021;37(1):1–10.
35 Mohanty S, Hall A, Mohanty P, et al. Being asymptomatic with atrial fibrillation: is it a genetic trait? *J Am Coll Cardiol*. 2016;67(13):677.
36 Kalarus Z, Mairesse GH, Sokal A, et al. Searching for atrial fibrillation: looking harder, looking longer, and in increasingly sophisticated ways. In: *An EHRA position paper*. *Vol. 25*. Europace; 2023.
37 Alonso A, Krijthe BP, Aspelund T, et al. Simple risk model predicts incidence of atrial fibrillation in a racially and geographically diverse population: the CHARGE-AF consortium. *J Am Heart Assoc*. 2013;2(2):e000102.
38 Lip GYH, Skjøth F, Nielsen PB, Larsen TB. Evaluation of the C2HEST risk score as a possible opportunistic screening tool for incident atrial fibrillation in a healthy population (from a nationwide Danish cohort study). *Am J Cardiol*. 2020;125(1):48–54.
39 Johnston BW, Chean CS, Duarte R, et al. Management of new onset atrial fibrillation in critically unwell adult patients: a systematic review and narrative synthesis. *Br J Anaesth*. 2022;128(5):759–771.
40 Shen Chean C, Mcauley D, Gordon A, Welters ID. Current practice in the management of new-onset atrial fibrillation in critically ill patients: a UK-wide survey. *PeerJ*. 2017;5:e3716. https://doi.org/10.7717/peerj.3716.
41 Joseph O'bryan L, Redfern OC, Bedford J, et al. Managing new-onset atrial fibrillation in critically ill patients: a systematic narrative review [cited 2024 Apr 29]; Available from: http://bmjopen.bmj.com/.
42 Bosch NA, Cimini J, Walkey AJ. Atrial fibrillation in the ICU. *Chest*. 2018;154(6):1424–1434.
43 O'Driscoll BR, Smith R. Oxygen use in critical illness. *Respir Care*. 2019;64(10):1293–1307.
44 Wu P, Gifford A, Meng X, et al. Mapping ICD-10 and ICD-10-CM Codes to phecodes: workflow development and initial evaluation. *JMIR Med Inform*. 2019;7(4):1–13.
45 Handelsman DJ, Hirschberg AL, Bermon S. Circulating testosterone as the hormonal basis of sex differences in athletic performance. *Endocr Rev*. 2018;39(5):803–829.
46 Qu X, Donnelly R. Sex hormone-binding globulin (Shbg) as an early biomarker and therapeutic target in polycystic ovary syndrome. *Int J Mol Sci*. 2020;21(21):1–17.
47 Chatterjee P, Cymberknop LJ, Armentano RL. Nonlinear systems in healthcare towards intelligent disease prediction. In: *Nonlinear Systems -Theoretical Aspects and Recent Applications*. 2019.
48 Gaspar HA, Hübel C, Breen G. Biological pathways and drug gene-sets: analysis and visualization. *Eur Neuropsychopharmacol*. 2019;29:S834.
49 Olier I, Amengual J, Vellido A. A variational Bayesian approach for the robust analysis of the cortical silent period from EMG recordings of brain stroke patients. *Neurocomputing*. 2011;74(9):1301–1314.
50 Feng J, Liu Z, Feng L. Identifying opportunities for sustainable business models in manufacturing: application of patent analysis and generative topographic mapping. *Sustain Prod Consum*. 2021;27:509–522.
51 Ortega-Martorell S, Bellfield RAA, Harrison S, Dyke D, Williams N, Olier I. Mapping the global free expression landscape using machine learning. *SN Appl Sci*. 2023;5(12):354.