Discovering and Validating Disease Subtypes for Heart Failure using Unsupervised Machine Learning Methods

Ghazaleh Fatemifar, R Thomas Lumbers, Daniel I Swerdlow, Spiros Denaxas
Institute of Health Informatics, University College London, London, UK
Farr Institute of Health Informatics Research, London, UK

Introduction

Notable heterogeneity exists in the clinical presentation of heart failure (HF) patients. Current subtype classifications are based on ejection fraction may not fully capture the aetiological and prognostic heterogeneity of HF.

The use of unsupervised machine learning (ML) approaches, such as cluster analysis, on large-scale observational data from electronic health records (EHR), can enable the discovery of novel subtypes and guide the characterization of their clinical manifestation. Clustering methods can group HF patients based on similarities between their clinical features without making a *priori* assumptions about the distribution of the data.

We sought to discover, characterize and replicate HF subtypes by applying a clustering method on a heterogeneous HF population derived from phenotypically rich EHR. Characterization of HF subtypes using EHR derived variable may enable more precise large-scale genomic analysis to inform better **prevention, diagnostic** and **treatment** strategies.

Aims

a. Use clustering methods to identify and characterize HF subtypes using clinical features extracted from phenotypically rich, longitudinal EHR data.
b. Evaluate identified disease subtypes in terms of cardiac-related mortality.

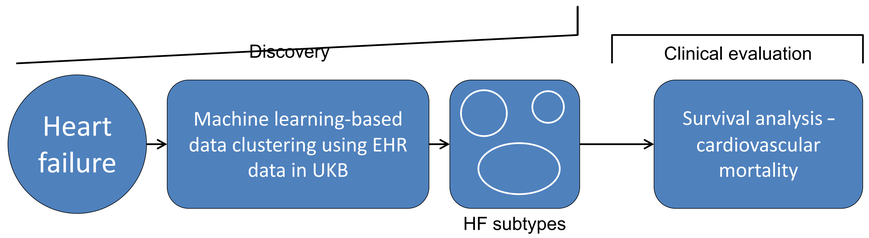

Figure 1. Discovery and validation of HF subtypes using ML methodologies.

Methods

We defined a cohort of HF patients using data from the UK Biobank (UKB). UKB is a cohort (n=500K) of middle aged participants with linked baseline, EHR and genetics data. We identified HF patients using:

a.nurse-validated patients-reported medical history (Non-cancer illness code = 1076)

b.hospital care diagnostic codes (ICD9: 402.01, 402.11, 404.01, 404.03, 404.11, 404.13, 404.91, 404.93, 428 and ICD10: I110, I130, I132, I50)

We created clinical feature vectors composed of demographics (e.g. sex, Townsend score) health behaviours (e.g. alcohol weekly units, smoking) and clinical characteristics (e.g. BMI, lung function, blood pressure) for all HF patients by extracting values at baseline. Principle component analysis (PCA) was used to reduce dimensionality **(Figure 2)**.

Clinical features were clustered using a partitioned method (K-medoids) with Gower's distance (used for mixed data types). The optimal number of clusters was derived using greatest silhouette coefficient. A Cox proportional hazards survival analysis was used to explore the differences between each subtype and cardiac-related mortality.

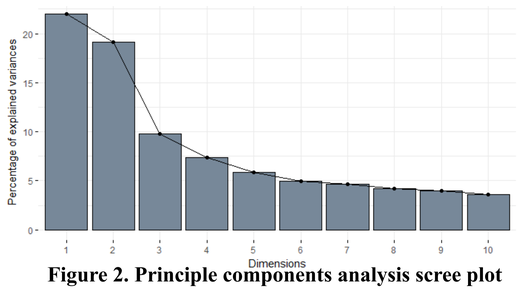

Figure 2. Principle components analysis scree plot

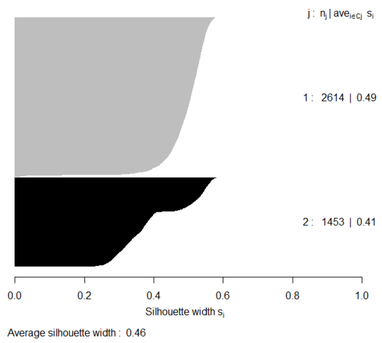

Figure 3: Silhouette coefficient calculated for each data point presented by HF cluster (k=2). Si indicates how close each point in one cluster is to points in another cluster; Si close to 1 indicates the patients are well clustered.

Results

We included **4067** patients in our study: 70% male and a mean age of 62.5 years (SD:6.12). Features that contributed the most to components 1-3 were selected for analysis based on inflection point as shown in **Figure 2**. We selected features based on previous analyses and know n risk factors **(REF). Table 1** highlights some of the features used to cluster patients after PCA.

We identified two patient clusters (**Figure 3**). This was based on silhouette analysis which indicated the optimal number of clusters. Demographic, health behaviours and clinical characteristics according to each cluster are presented in **Table 1**. Cluster features appear to map well to **HFpEF/HFrE**F risk factors **(REF). Cluster 1** (n=2614, 64% male) included patients with higher blood pressure, no/vey low presence of circulatory co-morbidities, more smokers and less deprivation. **Cluster 2 (**n=1453, 81% male) had patients with the highest anthropometric and grip strength measures, as well as the majority of patients with myocardial Infarction, angina and coronary artery disease.

We report a **24% increase in risk** of mortality for patients in cluster 2 when compared to cluster 1 (HR:1.24 95% CI: 1.003-1.54). **Figure 4** illustrates the cumulative hazards for cardiac mortality stratified by HF cluster.

		Cluster 1 (n=2614)	Cluster 2 (n=1453)		Feature extracted post-PCA
	Feature	Mean[SD]	Mean[SD]	P	
Demographic	**Sex = male (%)**	1672 [64]	1179 [81.1]	<0.001	
	Age	62.09 [6.34]	63.25 [5.61]	<0.001	
	Townsend score	-0.86 [3.33]	-0.45 [3.41]	<0.001	
	Smoke (ever) (count / %)	1489 [57]	1007 [69.3]	<0.001	
Biomarkers	**White blood count (109 cells/Litre)**	7.56 [3.05]	7.77 [1.92]	0.018	
Anthropometric	**Weight (kg)**	85.74 [19.02]	88.42 [17.06]	<0.001	
	BMI	29.54 [5.77]	30.26 [5.09]	<0.001	✓
	Waist (cm)	98.47 [15.28]	102 [13.41]	<0.001	✓
	Standing height (cm)	170.13 [9.51]	170.76 [8.81]	0.036	✓
Medical History	**Diastolic blood pressure (mmHg)**	83.19 [11.23]	77.13 [10.95]	<0.001	✓
	Systolic blood pressure (mmHg)	143.8 [20.19]	135.93 [20.52]	<0.001	
	Pulse rate (bpm)	72.9 [14.3]	66.98 [13.92]	<0.001	
	Hand grip strength right (kg)	31.58 [11.3]	32.77 [10.94]	0.001	✓
	Hand grip strength left (kg)	29.44 [11.35]	30.59 [11.32]	0.002	✓
Co-morbid Disease	**Angina (count/%)**	7 [0.27]	914 [62.9]	<0.001	✓
	Diabetes (count/%)	397 [15.19]	411 [28.28]	<0.001	
	Myocardial infarction (count/%)	0 [0]	1101 [75.77]	<0.001	✓
	Coronary artery disease (count/%)	35 [1.34]	1453 [100]	<0.001	✓

Table 1: Patient characteristics according to HF Cluster

Discussion and impact

Using ML, we identified **two** distinct subtypes for HF that differed with respect to cardiac mortality. These results demonstrate that distinct disease subtypes can be identified using unsupervised methods. This approach may facilitate more precise disease definition towards precision medicine approaches to improve patient care.

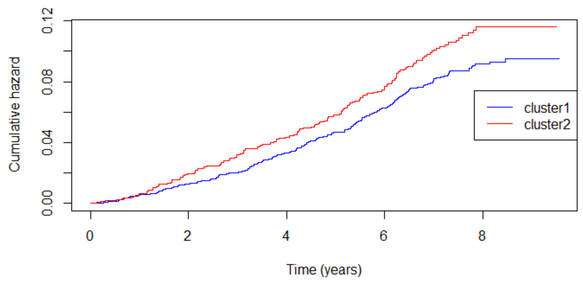

Figure 4: Kaplan-Meier curves for cardiac mortality by HF cluster