International Journal of Data Science and Analytics (2023) 15:49–66
https://doi.org/10.1007/s41060-022-00346-9

REGULAR PAPER

Data-driven versus a domain-led approach to k-means clustering on an open heart failure dataset

A. Jasinska-Piadlo[1,2] · R. Bond[2] · P. Biglarbeigi[2] · R. Brisk[1,2] · P. Campbell[3] · F. Browne[5] · D. McEneaney[1,4]

Received: 4 January 2022 / Accepted: 24 June 2022 / Published online: 25 July 2022

Abstract
Domain-driven data mining of health care data poses unique challenges. The aim of this paper is to explore the advantages and the challenges of a 'domain-led approach' versus a data-driven approach to a k-means clustering experiment. For the purpose of this experiment, clinical experts in heart failure selected variables to be used during the k-means clustering, whilst during the 'data-driven approach' feature selection was performed by applying principal component analysis to the multidimensional dataset. Six out of seven features selected by physicians were amongst 26 features that contributed most to the significant principal components within the k-means algorithm. The data-driven approach showed advantage over the domain-led approach for feature selection by removing the risk of bias that can be introduced by domain experts. Whilst the 'domain-led approach' may potentially prohibit knowledge discovery that can be hidden behind variables not routinely taken into consideration as clinically important features, the domain knowledge played an important role at the interpretation stage of the clustering experiment providing insight into the context and preventing far fetched conclusions. The "data-driven approach" was accurate in identifying clusters with distinct features at the physiological level. To promote the domain-led data mining approach, as a result of this experiment we developed a practical checklist guiding how to enable the integration of the domain knowledge into the data mining project.

Keywords Heart failure · k-means clustering · Domain knowledge · Domain-led data mining · Data science

1 Introduction

In recent years, there has been a growing interest in the application of advanced analytics and machine learning to healthcare data. This growing interest has been sparked by opportunities to analyse anonymised health care data and the advancement in the contemporary hardware and software technology [1]. The benefits of healthcare data analytics cannot be underestimated. The access to electronic health records (EHR), population-based registries, disease registries and data from clinical trials can lead to knowledge discovery once the clinical problem is well-defined and the target dataset is analysed in collaboration with clinical teams. The domain knowledge and clinical experience enables researchers to identify knowledge gaps and formulate clinically important research questions which advanced analytics can address during the data science process. Analogous to other fields, the domain knowledge plays a role with a varying degree at every stage of the data science project [2].

2 Aims and objectives

The aim of this paper is firstly to explore how domain knowledge influences the data mining process during a clustering

✉ A. Jasinska-Piadlo
jasinska_piadlo-a@ulster.ac.uk

1 Southern Health and Social Care Trust, CVD Research Unit, Craigavon Hospital, 68 Lurgan Road, Portadown BT63 5QQ, Northern Ireland

2 Faculty of Computing, Engineering and the Built Environment, Ulster University, Shore Road, Jordanstown BT37 0QB, Northern Ireland

3 Southern Health and Social Care Trust, Cardiology Department, Craigavon Hospital, 68 Lurgan Road, Portadown BT63 5QQ, Northern Ireland

4 Centre for Advanced Cardiovascular Research, Ulster University, Shore Road, Jordanstown BT37 0QB, Northern Ireland

5 AI Department, Datactics, 1 Lanyon Quay, Belfast BT1 3LG, Northern Ireland

experiment. We chose clustering analysis because this is a popular unsupervised ML used previously to define new 'subgroups' of patients with heart failure (HF) [3] [4] [5] [6]. Research has demonstrated that ML can result in improved phenotyping of patients with HF, leading to discovery of new clinical taxonomies [7] and the design of clinical trials testing new treatments combinations that were never used before in specific clusters of patients with HF [8].

Secondly, we will attempt to address the gap which exists between academic data scientists and health practitioners by introducing a framework to enable the integration of domain knowledge into the clustering analysis process. The main motivation to close this gap is to promote effective dialogue between domain experts and theoretical data scientists which will generate mutual gains. The prospect of discovering new patterns that could lead to the development of practical solutions for the medical industry should meet the needs of the domain experts, whereas the opportunity for the publication of the detailed report including data mining process and translation of the experimental work into industry may bring gains to academics. The interaction and effective dialogue between academia and medical industry starts with the mutual respect of the contributions of both parties.

3 Background

Over a decade ago, Cao et al. (2006) [9] identified the differences between data-driven data mining and domain-led data mining process leading to knowledge discovery. The most striking differences in the approaches to data mining were seen between academia and business [10]. According to Cao et al. [10], traditional data-driven data mining is focused on developing innovative approaches with the algorithm at the centre. This approach lets the data create research innovation and perhaps novel algorithms, whilst domain-driven data mining includes humans as a central part of the process and brings solutions to real-world business problems. Following this observation, there was a justified call from the organisers and participants of the 2007 ACM SIKGKDD International Workshop on Domain Driven Data Mining for a paradigm shift from "interesting hidden pattern mining" to "actionable knowledge discovery in varying data mining domains" [11].

A similar pattern of discrepancies in the goals of data mining performed by technologists versus clinicians has been presented by Jasinska et.al. (2021) in the extensive systematic literature of studies using ML on heart failure data sets [12]. There was observed tendency to overclaim the usefulness and applicability of the predictive models to real-world clinical problems, as well as there seem to be an 'unwritten' race in achieving better than previous authors AUC of the designed model. This systematic literature review provided examples of high-quality data science projects co-authored by domain experts and data scientists resulting in sound predictive models and novel clinical phenotypes [8] [13] [14] [15] [16] [17] [18] [19]. There were studies, where authors affiliated purely with information technology engaged with clinicians (domain experts) at various stages of the data mining process [20] [21] [22]. In most of those studies, clinicians were involved either at the stage of data extraction when the candidate features from data sets were being assessed as suitable for consideration in the prediction model or at interpretation of the results stage [22] [23] [19]. For example, Sun et al. (2012) [22] performed a series of interviews with cardiologists to search for clinically meaningful additional risk factors to be used in their predictive model. Saqlain at al. [20] asked cardiac specialists to make sure that chosen features were sufficient to get valuable results for the model and claimed that by that approach they were provided with a deep knowledge of cardiology and it helped them to understand the domain of the problem.

However, one key result of this review showed that a quarter of included papers (22 papers out of 81) were authored exclusively by researchers affiliated with either computer science, IT, business administration, translational research, health informatics, bio-medical informatics, quantitative health sciences, statistics or a related area [12]. Moreover, none of those 22 papers mentioned the incorporation of domain knowledge in the design and execution of the data science project [12]. This is a concerning pattern, illustrating that in the pure data-driven approach to data mining there was very limited (if any) integration of the domain knowledge into the data science process.

Undoubtedly, in the era of BigData and digital transformation, more than ever before, there is a need for the paradigm shift from the data-driven data mining to domain-driven knowledge discovery, particularly in the field of healthcare data.

4 Structure

In this paper, we explore a domain-led approach and a data-driven approach to performing k-means clustering. We compare the insights deducted from both approaches. During the domain-led approach, the variables (features) used for clustering were agreed and selected by clinical experts in heart failure management. These clinicians are co-authors of this paper (AJP, DM, PC, R. Brisk). The justification for the choice of feature is described in the Methods section. During the second approach, referred to as the data-driven approach, the variables (features) for clustering were determined using a 'pure' data-driven feature extraction process using principal component analysis (PCA) to reduce the dimensionality of the dataset whilst maintaining a degree of variance in the dataset. In the Methods section, we describe two approaches

to the feature selection stage. In Sect. 5.3, we describe the methods that were used in the domain-led feature selection and clustering experiment, whilst in the Sect. 5.4, we present the data-driven approach to feature extraction and clustering analysis. The Results section is divided into two subsections to provide the results from the two approaches. In the Discussion section, we synthesise the results of both approaches and present the advantages and disadvantages of data-driven and domain-led clustering analysis. We present flowcharts illustrating both of the approaches used in this experiment. We go beyond the synthesis of the results from both approaches and we propose a practical checklist that could be used by data scientists to ensure that domain knowledge is embedded in the data mining project focused on healthcare data.

5 Methods

5.1 Materials

For the purpose of this experiment, we chose an open access heart failure dataset available from the Physionet data repository curated by Zhang et al. (2021) [24] [23]. This dataset was collected with the goal of developing a predictive model for classifying emergency readmission of patients with heart failure using data from electronic health records (EHR). The data were collected during the time period of 2016–2019 in the Sichuan Hospital in China. This dataset contains 2008 instances (i.e. patient cases) and 168 variables (i.e. demographic and clinical features) describing the characteristics of patients with HF. Curators of the dataset, when identifying patients with HF, used the definition of the HF according to the European Society of Cardiology (ESC) [25], i.e. the presence of symptoms and/or signs of HF and the presence of (1) raised Brain Natriuretic Peptide (BNP) >35 pg/mL or NT–proBNP >125 pg/mL or (2) objective evidence of underlying functional or structural cardiac abnormalities evidenced by (3) stress test or (4) invasively measured elevated left ventricle (LV) filling pressure.

The data analysis in this paper was performed using MATLAB (version 2021b), with functions included in the Statistics and Machine Learning Toolbox [26].

5.2 Exploratory data analysis and data pre-processing

There are several limitations of the dataset [23] [24] used for this experiment. The dataset does not offer time series data over the hospitalisation period. Whilst there are 2008 patients in the dataset with 167 variables, there are several missing fields for variables that are considered important from a clinical domain perspective. As shown in Table 1 where we provide the percentage of missing values for selected features, only as little as 32% of patients included in this study, have data about their left ventricular ejection fraction (LVEF). LVEF is an important characteristic that is provided by an echocardiogram, an ultrasound heart scan (ECHO). The severity of the HF is defined by the range of the LVEF. Current international guideliness distinguish three types of HF. This includes, 1) HF with reduced ejection fraction (HFrEF) for LVEF <40%, 2) HF with mildly reduced ejection fraction (HFmrEF) with LVEF between 41–49%), and HF with preserved ejection fraction (HFpEF) with LVEF >50% [27]. LVEF range is used to categorise patients into the type of HF as well as to prescribe appropriate HF treatment. LVEF cut-off points are also used in selecting patients to participate in clinical trials. For the above reasons, we decided to perform the clustering analysis using only the data that contains patient records with known LVEF value and this was available for 635 patients (635 instances in the dataset).

Table 1 Number of missing instances for each variable in the dataset. % out of 2008 instances. LVEF—left ventricle ejection fraction, LVEDD—left ventricle end diastolic dimension, BNP—brain natriuretic peptide, GFR—glomerular filtration rate, CK—creatinine kinase

Variable	Missing	% Missing value
LVEF	1373	68%
LVEDD	679	33.80%
BNP	35	1.70%
Troponin	79	3.90%
Coagulation (7 variables)	34	1.70%
Haemoglobin	28	1.40%
Lipid Profile (4 variables)	198	10%
Total protein	102	5%
Electrolytes (5 variables)	11	0.50%
Creatinine/Urea	23	1%
GFR	63	3%
Lactate	241	12%
CK	241	12%

The dataset provides information about the re-hospitalisation due to HF and reports patient mortality. This dataset consists of 50 categorical and 117 numerical variables. Out of 117 numerical variables, only 68 variables had less than 10% missing values for pre-selected patients with known LVEF. Only numerical variables were used in the clustering analysis, because the k-means algorithm performs best on numerical data.

To assess data distributions, we used visual and statistical methods. Each variable was plotted on a quantile–quantile plot (qqplot), which displays the quantiles of the sample data versus the theoretical quantiles from a normal distribution [28]. On visual inspection, it was clear that the numerical variables did not follow a normal distribution in this

dataset. In addition to visual inspection, we used a one-sample Kolmogorov–Smirnov test (p-value < 0.05) to check if the data were normally distributed [29]. The null hypothesis of the normal distribution of the data was rejected for each variable.

Following the findings of the exploratory data analysis (EDA), in the pre-processing stage, we addressed missing values in the dataset by imputing the median value for 10% of missing data points. Due to not normally distributed nature of variables, we used the single imputation technique with median value for missing 10% values of each variable. We normalised the dataset by scaling all feature values to the range 0:1. We did not remove outliers from the dataset as this could potentially lead to the loss of important information about groupings in the data. The final dataset consisted of 635 instances with 68 variables.

5.3 Experiment 1: Domain-led approach to feature selection

The main difference between domain-led approach and the data-driven approach is the feature selection stage. In order to agree on variables to be passed into the k-means clustering algorithm, the clinical co-authors reviewed the variables and decided upon using the following features: "brain natriuretic peptide (BNP)", "haemoglobin", "mean corpuscular volume" (MCV), "creatinine enzymatic method", "sodium", "albumin" and "left ventricle ejection fraction (LVEF)". The decision to select these particular features was driven by clinical experience and knowledge of the outcomes of previous randomised controlled trials (RCTs) as well as observational studies in HF [30] [31]. Based on experience, those variables accurately characterise the severity of heart failure. Moreover, it has been shown that these features carry a prognostic value in the course of HF with regard to diagnosis, prognosis, quality of life and hospitalisation.

From the available variables, the clinicians selected brain natriuretic peptide (BNP) as the current standard for diagnosis and monitoring of HF. BNP levels correlate with the New York Heart Association (NYHA) classification of HF. BNP is a test of high specificity and sensitivity. BNP levels greater than 100 pg/mL have a specificity greater than 95% and a sensitivity greater than 98% when comparing patients without HF to all patients with HF [27]. It was a strong argument to choose BNP as one of the features for the clustering experiment.

Haemoglobin was chosen as indicator of anaemia in general and MCV as indicator of iron deficiency anaemia. It is known that up to 50% of patients with HF suffer from iron deficiency [32]. Using the variables "haemoglobin" and "mean corpuscular volume" we could potentially capture the severity of iron deficiency anaemia during the clustering experiment. Iron deficiency anaemia is evidenced by low haemoglobin level, low mean corpuscular volume of the red cell (MCV) and low iron serum level. Iron deficiency anaemia is associated with a lower quality of life, reduced exercise tolerance and increased mortality in HF patients [27]. RCTs (IRONMAN – NCT02642562, AFFIRM-AHF – NCT02937454, FAIR-HF2 – NCT03036462, HEART-FID — NCT03037931, FAIR-HFpEF – NCT03074591) [32] and meta-analyses [33] [34] have demonstrated that intravenous iron supplementation in HF patients with iron deficiency improves symptoms, quality of life and exercise tolerance (as measured by VO2 peak and 6 minute walk test (6MWT)), with an observed trend to reduction of hospitalisation rates. Creatinine was chosen as an indicator for the possible presence of cardio-renal syndrome. During the natural history of HF, patients develop cardio-renal syndrome which is the result of the poor renal perfusion secondary to low cardiac output present in HF. Patients with severe HF continue to develop chronic kidney disease that gradually progresses to irreversible renal failure [35]. Cardio-renal syndrome has a negative impact on HF patients' outcomes, and the stage of the kidney disease carries a prognostic value [35].

Hyponatraemia (low sodium level) is a strong predictor of the severity of HF and is strongly correlated with increased mortality [30] [31]. Hypoalbuminaemia (low albumin level) is commonly a sign of cachexia—malnutrition—which is frequently described in patients with HF despite normal or above normal Body Mass Index (BMI) [36]. Out of variables obtained by ECHO, left ventricle ejection fraction (LVEF) was selected as the most representative feature to characterise the severity of heart failure. We provided argument and justification for selecting LVEF in Sect. 5.2 on exploratory data analysis and data pre-processing.

In addition to above reasons for selecting specific variables, by having a prior knowledge of variables, clinicians intuitively avoided choosing the variables that are either the ratio of other variables or highly correlated variables. For example 'INR' is an International Normalised Ratio, which is derived from prothrombin time (PT) which is calculated as a ratio of the patient's PT to a control PT standardised for the potency of the thromboplastin reagent. This formula was developed by the World Health Organization [37]:

$$\text{INR} = \text{Patient}\,PT \div \text{Control}\,PT \tag{1}$$

Haematocrit is another example of the ratio calculated from the full blood count—it is a ratio between cell concentration-to-blood serum volume ratio.

Some of those variables (not all at once, or in one study) have been commonly chosen in previous studies using machine learning to perform clustering or classification tasks [3] [4] [5] [6]. Amhad et al. (2018) for example used 8 variables with k-means clustering, including the variables: age,

creatinine, haemoglobin, weight, heart rate, systolic blood pressure, mean arterial pressure, and income [8].

5.4 Experiment 2: data-driven approach to feature selection

High-quality clustering produces a number of clusters, which are typically characterised by high within-cluster similarity but high between cluster dissimilarity. This objective of high within-cluster similarity whilst maintaining high between cluster dissimilarity is particularly difficult to achieve whilst applying clustering to a highly dimensional dataset. In the data-driven approach, principal component analysis (PCA) was used to reduce the dimensionality of the dataset whilst maintaining a high degree of the variance in the dataset. The final dataset (635-by-68) was passed through a PCA function. PCA used singular value decomposition (SVD) algorithm based on a variance-covariance matrix. The algorithm centred the X (n-by-p matrix) by subtracting column means before computing SVD. PCA used all of the observations for the matrix n-by-p (635-by-68) and returned all 68 principal components. In the results section, we present the scree plot that shows the explained variance of each PC that was studied and we present the loadings that were considered the most important for each principal component.

In next step, we needed to decide how many PCs were to be used for the k-means clustering algorithm. We decided to use the first 22 PCs which were informed by the method described by Tabachnick et al. [38]. It has been suggested that while dealing with moderate size data and a large amount of variables, the number of PCs to be selected for further analysis could be decided based on a simple calculation by taking into account the number of variables. The range was chosen between number of variables divided by 5 and divided by 3. This way the analyst can decide on the optimal number of PCs out of the range: (p/5, p/3), where p is number of variables of the n-by-p data matrix. Tabachnick et al. describes the visual method of deciding upon the number of PCs to be considered; however, this method may not be accurate and might be prone to variability due to the subjectivity of the plot assessment [38].

Holland [39] proposed another method for selecting the principal components with the assumption that all variables contributed the same variance to the PC. In this case, it is recommended to select all PC that are equal or greater than 1/p, where the p is the number of variables used in the dataset. In our experiment, the cut-off point is 1.47. Figure 1 shows all 68 PCs with the horizontal line that is used as the cut-off point to accept the first 22 PCs. Beyond this point, the remaining PCs do not carry high enough loadings in the new environment, and hence they were not taken to the next stage of the experiment.

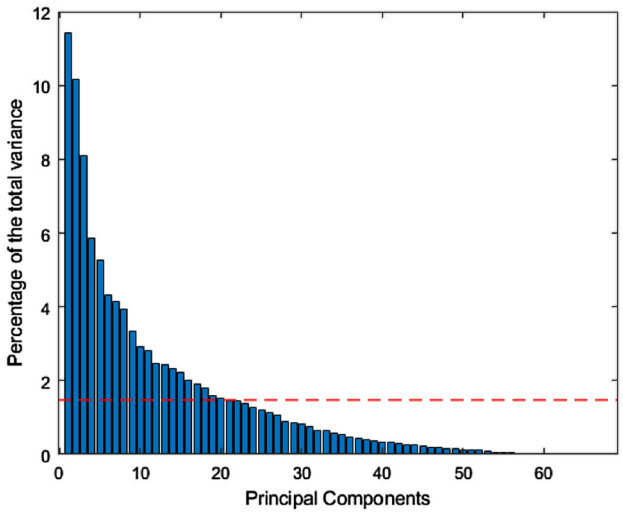

Fig. 1 Scree plot where each PC is represented as a bar in descending order of the percentage of the total variance explained by each principal component. PC after passing dataset consisting of 68 numerical variables for 635 patients. Horizontal line is a cut-off point equalled 1/68 (1.47)

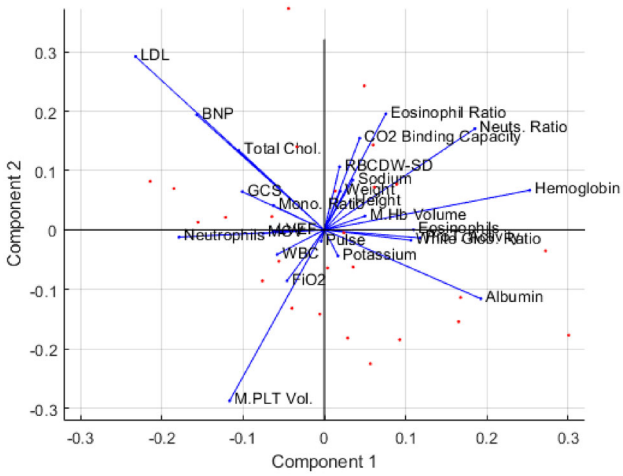

Fig. 2 Biplot of 26 variables with highest loadings (loading > 0.32) in each of the contributing to first 22 principal components

To further investigate the first 22 PCs, we used the loadings from the coefficient matrix to identify variables with loadings that are greater than 0.32. The decision to explore eigenvalues with loadings that are greater than 0.32 was supported by the rule of thumb described by Tabachinck et al. [38]. Tabachnick et al. state that variables with loadings greater than 0.32 contribute the most to a given PC. Using this rule of thumb, we identified 26 variables that had the highest contribution to the first 22 principal components. Figure 2 presents a biplot showing these 26 variables.

In the next stage, we used the 22 PCs to pass into the k-means clustering algorithm. Whilst we chose PCs that retained 81% of the variance in the dataset, we were able

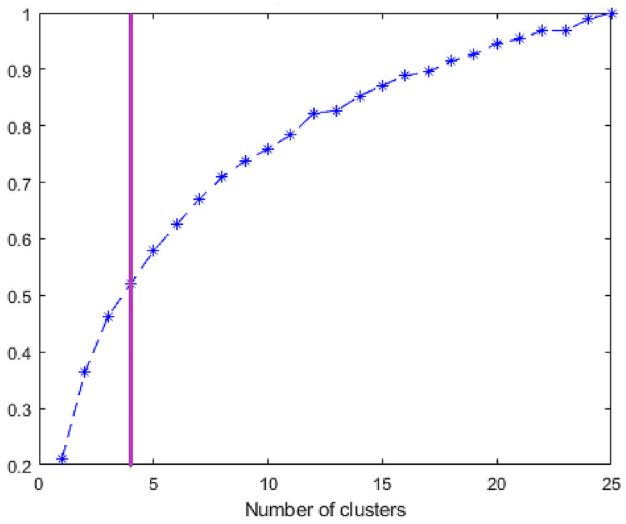

Fig. 3 Data-Driven approach. Graph shows the "elbow method" representing number of clusters when the algorithm is applied to 635-by-22 matrix (22 columns are representing first 22 PC). Vertical purple line identifies optimal number of clusters

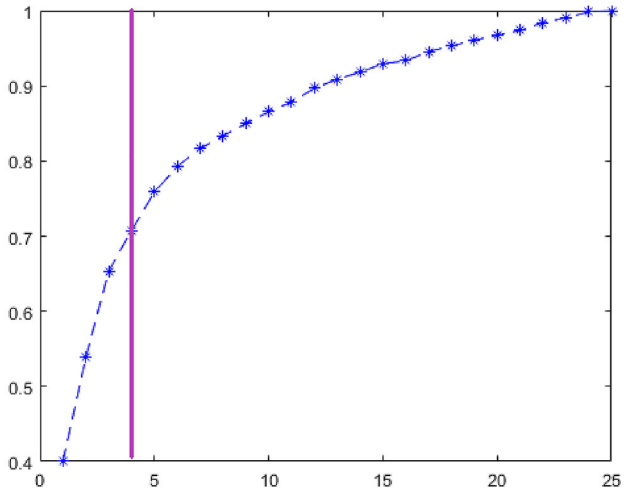

Fig. 4 Domain-led approach. The graph shows the "elbow method" which can be applied to chose the optimal number of clusters when the k-means clustering algorithm is applied to the 635-by-7 matrix (7 columns are representing variables chosen by domain experts to the domain-led clustering experiment). Vertical purple line identifies optimal number of cluster

to reduce the dimensionality from 68 features to new 22 features represented by PCs.

5.5 k-means clustering: optimal number of clusters

To identify the optimal number of clusters for k-means clustering in both approaches, we used a visualisation technique known as the 'elbow method' (Fig. 3, 4). We decided to divide the dataset into four clusters in both approaches. To assess the inter-cluster separation, we also used a silhouette criterion and silhouette graph for both methods [40], which confirmed that four clusters were to be optimal number of good quality clusters.

6 Results

6.1 Results of experiment 1: domain-led approach

Figure 5 shows a summary of the characteristics of each cluster derived by using the domain-led approach. Table 2 provides median value for each of seven variables used in the clustering experiment. Table 3 provides summary of comorbidities observed in each cohort. Utilising the domain-led approach, we identified the following clusters: *Cluster 1* was the cluster with the second most impaired heart function, as per median BNP of 1591 and median LVEF of 39% with the range (17–49%). This cluster was similar to Cluster 2 in terms of the prevalence of kidney disease, with the second lowest prevalence of lung disease.

Cluster 2 was the largest cluster and had the least impaired heart function as defined by the highest median LVEF 59% and the lowest median BNP level (308). This cluster had the highest albumin level and the lowest creatinine level, indicating overall least impaired heart function and least impaired kidney function. In terms of comorbidities, *Cluster 2* has the lowest prevalence of myocardial infarction and chronic kidney disease (15%).

Cluster 3 had most severely impaired heart function with the highest BNP 4486 and the lowest range of the LVEF 43% (5–65%). Cluster 3 had the highest prevalence of lung disease, liver disease and dementia.

Cluster 4 is the cluster with the highest prevalence of diabetes and chronic kidney disease. This cluster has the highest prevalence of cerebrovascular disease (stroke) (7.6%).

6.2 Results of experiment 2: data-driven approach

In Tables, 4, 5, 6 and in Fig. 6, we present the most distinctive features that characterise each of the clusters. In the supplementary material, we provide additional Table 9 which breaks down all the remainder of the characteristics for all 4 clusters that resulted from the data-driven clustering approach.

Cluster 1 has the highest value of MCV, MCHC, Mean PLT Volume, Eosinophil Count and Eosinophil Ratio. This cluster has the highest median value of Albumin, white globulin, sodium, prothrombin activity and CO2-binding capacity. *Cluster 1 BNP and LVEF:* the lowest BNP level and the second highest LVEF. In terms of 42 variables that were used to clustering, but with less contribution to the first 22 PC taken into the k-means clustering, *Cluster 1* has the high-

Table 2 Domain-led approach. Median value (Minimum - Maximum) of 7 variables selected by domain experts to be used in the k-means clustering of HF cohort

	Cluster 1 (n = 179)	Cluster 2 (n = 256)	Cluster 3 (n = 69)	Cluster 4 (n = 131)
LVEF	39 (17–49)	59.5 (39–82)	43 (5–65)	58 (22–76)
Creatinine Enzymatic Method	82 (38–293)	70.6 (35–308)	101 (43–553)	89 (32–963)
Brain Natriuretic Peptide	1591 (172–3134)	308 (2–2400)	4486 (2971–5000)	519 (36–3191)
Sodium	139 (115–148)	140 (122–148)	137 (123–144)	138 (120–148)
Albumin	36 (17–49)	38 (24–49)	36 (22–43)	33 (17–42)
Mean Corpuscular Volume	93 (69–111)	94.3 (52–117)	91 (69–107)	85 (58–108)
Haemoglobin	125 (78–181)	123 (86–163)	116 (123–144)	87 (31–136)

Table 3 Domain-led approach, characteristics of clusters, prevalence of clinical conditions in the cluster

	Cluster 1 (n = 179)	Cluster 2 (n = 256)	Cluster 3 (n = 69)	Cluster 4 (n = 131)
Myocardial Infarction	6.1%	5.9%	**13%**	6.1%
Congestive Heart Failure	100%	100%	100%	99.2%
Peripheral Vascular Disease	5%	5.5%	2.9%	**9.9%**
Cerebrovascular Disease	5%	5.9%	5.8%	**7.6%**
Dementia	6.7%	8.2%	**11.6%**	8.4%
COPD	9.5%	8.2%	**14.5%**	11.5%
Connective Tissue Disease	0.6%	0%	0%	0.8%
Peptic Ulcer Disease	1.7%	0.8%	1.4%	7.6%
Diabetes	24%	20.7%	20.3%	**33.6%**
Moderate To Severe CKD	17.9%	15%	33.3%	**36.6%**
Hemiplegia	1.1%	0.8%	1.4%	1.5%
Leukaemia	0%	0%	0%	0%
Malignant Lymphoma	0.6%	0%	0%	0%
Solid Tumour	1.7%	1.2%	2.9%	**4.6%**
Liver Disease	4.5%	1.2%	**7.2%**	2.3%

COPD chronic obstructive pulmonary disease, *CKD* chronic kidney disease Percentages in bold face indicate the highest prevalence of particular condition in the Cluster, when compared to other Clusters. High prevalence of this condition makes this a distinctive feature of the Cluster

Table 4 Data-driven approach. This table shows the prevalence of medical history documented for patients in each cluster

	Cluster 1 (n = 287)	Cluster 2 (n = 104)	Cluster 3 (n = 100)	Cluster 4 (n = 144)
Myocardial Infarction	6%	**11%**	3%	8%
Peripheral Vascular Disease	7%	5%	**10%**	3%
Cerebrovascular Disease	7%	6%	7%	4%
Dementia	7%	9%	8%	**10%**
COPD	7%	**14%**	11%	12%
Connective Tissue Disease	0%	0%	2%	0%
Peptic Ulcer Disease	1%	0%	**11%**	1%
Diabetes	21%	24%	27%	**30%**
Moderate To Severe CKD	14%	34%	**40%**	20%
Hemiplegia	1%	3%	0%	1%
Malignant Lymphoma	0%	0%	0%	1%
Solid Tumour	1%	2%	6%	2%
Liver Disease	1%	9%	3%	3%

COPD chronic obstructive pulmonary disease, *CKD* chronic kidney disease Percentages in bold face indicate the highest prevalence of particular condition in the Cluster, when compared to other Clusters. High prevalence of this condition makes this a distinctive feature of the Cluster

Table 5 Characteristics of clusters from the data-driven approach. Top 26 variables contributing the most to the first 22 PC passed through the k-means algorithm. Median (Minimum - Maximum) value provided for each cluster

	Cluster 1 (n = 287)	Cluster 2 (n = 104)	Cluster 3 (n = 100)	Cluster 4 (n = 144)
Pulse	80 (32–158)	86 (40–180)	82 (42–156)	91 (38–165)
Weight	51 (30–96)	50 (30–84)	50 (30–83)	50 (8–95)
Height	1.6 (1.4–1.8)	1.6 (1.4–1.7)	1.6 (1.2–1.7)	1.6 (1.4–1.8)
FiO2	33 (21–41)	33 (21–100)	33 (21–100)	33 (21–100)
LVEF	**54 (19–82)**	**40 (5–68)**	**58 (22–76)**	**52 (20–75)**
WBC	5.9 (2.4–11.3)	6.5 (2.9–16.6)	5.8 (2.4–25.4)	10.1 (4.4–26.3)
Monocytes Ratio	0.065 (0.03–0.2)	0.07 (0–0.1)	0.0675 (0–0.2)	0.06 (0–0.1)
RBCDW-SD	47 (34–65.3)	48.95 (38.3–77.8)	50.2 (36.3–83.2)	46.6 (34.4–71.2)
Mean Corpuscular Volume	**94 (63–118)**	**93 (69–111)**	**86 (59–108)**	**92 (64–106)**
Mean Hb Volume	31 (20–39)	31 (21–37)	27 (16–35)	31 (20–35)
Mean PLT Vol.	13 (9–17)	13 (9–16)	11 (8–16)	12 (8–17)
Eosinophil Ratio	0.017 (0–0.2)	0.005 (0–0.1)	0.009 (0–0.1)	0.003 (0–0.1)
Eosinophil Count	0.1 (0–1.2)	0.03 (0–0.3)	0.05 (0–0.4)	0.03 (0–0.4)
Haemoglobin	**120 (82–163)**	**120 (69–164)**	**79 (31–119)**	**128 (65–181)**
Neutrophil Ratio	0.7 (0.3–0.9)	0.8 (0.6–0.9)	0.8 (0.5–1)	0.8 (0.6–1)
Neutrophil Count	4 (1–7)	5 (2–14)	5 (1–21)	8 (3–24)
Prothrombin Activity	74 (9–142)	57 (12–91)	65 (8–98)	72 (12–131)
CO2-Binding Capacity	25 (11–37)	21 (12–32)	23 (10–39)	24 (11–43)
Potassium	4 (3–6)	4 (3–7)	4 (3–7)	4 (2–7)
Sodium	**141 (132–148)**	**138 (124–145)**	**139 (121–148)**	**137 (116–144)**
Brain Natriuretic Peptide	**488 (3–2987)**	**3435 (668–5000)**	**571 (36–5000)**	**667 (17–4192)**
Albumin	**38 (23–49)**	**36 (24–45)**	**33 (17–43)**	**37 (18–50)**
White Globulin Ratio	1.4 (1–2.6)	1.3 (0.8–2.4)	1.2 (0.4–2.1)	1.2 (0.6–2)
Total Cholesterol	3.8 (1.7–7.2)	3.5 (1.8–8.4)	3 (1.6–5.7)	4.5 (2–8.4)
LDL	1.83 (0.4–4.3)	1.78 (0.7–5.1)	1.3 (0.4–2.8)	2.36 (0.7–5.2)
Glasgow Coma Scale	15 (13–15)	15 (3–15)	15 (3–15)	15 (3–15)

Variables in bold face are the 6 out of 7 variables selected by clinicians to be passed through the clustering algorithm during the domain-driven approach. "creatinine enzymatic method" is not amongst variables contributing the most to the top 22 PCs, despite being selected by clinicians as important variable. "creatinine enzymatic method" did not show either high correlation with other variables nor high variance in contrast to variables that are presented in this table

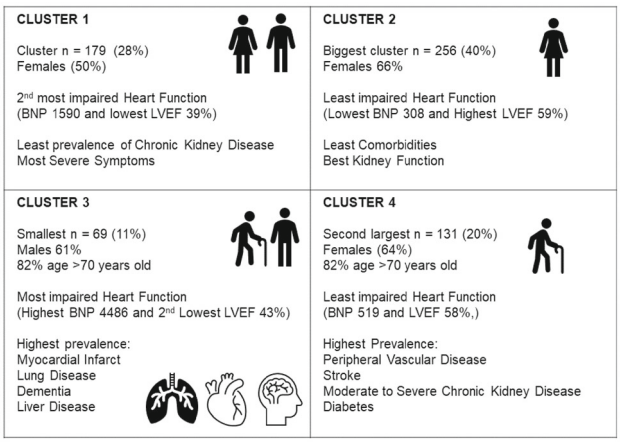

Fig. 5 Domain-led approach. Summary of most distinctive features in each cluster

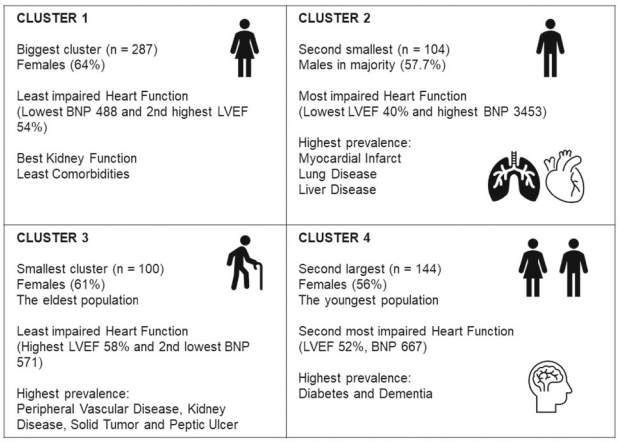

Fig. 6 Data-driven approach. Summary of most distinctive features of each cluster

Table 6 Data-driven Approach. This table shows demographic characteristics including age and gender as well as reported symptoms according to NYHA class (New York Heart Association Functional Classification) and Killip classification

	Cluster 1 (n = 287)	Cluster 2 (n = 104)	Cluster 3 (n = 100)	Cluster 4 (n = 144)
Female %	64.1%	42.3%	61%	56.9%
Male %	35.9%	57.7%	39%	43.1%
Age group				
21–29	0.3%	1%	0%	0.7%
30–39	0.7%	1%	0%	0%
40–49	1.7%	2.9%	0%	4.9%
50–59	5.6%	5.8%	4%	7.6%
60–69	22%	18.3%	9%	21.5%
70–79	37.6%	39.4%	38%	29.2%
80–89	27.9%	29.8%	38%	29.2%
90–110	4.2%	1.9%	11%	6.9%
NYHA				
I				
II	10.5%	0%	3%	2.1%
III	62.7%	41.3%	56%	43.1%
IV	26.8%	58.7%	41%	54.9%
Killip				
I	43.2%	20.2%	26%	22.2%
II	46.7%	64.4%	56%	49.3%
III	8.7%	11.5%	15%	25%
IV	1.4%	3.8%	3%	3.5%

est systolic and diastolic BP, resulting with the highest mean arterial pressure (MAP). This cluster has the highest BMI, lymphocyte count, monocyte count, basophil count, basophil ratio and the lowest platelet count with the highest platelet width distribution (PLT-WD) and the highest chloride. This cluster has the highest glomerular filtration rate (GFR), with the lowest creatinine (68), urea (6.2) and the lowest uric acid (386), and the lowest ALP, direct bilirubin and the lowest globulin. *Cluster 1* has the lowest INR, prothrombin time ratio with the lowest high sensitivity troponin. In summary, this cluster has the best kidney function and the best heart function when compared to the other clusters.

Cluster 2 has the lowest prothrombin activity and the lowest CO2-binding capacity. *Cluster 2 BNP and LVEF:* the highest BNP (3435) and the lowest LVEF (40%) *Cluster 2* has the lowest systolic and diastolic BP, with the lowest MAP. This cluster has the highest left ventricle end diastolic dimension LVEDD (62mm). This cluster has the highest urea and uricacod levels with the second highest creatinine level. This cluster has the highest Ddimer level, INR (1.36), APTT, prothrombin time and PT ratio, with the lowest fibrinogen level. IN terms of liver enzymes, *Cluster 2* has the highest GGT (59), ALT (34), and the highest total bilirubin (27.2), indirect bilirubin (15.6), and direct bilirubin (10.3). In summary, Cluster 2 has the worst heart function with the highest BNP, LVEDD and the lowest LVEF and the poorest liver function.

Cluster 3 has the lowest albumin (33), the lowest total cholesterol LDL ad white globulin levels. This cluster has the lowest haemoglobin (79) and the lowest MCV (86). *Cluster 3 BNP and LVEF*: second lowest BNP and the highest LVEF (54%). *Cluster 3* has the highest creatinine level with the lowest GFR. This cluster has the lowest red blood cell count, lowest hematoctrit, monocyte count, lymphocyte count and the lowest PLT hematocrit and the lowest PLT -DW. This Cluster has the lowest GGT, ALT, total bilirubin, total protein, triglycerides and HDL. In summary, this cluster comprises patients that have the worst kidney function, with severe anaemia evidenced by low haemoglobin and MCV, with the some stigmata f malnutrition evidenced by the lowest total protein, cholesterol and HDL.

Cluster 4 has the highest haemoglobin, white blood cell count, neutrophil count, total Hb volume, total cholesterol and the highest LDL. *Cluster 4 BNP and LVEF*: the second highest BNP (667) and the second lowest LVEF (52%) . *Cluster 4* has the highest red blood cell count with the highest monocyte count and hematoctrit. This cluster has the highest APTT and thrombin time with the highest fibrinogen. This cluster has the highest globulin level, total protein, triglyceride and HDL. This cluster has mostly moderate values for LVEDD, creatinine , urea, uric acid and GFR. In summary, this cluster has good heart function with normal kidney function.

7 Discussion

This experiment demonstrates that domain knowledge significantly reduces the data dimensionality of the feature set and plays important role in the interpretation of the clustering results.

Our aim was to explore how domain knowledge influences the stages of data science project and how it can help solve challenges posed by domain specific issues. Moreover, by example of this experiment we wanted to bring attention to the need of active involvement of domain experts in data mining process. The goal of this experiment was an attempt to address some of the gaps identified in domain-led data mining process [10].

7.1 Challenges of working with healthcare data

The healthcare sector produces one third of the globally stored digital data; hence, it seems obvious that clinical experts need to be involved in key stages of data science projects to unlock new insights and to integrate layers of clinical knowledge [41]. In the case of the healthcare sector,

governments have recognised the need to train the clinical workforce in data analytics to help improve the integration of domain knowledge into the data science process and to prepare clinical teams to embrace the opportunities arising from digital transformation [42]. There is an expectation that clinical teams become skilled in data analytics that is beyond their already acquired knowledge of biostatistics, which is an existing curriculum requirement in medical schools and postgraduate training programmes [42] [43]. The involvement of domain experts is expected at various stages of the data science project. Starting with (1) the problem definition, (2) proposing and curating a target dataset, (3) data cleaning, pre-processing and data transformation, (4) feature set and algorithm selection, (5) the evaluation and interpretation of learned knowledge and suggestion of practical use of the new knowledge to improve processes within the specialist domain.

Skills in the practical application of advanced analytics including artificial intelligence (AI) and machine learning (ML) together with the ability to critically appraise the results will enhance knowledge discovery and provide innovations for the healthcare sector [12]. Education and training in AI and ML will increase the uptake of modern technologies that still suffer from the 'black box' stigma. Explainable ML methods must be understandable to the end users, i.e. the clinicians; moreover, clinicians and analysts must use a common language and have a comparable set of analytical skills.

In this experiment, a particular challenge was posed by high dimensionality of the dataset, which is not an uncommon challenge when dealing with healthcare data [44] [45]. We dealt with this issue by using domain knowledge to 'hand pick' features to be passed through the clustering algorithm. The advantage of the feature selection performed by domain experts as opposed to feature extraction enabled by algorithms like PCA is the immediate interpretability of the former [46]. In order to assure that the data-driven method could be used by other clinical teams, we aimed to improve the explainability of the data-driven approach to physicians. It was important to present how the PCA algorithm operates and to indicate which variables contributed the most to the PCs passed through the k-means algorithm. When we analysed the make up of the top 22 PCs and once the variables contributing the most were identified, we found out that only six out of seven variables chosen by the physicians were among the 26 variables carrying the highest value in top 22 PCs (LVEF, MCV, haemoglobin, sodium, BNP and albumin). Interestingly, "creatinine enzymatic method" was not amongst variables contributing the most to the top 22 PCs, even though it is felt to be an important variable from a clinical perspective. One possible reason why creatinine did not appear in the top contributing features could be the fact that PCs are aggregates of correlated variables with high variance, whereas in this dataset "creatinine enzymatic method" did not show either high correlation with other variables nor high variance. On a closer look at the reminder of top 26 contributing variables, we noted that total white cell count, eosinophil count, neutrophil count, as well as other variables obtained during the analysis of the full blood count were among these 26 variables. Those indices are produced by the haematology analyser and in most cases they come from a single blood sample. Physicians would be aware that those indices are usually highly correlated; hence, it is not a surprise to see those variables in the top PCs, as all components in PCA represent an aggregation of the correlated variables. The data-driven approach identified 4 clusters and was effective in identifying smaller clusters with strongly distinctive features, in terms of comorbidity and underlying physiology. As a result of the domain-led approach, 4 clusters were identified; however, on a closer look, the clusters had a similar prevalence of the comorbidities. To date, in the literature, we could not find a standard measure for evaluating clustering or standard measures for evaluating unsupervised feature selection methods for clustering [47]. There are, however, some commonly used internal and external measures that can be used for the quality assessment of clusters generated by a clustering algorithm [40]. Clustering solutions can be assessed externally based on how much it resembles a set of classes, commonly known as ground-truth or 'expert classification' [40]. This 'expert classification' is nothing else but manual tagging with class labels by human experts. As shown in Figure 7, domain knowledge contributed significantly to all stages of the clustering experiment.

7.1.1 Interpretability of the results

This paper shows that domain knowledge plays an important role in providing an analysis of results obtained through clustering. The knowledge of the physiology, pathology and correlations between a set of variables allowed domain experts to reduce the number of variables to be used in the clustering experiment from 68 to 7. This approach was at risk of bias though, as domain experts opted to use the most commonly used variables to describe the stage of the HF in day to day clinical work. Those variables are well known as they carry prognostic values based on observational studies and RCTs. Such variables have been used for many years in clinical practice and it seemed justified to choose them to best describe clusters of HF patients. The choice appeared straightforward due to the fact that the variables, which are used to describe HF patients, are akin to a distinct language or code that is both universal and understood by clinicians. With a similar ease, we approached the interpretation of the clustering results. Knowing the normal range for all 68 variables by heart, it was a straightforward task to describe clusters of patients and draw conclusions regarding underlying pathological processes. For example, in the case of

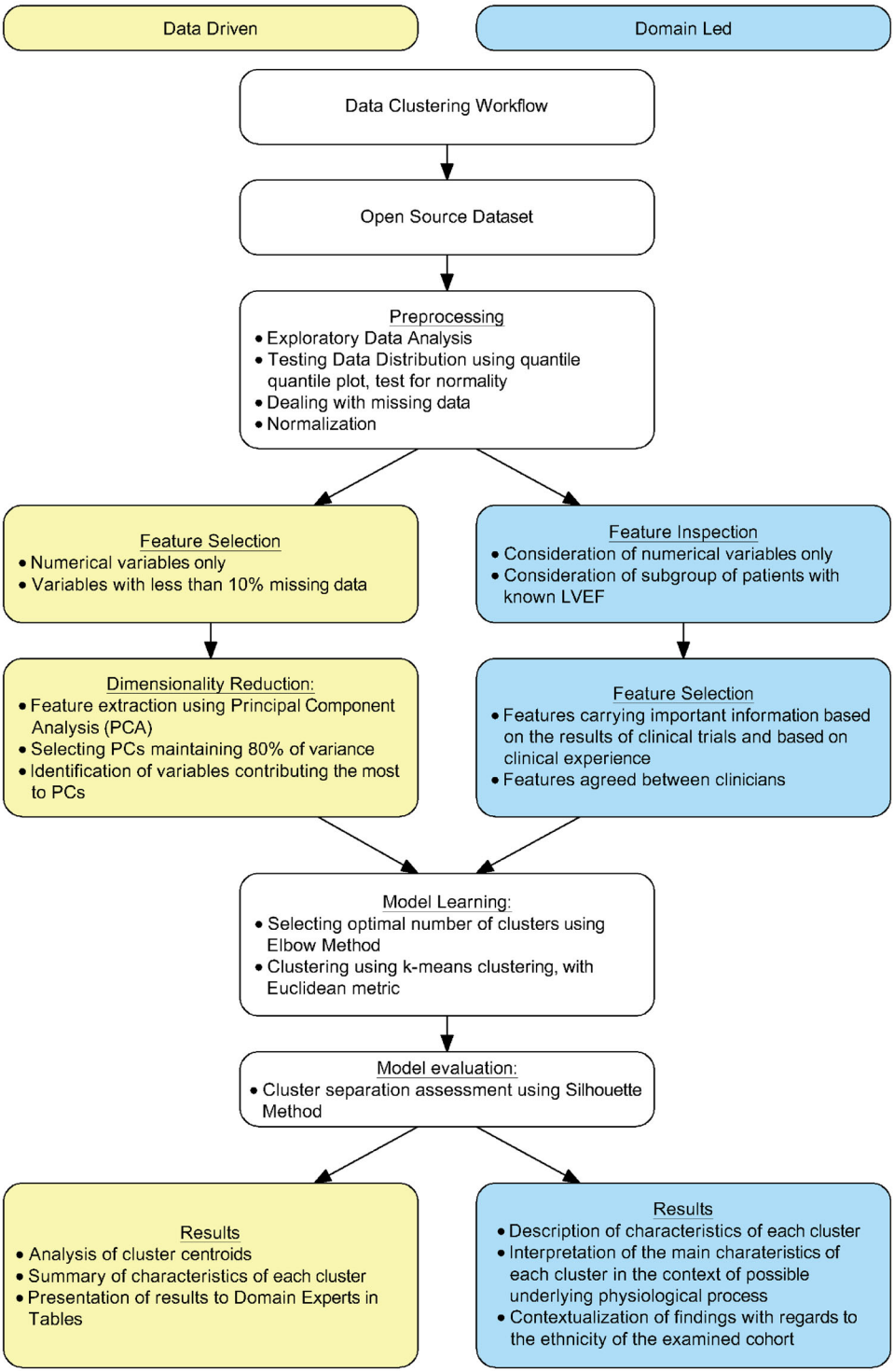

Fig. 7 This figure shows learning processes used during clustering experiment. The branches on the left (shown in yellow) present stages unique to the data-driven approach, whereas the branches on the right (shown in blue) present stages unique to the domain-led approach

Cluster 3 from the data-driven approach, this cluster had the lowest median value of haemoglobin, with the lowest MCV, signifying iron deficiency anaemia and features of malnutrition, with the lowest total protein level and albumin. It was not a surprise that in this cluster there have been the highest prevalence of the peptic ulcer disease (11% of the cluster 3) and the highest solid tumour presence (6% of the cluster 3). As clinical domain experts, we would not make the mistake of labelling this cluster of HF patients as "anaemic and malnourished" because due to the ability to contextualise the provided information, we know that peptic ulcer disease in itself, but especially presence of the malignancy—hidden here under the term 'solid tumour', can cause the iron deficiency anaemia and can lead to cancer related malnutrition.

What is most important though is that we are still aware that 'correlation is not causation', and our interpretation may be completely wrong either way. Another important aspect that requires attention when working with datasets of a selected population sample is the risk of bias that could be introduced into the dataset. It is well known to physicians that the prevalence of peptic ulcer disease is significantly higher in South Pacific populations compared to Western populations. Data from the Systematic Investigation of Gastrointestinal Disease in China showed that the prevalence of the peptic ulcer disease was substantially higher in the Shanghai population (17.2%) than in Western populations (4.1%) [48]. Again, domain knowledge proves critical in preventing analysts from drawing conclusions from an unbalanced dataset or a dataset that represents disease prevalence unique to the population in a certain geographic area.

7.1.2 "Actionability" of the results

In the previous paragraph, we discussed the significance of the interpretability of the clustering results. Interpretability is an excellent advantage of the domain-driven approach that risks, however, being lost or skewed in the data-driven approach. Even though the perceived advantage of the objectivity of the data-driven approach may be tempting on using this approach over the domain-led approach, what is important to emphasise is the "actionability" of the clustering results that is strongly linked with interpretability. "Actionability" is a natural byproduct of interpretability and they both should go "hand in hand" during the data mining process. The Domain Driven in Depth Pattern Discovery (DDID-PD) framework proposed by Cao et al. in addition to providing directions on how the domain knowledge should be put on top of the data-driven data mining framework emphasises how the actionability of the data mining can be enhanced. They use the terms of technical and business (domain) interestingness for the purpose of illustrating the process in which the actionable knowledge can be discovered. Authors of the DDID-PD framework, in a form of a mathematical equation, provide a literal prescription for the successful domain-driven data mining, exemplified by the cases of mining actionable correlations in the stock market. Following this framework, the actionable pattern can be discovered whilst two conditions are met: the technical interestingness and business interestingness. The DDID-PD framework captures the essence of the successful domain-driven data mining and is certainly general enough to be applied to other sectors. We see the applicability of the DDID-PD framework to the healthcare sector and the healthcare data. In terms of actionability of the clustering results we would like to suggest the following 3 levels of actionability:

1. Low level action is associated with the discovered taxonomy and labels for the clusters. For many years, taxonomies and classification methods have played a significant role in science and provided frameworks to present knowledge. In practical terms, the use of labels for representing the different types of patients (clusters) could allow new ways for monitoring temporal changes of these clusters/cohorts (in surveillance/epidemiology), for example monitoring the size of those cohorts or other characteristics and be an indicator of the population characteristics of a certain healthcare facility or region.
2. Intermediate level of action could be associated with designing new clinical research protocols looking at specific cohorts of patients derived from data by clustering experiments. Groups of patients with specific features could be studied with respect to the cause of the pathology and potential new treatments.
3. Significant level of action can be implemented by redesigning the healthcare services to enhance the detection of the health condition in specific clusters of patients. Tracking the quality of care, impact on quality of life, comorbidities and mortality statistics of specific cohorts of patients, with frequently occurring health problems and with specific health needs could be used for clinical auditing purposes, as an evidence for the quality improvement interventions, clinical pathways streaming and service redesign.

7.2 Importance of the "Domain Knowledge"

Whilst this experiment did not provide any groundbreaking knowledge about HF itself, it is a useful case study demonstrating how domain knowledge can help navigate analysts through a healthcare data mining project. As far back as 2002, Kopnas et al. concluded that " *in terms of the actors involved in the data mining process, domain experts should be in prominent positions within data analysis, data mining, data warehousing and data processing and should actively participate in and guide the process*" [2]. Based on available publications [12] and voices of data science experts from the industry [49], the importance of the first pillar of the Cross Industry Standard Process for Data Mining (CRISP-DM) framework, which is a "Business Understanding", seems to be undervalued.

As a learning point from this experiment, we would like to propose a practical checklist to enhance the engagement of domain experts and the application of the domain knowledge in the data mining project related to healthcare data. It is important within the healthcare industry that analysts have an adequate understanding of the "domain" and that the domain experts (clinicians) help to navigate the direction of the analysis and point towards questions relevant from the clinical perspective.

Table 7 Proposed checklist enabling the integration of domain knowledge into the data mining project. Continuation of the checklist is provided in Table 8

No.	Checklist item:	Sample answers based on the study in this paper.
1	Have you engaged with multiple domain experts, e.g. clinicians, and what qualifies them as experts?	Yes, we engaged with two cardiology consultants. In addition, the data scientist leading this experiment is a cardiology resident. Cardiology consultants have over 20-year experience of managing patients with heart failure.
2	How much of the domain problem do you understand and how much do you need to understand to complete the project?	Data scientist leading this experiment is a cardiology resident, with over 7-year experience in general cardiology.
3	What exactly is the domain problem to be investigated?	We want to discover what specific groups of HF patients exist according to the Physionet HF dataset.
4	How well do you understand the meaning of variables included in the dataset? How will you assess the quality of the dataset?	The data scientist leading this project is a clinician; hence, the meaning behind each variable was well understood. The expected normal range for the variable was also known. There was a significant amount of time and effort spent by the analyst reviewing the data to ensure that it was fit for purpose before any approach was used. The data preparation side highlights the need for expertise in understanding the data and it's limitations at the outset. In this experiment, we excluded variables with multiple missing values. We had to then impute the median value for missing data for variables with less than 10% missing values, along with normalisation. All these approaches have an impact on the downstream analysis and are important to highlight the need for this understanding and literacy. All the pre-processing was done before data were passed to further stages of the clustering experiment. Domain knowledge played important role at this stage.
4	Have you agreed a "common ground" between domain experts and data scientists? Are you aware that "the common ground" can be dynamic and change during the project? Have you prepared for multiple conversations to find a "new common ground"?	Multiple meetings had taken place between the data scientists and the clinical experts to ensure there were rigorous discussions. This includes explaining the purpose of clustering analysis to the domain experts, whilst the domain experts explained the importance of various clinical features. The common ground was considered to categorise HF patients using Unsupervised ML. Feature selection was discussed and revised throughout the course of the study.
5	What are the exact questions that you need to ask domain experts, e.g. what features/variables in the dataset best characterise these patients?	Cardiologists explained which features are important in assessing the clinical course of a patient with heart failure. Their view was supported by from clinical perspective.
6	What is the purpose of the data science project and what are the measures of the project success as set by the domain experts? (i.e. discovery of new patterns, delivery of detailed report, reduction of administrative burden in the organisation?)	This was an experimental project exploring how data-driven and domain-driven approach could influence the result of the clustering analysis.
7	How are you planning to embed the domain knowledge in this project? At what stages of the project will you seek domain experts input?	Clinicians raised the research question to get more information on existing clusters of HF patients. Clinicians were involved at the pre-processing stage—reviewing percentage of missing values and selecting a subgroup of patients with known LVEF for further analysis. At the feature selection stage, clinicians selected features carrying predictive value based on their own domain expertise. At the results analysis stage, clinicians interpreted the cluster centroids and put them in context.

Table 7 continued

No.	Checklist item:	Sample answers based on the study in this paper.
8	According to the domain experts, would the discovery of new patterns or cluster labels be clinically useful?	This experiment to certain degree illustrated epidemiological and demographic trends observed in cardiovascular diseases: the smallest cluster in the domain-driven approach consisted of majority of men, who had mostly impaired heart function with concurrent the highest prevalence of myocardial infarction and lung disease. We also observed the epidemiological trend that is typical for the geographic region from which the sample was derived—the data-driven approach identified a group of patients with a particularly high prevalence of anaemia and peptic ulcer disease. It is debatable though if the cluster labels are clinically useful.

Table 8 Continuation of Table 7—Continuation of the Proposed checklist enabling the integration of domain knowledge into the data mining project

No.	Checklist item:	Sample answers based on the the study in this paper.
9	If dealing with multiple domain experts how will you achieve the consensus between parties, if differences of opinion occur at the analysis stages?	Fortunately during this experiment, we did not encounter differences of opinion during the discussions between domain experts due to mutual understanding of the problem and agreed "common ground". We debated/discussed via video conferencing over the variables that carry a prognostic value and the variables that are useful in clinical practice.
10	Do domain experts know what will be expected of them during this clustering/data mining project?	We wanted to find subgroups of patients with unique features within patients included in the Physionet Dataset.
11	Do you need to provide basic ML-awareness training to domain experts so that they can appreciate what clustering analysis can achieve?	It was important that data scientist leading this project had received a training during the postgraduate program.
12	Have you considered a formative assessment methods to ensure that the data scientist have sufficient understanding of the domain? Put differently, how did you validate that the data scientist has understood and appreciated the domain knowledge?	On hindsight, more formal methods could have been used to validate that the technical researchers actually understood the domain knowledge being described (for e.g. the teach-back method could have been used).

In Tables 7 and 8, we propose a set of questions that the data scientist should be prepared to address prior to the initiation of the data mining project. This checklist is a result of the collaboration within our team of clinicians and data scientists. We realised that opportunities brought on by advantages in the computational abilities of current software and hardware pose a great temptation to use new machine learning (ML) techniques on healthcare data, especially those available in a public domain. Exploiting new ML techniques on healthcare data may be more effective when performed with the involvement of healthcare expert. Analogous to the trend of a co-design of clinical studies with the involvement of the public and patients representatives' it would seem natural to talk about the co-design and then the co-production of the domain-driven data mining. We hope that this practical checklist for data scientists will enable better integration of the domain knowledge into the data mining project. In addition to the set of questions in the checklist, we provide an example how our team integrated the domain knowledge during the clustering experiment while working on open source heart failure dataset.

Starting with question 1 of the checklist presented in Tables 7 and 8, it is useful to develop a partnership with clinicians, when working on healthcare data early on in the project. With regards to questions 2–4 it is important to understand the exact domain problem which is to be investigated. It is important to know whether the data science project is a part of a research project, Quality Improvement Project (QIP), Clinical Audit, or service evaluation project. If this project is a part of the research, it will be useful to know commonly occurring questions in health research. They can be grouped into 6 main themes: (1) characterising diseases and

describing their natural course, (2) investigating the impact of a disease on the general population as well as assessing correlations between diseases, (3) finding the cause of disease, (4) discovering new treatments or the best treatments combinations out of already existing treatments, (5) assessing the way to deliver the treatment to achieve best result for the patients, (6) learning about the health systems and the costs associated with diseases management.

With regards to question 5, domain experts provide a specific knowledge of the subject and will know aspects related to the data itself. It is important that data scientists leading the project takes an opportunity to find out from domain experts (1) how the data was collected (i.e. was the data manually imputed by clinical staff into database or was it recorded by monitoring devices and automatically saved to patients electronic records), (2) what the data values mean and what is the normal value range (i.e. does the low value of the variable indicates normal state or severe pathology, or in case of time series data, for example does the long history of a certain condition has an impact on long term outcome for the patient and could influence the accuracy of the predictive model if that was objective of the data mining project), (3) the accuracy of the data (is there a risk of error in the data caused by human error whilst data were imputed manually), (4) how to interpret the results of the analysis (i.e. is the result of analysis clinically relevant, do the results make sense to clinicians), (5) the business/domain issues, i.e. could the results alter the current processes in the healthcare organisation.

Mao et al. [50] present factors influencing effective collaboration between teams of bio-medical scientists and data scientists working in the Research IBM. "To find the right answer or to ask the right question?" is the conclusion drawn from interviews with 22 interviewees. It turned out that for bio-medical scientists the original set of questions very early on into the data mining project changes into set of different or "better" questions. This, however, causes a challenge for data scientists who need to adjust to the new "common ground" that is different from the initial "common" ground achieved at the start of the data science project. Mao et al. illustrate the dynamics of the data science project between domain experts (bio-medical scientists) and data scientists. They comment as well on differences of motivations behind the data science project for bio-medical scientists and data scientists. As one of the participants stated, "we are always reproducing predictive models with higher predictive capabilities in the field (...) however we are more interested in what intervention can be done rather than the prediction is accurate" [50]. In addition to detailed analysis of dynamics within the team, they provide an overview of technologies used to enable co-design, communication and collaboration between bio-medical scientits and data scientists (i.e Google Docs, Google Sheets , GitHub, Skype, email, Slack etc.).

7.3 Limitations

Our experiment has limitations, and we will try to address them in future work on larger datasets. To deal with missing values, we used the single imputation method of using the median value for missing variables. Given the fact there was only a small percentage of missing values (only variables with less than 10% missing values were used in experiment) and that these variables did not follow the normal distribution, the single imputation method using the median value would be an appropriate method. Even though Jiang et al. [51] used mean imputation to impute missing data in features prior to performing unsupervised clustering on heart failure dataset, this method may be seen as limitation that that we have not used more sophisticated methods such as kernel density, IDW, K-nearest neighbours to deal with missing values.

In our analysis, we used PCA, which is not free from disadvantages. According to Dormann et al. [52], PCA presumes a multinormal distribution of data and does not cope well with outliers. Due to the nature of the clinical data, we dealt with a dataset that has a multivariate distribution. It has been suggested, however, that in practice, PCA is a relatively robust technique if it is used for continuous variables that are not strongly skewed and does not have many outliers. [52]. In future work, we will explore factor analysis as a method for dimensionality reduction as in contrast to PCA, factor analysis is performed on mutual variance (i.e. shared variance) of observed variables. In PCA, however, all the variances in the given variables are taken into consideration and contribute to the end result [38]. Another disadvantage of PCA is the fact that all components in PCA are the aggregates of correlated variables and they all co-produce a particular component. However, in the case of factor analysis, a factor carries information about the processes contributing to the production of correlations between variables that contribute to each factor [38]. Another limitation of our study is related to the ML method that we selected to perform the clustering experiment. K-means clustering is a popular method; however, we should perhaps experiment using the k-medoids clustering method. k-medoids is a partitioning method that is best suited for domains requiring robustness to outliers, inconsistent distance metrics, or the dataset with no clear definition of mean or median [53]. The k-medoids algorithm returns medoids which are the actual data points in the dataset. This facilitates the use the algorithm in situations where the mean of the data does not exist within the data set. This is the main difference between k-medoids and k-means where the centroids returned by k-means may not be within the data set. Hence k-medoids is useful for clustering categorical data where a mean is impossible to define or interpret.

8 Conclusions

During this experiment, we demonstrated that the k-means clustering algorithm identified groups of HF patients with distinct features at the physiological level (as evidenced by median blood test results, ECHO findings and clinical observations). The findings at the physiology level were likely to be the accurate reflection of the 'labels' given by medical diagnoses as documented for each patient in the dataset. The data-driven approach that utilised PCA seemed more accurate in identifying smaller clusters with distinct features at the physiological level. During this experiment, we compared how domain-led feature selection compares to the data-driven approach. From one perspective, the data-driven approach had the advantage over the domain-led approach for feature extraction as it removed a risk of bias that can be introduced by humans (domain experts). The domain-led approach may potentially prohibit knowledge discovery that can be hidden behind features that are not routinely taken into consideration by physicians as important variables. The domain knowledge played an important role at the interpretation stage of the clustering experiment providing insight into the context and preventing far fetched conclusions. Having carried out this experiment, we have realised the importance of ensuring that the data scientist has appreciated the domain knowledge and fully understood the associated concepts. Therefore the future work may include a framework that would ensure that the data scientist understand the domain knowledge. For example, Delphie Technique could be used alongside a group of experts to form a consensus on what concepts and knowledge would need to be fully understood for data scientist to carry out the domain-led data mining project. Once that consensus is formed, those concepts then can be delivered in a form of training and there will need to be some form of assessment to ensure that knowledge exchange has successfully taken place. This kind of work is much needed because it would provide a consistency across domain-led data mining and would also help reduce the possibility of data scientist misunderstanding concepts and knowledge from the application area. We propose a checklist of questions that should prompt data scientist to actively seek the involvement of domain knowledge expert. This checklist can be further improved and become an agile document, updated when new concepts to enhancing domain-driven data mining arise. Moving forward embedding domain experts in the data analytics process will not only enhance the accuracy of conclusions but will be core to closing gaps in between academic data scientists and clinicians.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s41060-022-00346-9.

Author Contributions All authors contributed to the conception and design. Data analysis was performed by A. J-P with input from P. B, R. B, D. M and P. C. The first draft of the manuscript was written by A. J-P. The draft has been critically revised by R. B, P. B, R. B, P. C, F. B and D. M. All authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding Dr Jasinska-Piadlo was awarded a Doctoral Fellowship Award by Public Health Agency and Research and Development Department of the Health and Social Care in Northern Ireland, UK.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Reddy, C.K., Aggarwal, C.C.: Healthcare Data Analytics, vol. 36. CRC Press, Boca Raton (2015)
2. Kopanas,I., Avouris,N.M., Daskalaki,S.: in *Hellenic Conference on Artificial Intelligence* (Springer, 2002), pp. 288–299
3. Nagamine, T., Gillette, B., Pakhomov, A., Kahoun, J., Mayer, H., Burghaus, R., Lippert, J., Saxena, M.: Multiscale classification of heart failure phenotypes by unsupervised clustering of unstructured electronic medical record data. Sci. Rep. **10**(1), 1 (2020)
4. Gu, J., Pan, J.A., Lin, H., Zhang, J.F., Wang, C.Q.: Characteristics, prognosis and treatment response in distinct phenogroups of heart failure with preserved ejection fraction. Int. J. Cardiol. **323**, 148 (2021)
5. Schrub, F., Oger, E., Bidaut, A., Hage, C., Charton, M., Daubert, J.C., Leclercq, C., Linde, C., Lund, L., Donal, E.: Heart failure with preserved ejection fraction: a clustering approach to a heterogenous syndrome. Arch. Cardiovasc. Dis. **113**(6–7), 381 (2020)
6. Kaptein, Y.E., Karagodin, I., Zuo, H., Lu, Y., Zhang, J., Kaptein, J.S., Strande, J.L.: Identifying Phenogroups in patients with subclinical diastolic dysfunction using unsupervised statistical learning. BMC Cardiovasc. Disord. **20**(1), 1 (2020)
7. Gevaert,A.B., Tibebu,S., Mamas,M.A., Ravindra,N.G., Lee,S.F., Ahmad, T., Ko,D.T., Januzzi Jr, J.L., Van Spall,H.G.: Clinical phenogroups are more effective than left ventricular ejection fraction categories in stratifying heart failure outcomes, ESC Heart Failure (2021)
8. Ahmad, T., Lund, L.H., Rao, P., Ghosh, R., Warier, P., Vaccaro, B., Dahlström, U., O'connor, C.M., Felker, G.M., Desai, N.R.: Machine learning methods improve prognostication, identify clinically distinct phenotypes, and detect heterogeneity in response to

therapy in a large cohort of heart failure patients. J. Am. Heart Assoc. **7**(8), e008081 (2018)
9. Cao, L., Zhang, C.: Domain-driven data mining: a practical methodology. Int. J. Data Warehous. Min. (IJDWM) **2**(4), 49 (2006)
10. Cao, L., Zhang, C., Yang, Q., Bell, D., Vlachos, M., Taneri, B., Keogh, E., Philip, S.Y., Zhong, N., Ashrafi, M.Z., et al.: Domain-driven, actionable knowledge discovery. IEEE Intell. Syst. **22**(4), 78 (2007)
11. Yu, P., Zhang, C., Williams, G., Cao, L.: in *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining* (2007), p. 1
12. Jasinska-Piadlo, A., Bond, R., Biglarbeigi, P., Brisk, R., Campbell, P., McEneaney, D.: What can machines learn about heart failure? A systematic literature review, Int. J. Data Sci. Anal. (2021)
13. Ben-Assuli, O., Heart, T., Shlomo, N., Klempfner, R.: Bringing big data analytics closer to practice: a methodological explanation and demonstration of classification algorithms. Health Policy Technol. **8**(1), 7 (2019)
14. Blackstone, E.H., Rajeswaran, J., Cruz, V.B., Hsich, E.M., Koprivanac, M., Smedira, N.G., Hoercher, K.J., Thuita, L., Starling, R.C.: Continuously updated estimation of heart transplant waitlist mortality. J. Am. Coll. Cardiol. **72**(6), 650 (2018)
15. Ben-Assuli, O., Heart, T., Vest, J.R., Ramon-Gonen, R., Shlomo, N., Klempfner,R.: Profiling Readmissions Using Hidden Markov Model-the Case of Congestive Heart Failure, Information Systems Management pp. 1–13 (2020)
16. Frizzell, J.D., Liang, L., Schulte, P.J., Yancy, C.W., Heidenreich, P.A., Hernandez, A.F., Bhatt, D.L., Fonarow, G.C., Laskey, W.K.: Prediction of 30-day all-cause readmissions in patients hospitalized for heart failure: comparison of machine learning and other statistical approaches. JAMA Cardiol. **2**(2), 204 (2017)
17. Taslimitehrani, V., Dong, G., Pereira, N.L., Panahiazar, M., Pathak, J.: Developing EHR-driven heart failure risk prediction models using CPXR (Log) with the probabilistic loss function. J. Biomed. Inform. **60**, 260 (2016)
18. Liu, R., Zolfaghar, K., Chin, S.c., Roy, S.B. , Teredesai, A.: in *2014 IEEE International Conference on Data Mining* (IEEE, 2014), pp. 911–916
19. Xiao, C., Ma, T., Dieng, A.B., Blei, D.M., Wang, F.: Readmission prediction via deep contextual embedding of clinical concepts. PloS One **13**(4), e0195024 (2018)
20. Saqlain, M., Athar, A., Saqib, N.A., Khan, M.A.: Developing a classification model for an effective treatment of heart failure. Int. J. Comput. Sci. Inf. Secur. **14**(8), 413 (2016)
21. Sideris, C., Pourhomayoun, M., Kalantarian, H., Sarrafzadeh, M.: A flexible data-driven comorbidity feature extraction framework. Comput. Biol. Med. **73**, 165 (2016)
22. Sun, J., Hu, J., Luo, D., Markatou, M., Wang, F., Edabollahi, S., Steinhubl,S.E., Daar, Z., Stewart, W.F.: in *AMIA Annual Symposium Proceedings*, vol. 2012 (American Medical Informatics Association, 2012), vol. 2012, p. 901
23. Zhang, Z., Cao, L., Zhao, Y., Xu, Z., Chen, R., Lv, L., Xu, P.: Hospitalized patients with heart failure: integrating electronic healthcare records and external outcome data,
24. Goldberger, A.L., Amaral, L.A., Glass, L., Hausdorff, J.M., Ivanov, P.C., Mark, R.G., Mietus, J.E., Moody, G.B., Peng, C.K., Stanley, H.E.: PhysioBank. PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. Circulation **101**(23), e215 (2000)
25. Ponikowski, P., Voors, A.A., Anker, S.D., Bueno, H., Cleland, J.G., Coats, A.J., Falk, V., González-Juanatey, J.R., Harjola, V.P. , Jankowska, E.A.: et al., ESC Scientific Document Group. 2016 ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure: The Task Force for the diagnosis and treatment of acute and chronic heart failure of the European Society of Cardiology (ESC) Developed with the special contribution of the Heart Failure Association (HFA) of the ESC, Eur Heart J **37**(27), 2129 (2016)
26. MATLAB, *version R2021b*. Natick, Massachusetts (2021)
27. McDonagh, T.A., Metra, M., Adamo, M., Gardner, R.S., Baumbach, A., Böhm, M., Burri, H., Butler, J., Čelutkienė, J., Chioncel, O., et al.: 2021 ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure: developed by the Task Force for the diagnosis and treatment of acute and chronic heart failure of the European Society of Cardiology (ESC) With the special contribution of the Heart Failure Association (HFA) of the ESC. Eur. Heart J. **42**(36), 3599 (2021)
28. MATLAB. Quantile-quantile plot (2021). https://uk.mathworks.com/help/stats/qqplot.html
29. MATLAB. One-sample kolmogorov-smirnov test (2021). https://uk.mathworks.com/help/stats/kstest.html?s_tid=srchtitle_kstest_1
30. Sato, N., Gheorghiade, M., Kajimoto, K., Munakata, R., Minami, Y., Mizuno, M., Aokage, T., Asai, K., Sakata, Y., Yumino, D., et al.: Hyponatremia and in-hospital mortality in patients admitted for heart failure (from the ATTEND registry). Am. J. Cardiol. **111**(7), 1019 (2013)
31. Bettari, L., Fiuzat, M., Felker, G.M., O'Connor, C.M.: Significance of hyponatremia in heart failure. Heart Fail. Rev. **17**(1), 17 (2012)
32. Seferovic, P.M., Ponikowski, P., Anker, S.D., Bauersachs, J., Chioncel, O., Cleland, J.G., de Boer, R.A., Drexel, H., Ben Gal, T., Hill, L.: et al., Clinical practice update on heart failure 2019: pharmacotherapy, procedures, devices and patient management. An expert consensus meeting report of the Heart Failure Association of the European Society of Cardiology, European journal of heart failure **21**(10), 1169 (2019)
33. Osman, M., Syed, M., Balla, S., Kheiri, B., Faisaluddin, M., Bianco, C.: A Meta-analysis of Intravenous Iron Therapy for Patients with Iron Deficiency and Heart Failure., Am. J. Cardiol. (2020)
34. Avni, T., Leibovici, L., Gafter-Gvili, A.: Iron supplementation for the treatment of chronic heart failure and iron deficiency: systematic review and meta-analysis. Eur. J. Heart Fail. **14**(4), 423 (2012)
35. Núñez, J., Miñana, G., Santas, E., Bertomeu-González, V.: Cardiorenal syndrome in acute heart failure: revisiting paradigms. Revista Española de Cardiología (English Edition) **68**(5), 426 (2015)
36. Horwich, T.B., Kalantar-Zadeh, K., MacLellan, R.W., Fonarow, G.C.: Albumin levels predict survival in patients with systolic heart failure. Am. Heart J. **155**(5), 883 (2008)
37. Shikdar, S., Vashisht, R., Bhattacharya, P.T.: International normalized ratio (INR), (2018)
38. Tabachnick, B.G., Fidell, L.S., Ullman, J.B.: Using Multivariate Statistics, vol. 5. Pearson Boston, MA (2007)
39. Holland, S.M.: Principal components analysis (PCA), Department of Geology, University of Georgia, Athens, GA pp. 30,602–2501 (2008)
40. Solorio-Fernández, S., Carrasco-Ochoa, J.A., Martínez-Trinidad, J.F.: A new hybrid filter-wrapper feature selection method for clustering based on ranking. Neurocomputing **214**, 866 (2016)
41. McKinsey. Transforming healthcare with ai: The impact on the workforce and organisations (2020). https://eithealth.eu/wp-content/uploads/2020/03/EIT-Health-and-McKinsey_Transforming-Healthcare-with-AI.pdf
42. Topol ,E.: The topol review: preparing the healthcare workforce to deliver the digital future (2019). https://www.hee.nhs.uk/our-work/topol-review
43. England, H.E.: The topol programme for digital fellowships in healthcare (2019). https://topol.hee.nhs.uk/digital-fellowships/
44. Wang, Z., Yao, L., Li, D., Ruan, T., Liu, M., Gao, J.: International journal of medical informatics **115**, 10 (2018)

45. Escamilla, A.K.G., El Hassani, A.H., Andres, E.: Dimensionality Reduction in Supervised Models-based for Heart Failure Prediction, (2019)
46. Boutsidis, C., Drineas, P., Mahoney, M.W.: in *Advances in Neural Information Processing Systems* (2009), pp. 153–161
47. Iglesias, F., Zseby, T., Zimek, A.: Clustering refinement. Int. J. Data Sci. Anal. **12**(4), 333 (2021)
48. Li, Z., Zou, D., Ma, X., Chen, J., Shi, X., Gong, Y., Man, X., Gao, L., Zhao, Y., Wang, R., et al.: Epidemiology of peptic ulcer disease: endoscopic results of the systematic investigation of gastrointestinal disease in China. Off. J. Am. Coll. Gastroenterol. ACG **105**(12), 2570 (2010)
49. T.R.S. (Charity), Dynamics of data science skills: how can all sectors benefit from data science talent?, (2019)
50. Mao, Y., Wang, D., Muller, M., Varshney, K.R., Baldini, I., Dugan, C., Mojsilović, A.: How data scientistswork together with domain experts in scientific collaborations: to find the right answer or to ask the right question?, Proceedings of the ACM on Human-Computer Interaction **3**(GROUP), 1 (2019)
51. W. Jiang, S. Siddiqui, S. Barnes, L.A. Barouch, F. Korley, D.A. Martinez, M. Toerper, S. Cabral, E. Hamrock, S. Levin, Readmission risk trajectories for patients with heart failure using a dynamic prediction approach: retrospective study, JMIR Medical Informatics **7**(4) (2019)
52. Dormann, C.F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., Marquéz, J.R.G., Gruber, B., Lafourcade, B., Leitão, P.J., et al.: Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. Ecography **36**(1), 27 (2013)
53. Lopes, H.E.G., Gosling, M.d.S.: Cluster analysis in practice: dealing with outliers in managerial research, Revista de Administração Contemporânea **25** (2020)