Data-driven machine-learning analysis of potential embolic sources in embolic stroke of undetermined source

G. Ntaiosa,*, S. F. Wengb,c,*, K. Perlepea, R. Akyeac, L. Conдонc, D. Lambroua, G. Sirimarcod, D. Strambod, A. Eskandari, E. Karagkiozia, A. Vemmoue, E. Korompokie, E. Manios, K. Makaritsisa, K. Vemmose and P. Michelid

Department of Internal Medicine, Faculty of Medicine, School of Health Sciences, University of Thessaly, Larissa, Greece; bNational Institute for Health Research School for Primary Care Research, University of Nottingham, Nottingham, UK; cPrimary Care Stratified Medicine, Division of Primary Care, School of Medicine, University of Nottingham, Nottingham, UK; dStroke Center and Neurology Service, Department of Clinical Neurosciences, Centre Hospitalier Universitaire Vaudois and University of Lausanne, Lausanne, Switzerland; eDepartment of Clinical Therapeutics, Medical School of Athens, Alexandra Hospital, Athens, Greece; and fDivision of Brain Sciences, Department of Stroke Medicine, Imperial College, London, UK

Keywords: embolic stroke of undetermined source, hierarchical clustering, machine learning, potential embolic source, stroke

Background and purpose: Hierarchical clustering, a common ‘unsupervised’ machine-learning algorithm, is advantageous for exploring potential underlying aetiology in particularly heterogeneous diseases. We investigated potential embolic sources in embolic stroke of undetermined source (ESUS) using a data-driven machine-learning method, and explored variation in stroke recurrence between clusters.

Methods: We used a hierarchical k-means clustering algorithm on patients’ baseline data, which assigned each individual into a unique clustering group, using a minimum-variance method to calculate the similarity between ESUS patients based on all baseline features. Potential embolic sources were categorised into atrial cardiopathy, atrial fibrillation, arterial disease, left ventricular disease, cardiac valvulopathy, patent foramen ovale (PFO) and cancer.

Results: Among 800 consecutive ESUS patients (43.3% women, median age 67 years), the optimal number of clusters was four. Left ventricular disease was most prevalent in cluster 1 (present in all patients) and perfectly associated with cluster 1. PFO was most prevalent in cluster 2 (38.9% of patients) and associated significantly with increased likelihood of cluster 2 [adjusted odds ratio: 2.69, 95% confidence interval (CI): 1.64–4.41]. Arterial disease was most prevalent in cluster 3 (57.7%) and associated with increased likelihood of cluster 3 (adjusted odds ratio: 2.21, 95% CI: 1.43–3.13). Atrial cardiopathy was most prevalent in cluster 4 (100%) and perfectly associated with cluster 4. Cluster 3 was the largest cluster involving 53.7% of patients. Atrial fibrillation was not significantly associated with any cluster.

Conclusions: This data-driven machine-learning analysis identified four clusters of ESUS that were strongly associated with arterial disease, atrial cardiopathy, PFO and left ventricular disease, respectively. More than half of the patients were assigned to the cluster associated with arterial disease.

Introduction

Approximately 17% of all ischemic stroke patients have an embolic stroke of undetermined source (ESUS) (i.e. a stroke without an apparent cause despite recommended diagnostic work-up) [1]. Numerous underlying pathologies may serve as embolic sources in patients with ESUS such as atherosclerotic plaques in the
carotids and the aortic arch, covert atrial fibrillation (AF), patent foramen ovale (PFO), left ventricular (LV) disease, atrial cardiopathy, cancer and cardiac valvular disease [1]. Recently, we showed that there is significant overlap of potential embolic sources (PESs) in patients with ESUS. In a cohort of consecutive ESUS patients, 65.5% had two or more PESs and 31.1% had three or more PESs, whereas on average, each patient had two PESs [2]. In this context, it is frequently difficult to identify the actual source of embolism in an ESUS patient, when several PESs coexist [3].

Clustering algorithms, a common ‘unsupervised’ machine-learning, can be used to identify groups (clusters) of similar individuals based on the sum of the combined values of their measured characteristics [4]. In hierarchical clustering, the results are easily reproducible, and this process is fixed once clusters are assigned, so participants cannot be reclassified into a different cluster. This contrasts with standard regression methods, which are used to identify associations between response and explanatory variables. This belongs to supervised learning, which can be used for multiple testing to determine significant differences between groups that need to be specified a priori. Each test is independent of the other tests, which results in groups that are only relevant to the particular variable tested. Clustering takes into account all variables, providing a way to holistically represent the entirety of the data collected [5]. Therefore, this process is extremely advantageous for exploring the potential underlying aetiology in particularly heterogeneous diseases such as ESUS.

In this context, we investigated the potential sources of embolism in ESUS patients using a data-driven, machine-learning analytical method and explored variation in rates of stroke recurrence between clusters.

Methods

Patient population

We analysed complete data from consecutive patients with ESUS recruited in three prospective stroke registries: ASTRAL (Acute Stroke Registry and Analysis of Lausanne), Athens Stroke Registry, and Larissa Stroke Registry [6–8]. A standard pro forma template was used to collect all clinical, demographic, biometric, biomarkers and outcome data. The use of the registry data for research was approved by local institutional review boards, and the study is registered at ClinicalTrials.gov (NCT02766205). Full descriptive study procedures and methods were previously published [2].

The definition of ESUS was based on the Cryptogenic Stroke/ESUS International Working Group criteria: non-lacunar brain infarct in the absence of (i) extracranial atherosclerosis causing ≥50% luminal stenosis in arteries supplying the area of ischemia or (ii) major-risk cardioembolic source or (iii) any other specific cause of stroke (e.g. arthritis, dissection, migraine/vasospasm or drug misuse) [1].

Patient features included in the analysis and methodology of cluster generation

The clustering methods utilised all baseline features detailed in the Supplementary Table S1, which included demographics, lifestyle factors, clinical symptoms/signs during the qualifying ESUS, comorbidities, biometrics, biomarkers, vascular imaging, brain imaging, electrocardiogram and echocardiography.

To identify groups of patients with similar characteristics (i.e. clusters), we used a combined k-means and hierarchical agglomerative approach to generate clusters, called hierarchical k-means clustering [9]. This process allows for the k-means-based approach to accelerate or speed up a traditional k-means algorithm in both training and query phases, which allows for a much larger number of centroids to be used, which in turn leads to much better learning [9]. In this process, we pick some k to be the branching factor, which defines the number of clusters at each level of the clustering hierarchy. We then cluster the set of points into k clusters using a standard k-means algorithm. Finally, we recursively clustered each subcluster until we determine a small fixed number of points. Therefore, using all the baseline data provided from ESUS patients, the algorithm could assign each individual into a unique cluster.

To determine the optimum number of clusters, we used a combined approach using 30 different clustering indices, which includes common methods including ‘elbow’, ‘average silhouette’, or ‘gap statistics’. The optimal number of clusters was determined from the highest frequency of selection from all 30 indices [10]. To visualize the clustering process, we generated a dendrogram (a tree diagram) to illustrate the arrangement of the clusters produced [11]. Each branching creates a unique participant cluster, with the size of the clusters determined by the height of the branches. Separately, we also conducted a principle components analysis by plotting the first two principle components on a coordinate to observe the clusters between each ESUS patient by his/her respective assigned cluster group. These principal components were derived using the orthogonal transformation (eigenvectors and eigenvalues) to reduce down the dimensionality of the original data, from all the clinical features collected on ESUS patients.
analyses and data visualisation tools were conducted using statistical software R (The R Foundation for Statistical Computing, Vienna, Austria) using packages cluster, NbClust, factoextra, dendxtend and ggplot2.

**Description of clusters: summary characteristics and prevalence of potential embolic sources**

Descriptive characteristics of each cluster were provided, reporting number (%) and median [interquartile range (IQR)] for categorical and continuous variables, respectively. We further profiled each cluster by determining the prevalence of each PES within each cluster. Patients were also categorised by the number of PESs: zero to one PES, two PESs, or three or more PESs.

Potential embolic sources were categorised as follows: atrial cardiopathy, AF, arterial disease, LV disease, cardiac valvular disease, PFO, and cancer, as previously described in detail [2]. In particular, based on previously published associations with the risk of stroke, atrial cardiopathy was diagnosed if the echocardiogram reported left atrial dilatation or increased left atrial diameter (>38 mm for women and >40 mm for men) [12,13], or if supraventricular extrasystoles were present at the 12-lead electrocardiograms performed during hospitalization [14,15]. We diagnosed arterial disease in case of presence of any ipsilateral atherosclerotic carotid plaque causing luminal stenosis of <50% [16–18] or aortic arch atherosclerosis [19–22] based on the imaging reports. We did not review the images. We did not include contralateral carotid atherosclerosis in this PES. LV disease was diagnosed if low LV ejection fraction (<35%) or LV hypertrophy or left-sided heart failure was reported at the echocardiogram, or if LV hypertrophy was identified at the electrocardiogram (Sokolow index ≥ 35 mm) [23]. We diagnosed cardiac valvular disease if moderate-to-severe stenosis or regurgitation of the mitral or aortic valve was reported at the echocardiogram. AF was assessed during on-site patient visits at the outpatient clinic and/or by contact with the patient and/or the next of kin or the patient’s primary physician; it was considered present if confirmed by an electrocardiogram performed for any reason including palpitations, irregular pulse on clinical examination, in-hospital surveillance or portable outpatient monitoring.

**Clinical end-points during follow-up**

We evaluated the risk of stroke recurrence over the 10 years of follow-up by cluster. Stroke during follow-up was ascertained by on-site patient visits at the outpatient’s clinic, contact with the patient’s next of kin, or the patient’s primary physician. Where possible, the outcome had been adjudicated by reviewing the patient’s medical notes and imaging outcomes.

**Statistical analysis**

Comparisons across clusters were conducted using the non-parametric Kruskal-Wallis test for continuous variables and $\chi^2$ tests for categorical variables [24,25]. Prior to the clustering analysis, data that were missing at random were imputed using multiple imputation using chained equations [26]. To quantify the contribution of each PES to each cluster, we applied logistic regression to determine the association between each PES with the derived cluster. In this analysis, the PES was the exposure variable, and the cluster grouping was the outcome variable (coded as 1, belonging to the cluster; or coded as 0, belonging to other clusters). All models were adjusted for sex, age, dyslipidaemias, diabetes mellitus, smoking, coronary artery disease, and National Institutes of Health Stroke Scale (NIHSS) score at admission. The PESs in each cluster were then ranked by significance and by the effect size, with 95% confidence intervals (CIs) provided. In this way, we were able to profile each cluster and associate them to specific PESs.

Subsequently, we evaluated the 10-year follow-up of stroke recurrence by cluster. Incidence rates (per 1000 person-years) and 95% CIs were provided. To obtain estimates for the association between cluster groups and stroke recurrence, we performed Cox proportional hazards regression analysis, with informative censoring of the survival time when patients were lost to follow-up or died. The cluster with the lowest event rate for stroke recurrence was used as the reference group.

Furthermore, we quantified the dose-response relationship of having multiple PESs compared to a single or no PES using Cox proportional hazards. Similar to the logistic regression analyses, all hazard ratios were adjusted for sex, age, hypertension, dyslipidaemia, diabetes mellitus, smoking, coronary artery disease, and NIHSS at admission. All hazard models were assessed for proportional hazards using Schoenfeld residuals. $P$ values <0.05 were considered statistically significant.

**Results**

A total of 800 ESUS patients (43.3% women) were included in the analysis. The median age of patients was 67 years (IQR = 54–77 years).
Visualization of the hierarchical clustering analysis

From 30 clustering indices, it was found that the optimal number of clusters is four (Fig. S1). The arrangement of the four clusters during the clustering process is illustrated at the dendrogram (Fig. S2).

The principal components analysis identified that 82% of all principal components were needed to explain 100% variation of the original ESUS data (Fig. S3a), which suggests that there is substantial heterogeneity between ESUS patients in clinical features, as a high number of principal components are needed to explain significant variation of the original data. By plotting the first two principal components, which only explains up to 16% of the variation in the original data, visual separation can be seen between clusters from the hierarchical clustering process (Fig. S3b). Cluster sizes were as follows: 44 patients (53.8%) in cluster 3, and 177 patients (5.5%) in cluster 1, 149 patients (18.6%) in cluster 2, 430 patients (57.7%) in cluster 4. There were overlap between cluster 1 and cluster 2. However, clusters 2, 3, and 4 all remained quite distinct, with a large degree of separation and very little overlap.

Characteristics of the cluster groups

The baseline characteristics of the patients in the four clusters are summarized in Table 1. There were significant differences between clusters in terms of gender, baseline age, NIHSS at admission, hypertension, diabetes mellitus, coronary artery disease, previous stroke and antithrombotic treatment at discharge.

The prevalence of each PES stratified by cluster is summarised in Table 2. There were significant differences between clusters in the prevalence of atrial fibrillation, atrial cardiopathy, arterial disease, LV disease, PFO and cancer. LV disease was most prevalent in cluster 1 (100%). PFO was most prevalent in cluster 2 (38.9%). Arterial disease was most prevalent in cluster 3 (57.7%). Atrial cardiopathy was most prevalent in cluster 4 (100%).

| Table 1 Baseline characteristics and outcomes of patients, stratified by cluster |
|-----------------------------------------------|-----------------|-----------------|-----------------|-----------------|-----------|
| Age, years Median (IQR)                       | 67 (62–75)      | 62.9 (46.7–74.6)| 66 (51.7–76.8) | 71 (64–79)      | <0.001    |
| Female sex n (%)                              | 14 (32.8)       | 51 (34.2)       | 205 (47.7)     | 75 (42.4)       | 0.013     |
| NIHSS score Median (IQR)                      | 9 (3–21)        | 3 (2–6)         | 7 (3–14)       | 6 (3–13)        | <0.001    |
| Hypertension n (%)                            | 30 (68.2)       | 80 (53.7)       | 250 (58.1)     | 136 (76.8)      | <0.001    |
| Dyslipidaemia n (%)                           | 26 (59.1)       | 92 (61.7)       | 294 (68.4)     | 113 (63.8)      | 0.319     |
| Diabetes mellitus n (%)                       | 19 (43.2)       | 21 (14.1)       | 66 (15.4)      | 42 (23.7)       | <0.001    |
| Smoking n (%)                                 | 16 (36.4)       | 53 (35.6)       | 183 (42.6)     | 57 (32.2)       | 0.087     |
| Coronary artery disease n (%)                 | 13 (29.6)       | 17 (11.4)       | 41 (9.53)      | 47 (26.6)       | <0.001    |
| Previous stroke n (%)                         | 3 (6.8)         | 21 (14.1)       | 83 (19.3)      | 21 (11.9)       | 0.039     |
| Antplatelet at discharge n (%)                | 32 (72.7)       | 137 (92.0)      | 402 (93.5)     | 161 (91.0)      | <0.001    |
| Anticoagulant at discharge n (%)              | 8 (18.2)        | 10 (6.7)        | 31 (7.2)       | 14 (7.9)        | <0.001    |
| Death at follow-up n (%)                      | 22 (50.0)       | 15 (10.1)       | 56 (13.0)      | 51 (28.8)       | <0.001    |
| Stroke recurrence n (%)                       | 3 (6.8)         | 14 (9.4)        | 45 (10.5)      | 39 (22.0)       | <0.001    |
| Stroke recurrence (event rate)a               | 21.7 (7.0–67.3) | 29.3 (17.3–49.5)| 29.5 (22.0–39.5)| 50.1 (36.6–68.6)| <0.001    |

aPer 1000 patient years. IQR, interquartile range; NIHSS, National Institute of Health Stroke Scale.

| Table 2 Prevalence of potential embolic sources and degree of their overlap stratified by cluster |
|-----------------------------------------------|-----------------|-----------------|-----------------|-----------------|-----------|
| PES sources                                   | Cluster 1 (n = 44) | Cluster 2 (n = 149) | Cluster 3 (n = 430) | Cluster 4 (n = 177) | P value  |
| Number of PES sources Median (IQR)            | 2 (2–3)         | 2 (1–3)         | 2 (1–3)         | 2 (2–3)         | <0.001   |
| Number with 2 PES sources n (%)               | 20 (45.5)       | 42 (28.2)       | 134 (31.2)      | 81 (45.8)       | <0.001   |
| Number with ≥3 PES sources n (%)              | 17 (38.6)       | 43 (28.7)       | 119 (27.7)      | 69 (39.0)       | <0.001   |
| Atrial fibrillation n (%)                     | 8 (18.2)        | 13 (8.7)        | 10 (9.3)        | 59 (33.3)       | <0.001   |
| Atrial cardiopathy n (%)                      | 32 (72.7)       | 48 (32.2)       | 103 (24.0)      | 177 (100)       | <0.001   |
| Arterial disease n (%)                        | 9 (20.5)        | 68 (45.6)       | 248 (57.7)      | 63 (35.6)       | <0.001   |
| Left ventricular disease n (%)                | 44 (100)        | 77 (51.7)       | 223 (51.9)      | 91 (51.4)       | <0.001   |
| Cardiac valvar disease n (%)                  | 6 (13.6)        | 13 (8.7)        | 41 (9.5)        | 9 (5.1)         | 0.198    |
| PFO n (%)                                      | 1 (2.3)         | 58 (38.9)       | 101 (23.5)      | 10 (5.7)        | <0.001   |
| Cancer n (%)                                   | 2 (4.6)         | 13 (8.7)        | 50 (11.6)       | 9 (5.1)         | 0.051    |

IQR, interquartile range; PES, potential embolic sources; PFO, patent foramen ovale.
Association between cluster grouping and PES

Using multivariable logistic regression models, we determined the association between each PES and cluster membership. The adjusted odds ratios and 95% CIs for each cluster are presented in Table 3. LV disease was perfectly associated with cluster 1 membership. PFO was significantly associated with increased likelihood of cluster 2 membership (adjusted odds ratio = 2.69, 95% CI: 1.64–4.41). Arterial disease was significantly associated with increased likelihood of cluster 3 membership (adjusted odds ratio = 2.21, 95% CI: 1.43–3.13). Atrial cardiopathy was perfectly associated with cluster 4 membership.

Risk of stroke recurrence across clusters

The mean and median follow-up duration was 3.7 years (SD = 3.7) and 2.1 years (IQR = 0.8–5.8 years). Over 2922 person-years, there were 101 recurrent strokes, corresponding to an overall rate of 34.6 per 1000 person-years (95% CI: 28.4–42.0). The risk of stroke recurrence was not different across clusters in adjusted models Table 4 and Fig. 1.

Discussion

This data-driven machine-learning analysis of consecutive ESUS patients identified four clusters of patients based on their baseline characteristics. The largest cluster, which included more than half of the overall population, was associated with the presence of arterial disease. Two clusters of medium size including approximately 15% to 20% of the overall population were associated with atrial cardiopathy and PFO, respectively. A small cluster, which included only 5% of the overall population, was associated with LV disease. Atrial fibrillation was not associated with any cluster. The risk of stroke recurrence was similar across clusters.

During recent years, there has been emerging evidence supporting an important etiological association between ESUS and atherosclerotic plaques. A recent analysis of the NAVIGATE-ESUS (New Approach Rivaroxaban Inhibition of Factor Xa in a Global Trial versus ASA to Prevent Embolism in Embolic Stroke of Undetermined Source) trial [27] as well as several other studies [28–40] showed that the prevalence of carotid plaques is higher ipsilateral to the infarct than contralateral in patients with ESUS. In addition, the AF-ESUS (Prediction of Atrial Fibrillation after Embolic Stroke of Undetermined Source) study showed that new incident AF is less frequently detected in patients with ESUS and carotid plaques compared to those without [18]. Similarly, in young adults with cryptogenic stroke, carotid plaques were associated with the absence of PFO [41]. Both latter

Table 3 Multivariable logistic regression to determine association and effect size for each potential embolic source and cluster membership

| Cluster | Odds ratio  | 95% CI | Association |
|---------|-------------|--------|-------------|
| Cluster 1 | | | |
| Left ventricular disease | Perfectly associated with cluster* | | Positive association |
| Arterial disease | 0.22 | 0.09–0.53 | Negative association |
| Atrial cardiopathy | 1.82 | 0.81–4.05 | No association |
| Cardiac valvular disease | 1.35 | 0.48–3.79 | No association |
| Atrial fibrillation | 0.85 | 0.33–2.18 | No association |
| Cancer | 0.53 | 0.11–2.48 | No association |
| PFO | 0.24 | 0.03–1.95 | No association |
| Cluster 2 | | | |
| PFO | 2.69 | 1.64–4.41 | Positive association |
| Atrial fibrillation | 0.65 | 0.34–1.28 | No association |
| Cardiac valvular disease | 1.49 | 0.76–2.94 | No association |
| Left ventricular disease | 1.17 | 0.76–1.81 | No association |
| Arterial disease | 1.16 | 0.72–1.84 | No association |
| Cancer | 1.14 | 0.59–2.23 | No association |
| Atrial cardiopathy | 0.67 | 0.43–1.03 | No association |
| Cluster 3 | | | |
| Arterial disease | 2.12 | 1.43–3.13 | Positive association |
| Atrial fibrillation | 0.14 | 0.10–0.20 | Negative association |
| Cancer | 1.63 | 0.90–2.96 | No association |
| Cardiac valvular disease | 1.62 | 0.91–2.90 | No association |
| Atrial fibrillation | 0.88 | 0.53–1.46 | No association |
| Left ventricular disease | 0.84 | 0.58–1.21 | No association |
| PFO | 0.69 | 0.44–1.11 | No association |
| Cluster 4 | | | |
| Atrial cardiopathy | Perfectly associated with cluster* | | Positive association |
| Left ventricular disease | 0.38 | 0.23–0.63 | Negative association |
| Cardiac valvular disease | 0.32 | 0.14–0.72 | Negative association |
| Atrial fibrillation | 1.46 | 0.83–2.55 | No association |
| Arterial disease | 0.63 | 0.37–1.08 | No association |
| Cancer | 0.47 | 0.20–1.11 | No association |
| PFO | 0.47 | 0.20–1.13 | No association |

The adjusted odds ratios and 95% CIs for each cluster are ranked by significance and effect size. The regression model has been adjusted for sex, age, hypertension, dyslipidaemia, diabetes mellitus, smoking, coronary artery disease and National Institute of Health Stroke Scale score at admission. *One hundred percent of individuals within the cluster had the condition. CI, confidence interval; PES, potential embolic source.
studies show that carotid plaques act as a competing stroke aetiology to other established stroke aetiologies, and hence, support their role as an underlying cause of ESUS. Moreover, a recent analysis of consecutive emboli retrieved during mechanical thrombectomy showed that the emboli from patients with large artery atherosclerotic and cryptogenic strokes had a similar proportion of platelet-rich clots, which was significantly higher compared with thrombi from patients with cardioembolic stroke [42]. The results of the present study provide further arguments in support of an important association between ESUS and atherosclerotic plaques.

The concept of atrial cardiopathy has emerged during the recent years as an important source of embolism in patients with ESUS [43]. There is growing body of evidence indicating that thrombi may be formed in the diseased left atrium, even in the absence of atrial fibrillation. Atrial cardiopathy has been assessed in various ways using several indices including biomarkers [44–46], cardiac magnetic resonance imaging [47] and electrocardiographic indices [48–50]. The present analysis adds to the evidence that supports an important causative role of atrial cardiopathy in ESUS and indicates that atrial cardiopathy could be the cause of stroke in 15% to 20% of the overall ESUS population. The ongoing ARCADIA trial (AtRial Cardiopathy and Antithrombotic Drugs In Prevention After Cryptogenic Stroke) currently investigates whether patients with ESUS and atrial cardiopathy respond better to apixaban compared to aspirin for secondary stroke prevention [51].

Although older randomized trials were neutral [52], several recent randomized trials showed that percutaneous PFO closure is associated with a large reduction of recurrent stroke rates in patients with ESUS [53], supporting an important etiological association between PFO and ESUS. The results of our analysis are in line with this, as we identified a cluster of patients (18% of the overall cohort) that is associated with PFO.

Several observational studies and randomized trials showed that AF can be detected in 30% of ESUS patients during follow-up, suggesting a strong causal association between AF and ESUS [54–58]. However, there has been emerging evidence questioning the strength of this association, especially for short-lasting episodes detected remotely after ESUS [59]. The rate of AF detection in ESUS patients was similar with

| Table 4 Multivariable regression analysis of the association between the phenotype clusters and stroke recurrence |
|---------------------------------------------------------------|
| Hazard ratio | 95% CI |
| Cluster 1 | Reference | Reference |
| Cluster 2 | 1.57 | 0.43–5.72 |
| Cluster 3 | 1.41 | 0.42–4.72 |
| Cluster 4 | 2.14 | 0.65–7.07 |

The association has been adjusted for sex, age, hypertension, dyslipidaemia, diabetes mellitus, smoking, coronary artery disease, and National Institute of Health Stroke Scale score at admission. CI, confidence interval.

Figure 1 Ten-year survival estimates of stroke recurrence in patients with embolic stroke of undetermined source, according to the assigned phenotype clusters. IR, incidence rate. [Colour figure can be viewed at wileyonlinelibrary.com]
other non-ESUS stroke patients [56], as well as with older patients without previous stroke [60]. In addition, ESUS patients are phenotypically different compared with stroke patients with AF, with the former being younger with milder strokes [57,58,61,62]. Moreover, the majority of ischemic strokes do not occur proximal to recent episodes of atrial tachycardia or AF, as shown in patients with implantable cardiac monitoring devices in the ASSERT (Asymptomatic Atrial Fibrillation and Stroke Evaluation in Pacemaker Patients and the Atrial Fibrillation Reduction Atrial Pacing Trial) [63] and TRENDS (A Prospective Study of the Clinical Significance of Atrial Arrhythmias Detected by Implanted Device Diagnostics) [64] studies. Finally, if the association between AF and ESUS was strong, direct oral anticoagulants would have probably reduced stroke recurrence rates compared with aspirin, in line with the 55% reduction in stroke risk conferred by apixaban versus aspirin reported in the AVERROES (Apixaban Versus Acetylsalicylic Acid to Prevent Stroke in Atrial Fibrillation Patients Who Have Failed or Are Unsuitable for Vitamin K Antagonist Treatment) trial [65]. However, this was not the case in the NAVIGATE-ESUS and RE-SPECT ESUS (Dabigatran Etxilate for Secondary Stroke Prevention in Patients With Embolic Stroke of Undetermined Source) trials, in which patients assigned to direct oral anticoagulants and aspirin had similar recurrence rates [61,62]. The present analysis adds to the aforementioned evidence that support the argument that AF is not so strongly associated with ESUS as it was initially believed.

The main strength of this study is its design: the data-driven hierarchical-clustering analysis allowed the categorization of patients into distinct clusters based on their baseline characteristics, without pre-specification of variables, and then coupling of these clusters with PES. This is a particularly advantageous method in cases of datasets with large degree of heterogeneity between individuals. The categorization of patients into clusters rather than PESs is advantageous and more informative, as there is large overlap of PESs in patients with ESUS. For example, a previous analysis in the same cohort showed that LV disease was present in 54.4% of the overall cohort [2]. However, the present analysis showed that the cluster, which was associated with LV disease, included only 5% of the overall cohort. This suggests that for the majority of patients with LV disease, this would represent an innocent bystander rather than the actual embolic source. On the other hand, our study is limited by the risk of registration bias within and between the participating registries and differences in the workup of patients during the in-hospital phase. Finally, the clustering algorithms are empirical methods, which also may be limited by the sample size of the data and number of clinical features collection to determine cluster associations, as the analysis was not specifically powered to determine potential associations with future outcomes. Future research should explore whether these findings are consistent in a much larger sample of ESUS patients.

**Conclusion**

This data-driven, machine-learning, hierarchical-clustering analysis identified four clusters of ESUS patients that were associated with arterial disease, atrial cardiopathy, PFO and LV disease, respectively. AF was not associated with any cluster. The risk of stroke recurrence was not different across clusters.

**Disclosure of conflicts of interest**

G.N.: received through his institution a research grant for the Prediction of AF in ESUS (AF-ESUS) study (ClinicalTrials.gov Identifier: NCT02766205), which is an investigator-initiated study supported by Pfizer through the BMS/Pfizer European Thrombosis Investigator Initiated Research Program (ERISTA). Speaker fees/advisory boards/research support from Amgen, Bayer, BMS/Pfizer, Boehringer-Ingelheim, Elpen, Galenica, Sanofi, Winmedica. All fees are paid to his institution (University of Thessaly). KP.: received travel support by Pfizer. G.S.: received research grant from Swiss Heart Foundation, congress travel support from Bayer and Shire, and served on scientific advisory boards for Amgen and Daiichi-Sankyo. All fees are paid to his institution (CHUV). E.K: Speaker fees/honoraria for advisory boards from Pfizer, Amgen. P.M.: received within the last 2 years through his institution research grants from the Swiss Heart Foundation and BMS; speaker fees from Boehringer-Ingelheim, Medtronic and Amgen; consulting fees from Medtronic and Amgen, and honoraria from scientific advisory boards from Boehringer-Ingelheim, Pfizer and BMS. All this support goes to his institution (CHUV) and is used for stroke education and research. S.W.: Member of the Clinical Practice Research Datalink (CPRD) Independent Scientific Advisory Committee (ISAC), academic advisor to Quealth Ltd. and has received honorarium from AMGEN. All other authors: none.

This study was performed in the dataset of the Prediction of AF in ESUS (AF-ESUS) study (ClinicalTrials.gov Identifier: NCT02766205), which was an investigator-initiated study supported by Pfizer through the BMS/Pfizer European Thrombosis Investigator Initiated Research Program. The Swiss Cardiology Foundation supported data collection in the ASTRAL (Acute STroke Registry and Analysis of Lausanne) registry.
Data availability statement
The data that support the findings of this study are available from the corresponding author upon reasonable request.

Supporting Information
Additional Supporting Information may be found in the online version of this article:

Figure S1. Determining the optimal number of clusters by frequency among 30 clustering indices.

Figure S2. To visualise the clustering process, we generated a dendrogram (tree diagram) to illustrate the arrangement of the clusters produced the hierarchical k-means clustering analysis.

Figure S3. Principal components analysis of collected baseline clinical features from 800 ESUS patients.

Table S1. Baseline characteristics of patients.

References
1. Hart RG, Diener HC, Coutts SB, et al. Embolic strokes of undetermined source: the case for a new clinical construct. Lancet Neurol 2014; 13: 429–438.
2. Ntaios G, Perlepe K, Lambrou D, et al. Prevalence and overlap of potential embolic sources in patients with embolic stroke of undetermined source. J Am Heart Assoc 2019; 8: e012858.
3. Ntaios G, Hart RG. Embolic stroke. Circulation 2017; 136: 2403–2405.
4. Kimes PK, Liu Y, Neil Hayes D, Marron JS. Statistical significance for hierarchical clustering. Biometrics 2017; 73: 811–821.
5. Murtagh F, Contreras P. Algorithms for hierarchical clustering: an overview. WIRES Data Min Knowl Discov 2011; 2: 86–97.
6. Michel P, Odier C, Rutgers M, et al. The Acute StROKE Registry and Analysis of Lausanne (ASTRAL): design and baseline analysis of an ischemic stroke registry including acute multimodal imaging. Stroke 2010; 41: 2491–2498.
7. Vemmkos K, Ntaios G, Spengos K, et al. Association between obesity and mortality after acute first-ever stroke: the obesity-stroke paradox. Stroke 2011; 42: 30–36.
8. Ntaios G, Vemmkos K, Lip GY, et al. Risk stratification for recurrence and mortality in embolic stroke of undetermined source. Stroke 2016; 47: 2278–2285.
9. Peterson AD, Ghosh AP, Maitra R. Merging K-means with hierarchical clustering for identifying general-shaped groups. Stat (International Statistical Institute) 2018; 7: e172.
10. Charrad M, Ghazzali N, Boiteau V, Nbiclust Niknafs A. An R package for determining the relevant number of clusters in a data set. J Stat Softw 2014; 61: 1–36.
11. Galili T. dendextend: an R package for visualizing, adjusting and comparing trees of hierarchical clustering. Bioinformatics 2015; 31: 3718–3720.
of embolic stroke of undetermined source. Stroke 2018; 49: 3054–3056.
30. Coutinho JM, Derkatch S, Potvin AR, et al. Non-stenotic carotid plaque on CT angiography in patients with cryptogenic stroke. Neurology 2016; 87: 665–672.
31. Freilinger TM, Schindler A, Schmidt C, et al. Prevalence of nonstenosing, complicated atherosclerotic plaques in cryptogenic stroke. JACC Cardiovasc Imag 2012; 5: 397–405.
32. Hyafil F, Schindler A, Sepp D, et al. High-risk plaque features can be detected in non-stenotic carotid plaques of patients with ischaemic stroke classified as cryptogenic using combined (18)F-FDG PET/MR imaging. Eur J Nuclear Med Mol Imag 2016; 43: 270–279.
33. Buon R, Guidolin B, Jaffre A, et al. Carotid ultrasound for assessment of nonobstructive carotid atherosclerosis in young adults with cryptogenic stroke. J Stroke Cerebrovasc Dis 2018; 27: 1212–1216.
34. Siegler JE, Thon JM, Woo JH, Do D, Messe SR, Cucchiara B. Prevalence of nonstenotic carotid plaque in stroke due to atrial fibrillation compared to embolic stroke of undetermined source. J Stroke Cerebrovasc Dis 2019; 28: 104289.
35. Rajalingam R, Jalini S, Pikula A. Extracranial and intracranial non-stenotic carotid atherosclerotic plaques in ESUS patients (P5.221). Neurology 2018; 90(P5):221.
36. Gupta A, Gialdini G, Lerario MP, et al. Magnetic resonance angiography detection of abnormal carotid artery plaque in patients with cryptogenic stroke. J Am Heart Assoc 2015; 4: e002012.
37. Kamel H, Gialdini G, Baradaran H, et al. Cryptogenic stroke and nonstenosing intracranial calcified atherosclerosis. J Stroke Cerebrovasc Dis 2017; 26: 863–870.
38. Gupta A, Gialdini G, Giambrone AE, et al. Association between nonstenosing carotid artery plaque on MR angiography and acute ischemic stroke. JACC Cardiovasc Imag 2016; 9: 1228–1229.
39. Cheung HM, Moody AR, Singh N, Bitar R, Zhan J, Leung G. Late stage complicated atheroma in low-grade stenotic carotid disease: MR imaging depiction–prevalence and risk factors. Radiology 2011; 260: 841–847.
40. Singh N, Moody AR, Panzov V, Gladstone DJ. Carotid intraplaque hemorrhage in patients with embolic stroke of undetermined source. J Stroke Cerebrovasc Dis 2018; 27: 1956–1959.
41. Jaffre A, Guidolin B, Ruidavets JB, Nasr N, Larrue V. Non-obstructive carotid atherosclerosis and patent foramen ovale in young adults with cryptogenic stroke. Eur J Neurol 2017; 24: 663–666.
42. Fitzgerald S, Dair D, Wang S, et al. Platelet-rich emboli in cerebral large vessel occlusion are associated with a large artery atherosclerosis source. Stroke 2019; 50: 1907–1910.
43. Elkind MSV. Atrial cardiopathy and stroke prevention. Curr Cardiol Rep 2018; 20: 103.
44. Llombart V, Antolin-Fontes A, Bustamante A, et al. B-type natriuretic peptides help in cardioembolic stroke diagnosis: pooled data meta-analysis. Stroke 2015; 46: 1187–1195.
45. Berntsson J, Zia E, Borne Y, Melander O, Hedblad B, Engstrom G. Plasma natriuretic peptides and incidence of subtypes of ischemic stroke. Cerebrovasc Dis 2014; 37: 444–450.
46. Folsom AR, Nambi V, Bell EJ, et al. Troponin T, N-terminal pro-B-type natriuretic peptide, and incidence of stroke: the atherosclerosis risk in communities study. Stroke 2013; 44: 961–967.
47. Yaghhi S, Liberman AL, Atalay M, et al. Cardiac magnetic resonance imaging: a new tool to identify cardiovascular sources in ischaemic stroke. J Neurol Neurosurg Psychiatry 2017; 88: 31–37.
48. Kamel H, Bartz TM, Longstreth WT Jr, et al. Association between left atrial abnormality on ECG and vascular brain injury on MRI in the Cardiovascular Health Study. Stroke 2015; 46: 711–716.
49. Kamel H, Hunter M, Moon YP, et al. Electrocardiographic left atrial abnormality and risk of stroke: Northern Manhattan study. Stroke 2015; 46: 3208–3212.
50. Kamel H, O’Neal WT, Okin PM, Loehr LR, Alonso A, Soliman EZ. Electrocardiographic left atrial abnormality and stroke subtype in the atherosclerosis risk in communities study, Ann Neurol 2015; 78: 670–678.
51. Kamel H, Longstreth WT Jr, Tirschwell DL, et al. The AtRial Cardiopathy and Antithrombotic Drugs In prevention After cryptogenic stroke randomized trial: Rationale and methods. Int J Stroke 2019; 14: 207–214.
52. Nuios G, Papavasileiou V, Makaritsis K, Michel P. PFO closure vs. medical therapy in cryptogenic stroke or transient ischemic attack: a systematic review and meta-analysis. Int J Cardiol 2013; 169: 101–105.
53. Nuios G, Papavasileiou V, Sigris D, et al. Closure of patent foramen ovale versus medical therapy in patients with cryptogenic stroke or transient ischemic attack: updated systematic review and meta-analysis. Stroke 2018; 49: 412–418.
54. Sanna T, Diener HC, Passman RS, et al. Cryptogenic stroke and underlying atrial fibrillation. N Engl J Med 2014; 370: 2478–2486.
55. Gladstone DJ, Spring M, Dorian P, et al. Atrial fibrillation in patients with cryptogenic stroke. N Engl J Med 2014; 370: 2467–2477.
56. Wachter R, Groschel K, Gelbrich G, et al. Holter-electrocardiogram-monitoring in patients with acute ischaemic stroke (Find-AFRANDOMISED): an open-label randomised controlled trial. Lancet Neurol 2017; 16: 282–290.
57. Hart RG, Catanese L, Perera KS, Ntaios G, Connolly SJ. Embolic stroke of undetermined source: a systematic review and clinical update. Stroke 2017; 48: 867–872.
58. Nuios G, Papavasileiou V, Milionis H, et al. Embolic strokes of undetermined source in the Athens stroke registry: a descriptive analysis. Stroke 2015; 46: 176–181.
59. Schnabel RB, Haesler KG, Healey JS, et al. Searching for atrial fibrillation poststroke: a white paper of the AF-SCREEN international collaboration. Circulation 2019; 140: 1834–1850.
60. Healey JS, Alings M, Ha A, et al. Subclinical atrial fibrillation in older patients. Circulation 2017; 136: 1276–1283.
61. Hart RG, Sharma M, Mundl H, et al. Rivaroxaban for stroke prevention after embolic stroke of undetermined source. N Engl J Med 2018; 378: 2191–2201.
62. Diener HC, Sacco RL, Easton JD, et al. Dabigatran for prevention of stroke after embolic stroke of undetermined source. N Engl J Med 2019; 380: 1906–1917.
63. Brambatti M, Connolly SJ, Gold MR, et al. Temporal relationship between subclinical atrial fibrillation and embolic events. *Circulation* 2014; 129: 2094–2099.

64. Daoud EG, Glotzer TV, Wyse DG, et al. Temporal relationship of atrial tachyarrhythmias, cerebrovascular events, and systemic emboli based on stored device data: a subgroup analysis of TRENDS. *Heart Rhythm* 2011; 8: 1416–1423.

65. Connolly SJ, Eikelboom J, Joyner C, et al. Apixaban in patients with atrial fibrillation. *N Engl J Med* 2011; 364: 806–817.