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

George Ntaios MD[1,*], Stephen F. Weng PhD[2,3,*], Kalliopi Perlepe MD[1], Ralph Akyea MPH[3], Laura Condon PhD[3], Dimitrios Lambrou PhD[1], Gaia Sirimarco MD[4], Davide Strambo MD[4], Ashraf Eskandari MD[4], Efstathia Karagkiozi BSc[1], Anastasia Vemmou MD[5], Eleni Korompoki MD[5,6], Efstathios Manios MD[5], Konstantinos Makaritsis MD[1], Konstantinos Vemmos MD[5], Patrik Michel MD[4]

1. Department of Internal Medicine, Faculty of Medicine, School of Health Sciences, University of Thessaly, Larissa, Greece
2. NIHR School for Primary Care Research, University of Nottingham, Nottingham, United Kingdom
3. Primary Care Stratified Medicine (PRISM), Division of Primary Care, School of Medicine, University of Nottingham, Nottingham, United Kingdom
4. Stroke Center and Neurology Service, Department of Clinical Neurosciences, Centre Hospitalier Universitaire Vaudois and University of Lausanne, Lausanne, Switzerland
5. Department of Clinical Therapeutics, Medical School of Athens, Alexandra Hospital, Athens, Greece
6. Division of Brain Sciences, Department of Stroke Medicine, Imperial College, London, United Kingdom

* these authors contributed equally

Abstract/total word count:

Tables/Figures/supplemental material: 4/1/1

Corresponding author

George Ntaios MD, MSc (Stroke Medicine), PhD

Department of Internal Medicine, Faculty of Medicine, School of Health Sciences, University of Thessaly, Larissa, Greece. T: +30-241-3502888, F: +30-241-3501557, E-mail: gntaios@med.uth.gr

DR. GEORGE NTAIOS (Orcid ID: 0000-0002-0629-9248)

DR. ELENI KOROMPOKI (Orcid ID: 0000-0003-1540-8650)

Article type : Original Article

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

George Ntaios MD[1,*], Stephen F. Weng PhD[2,3,*], Kalliopi Perlepe MD[1], Ralph Akyea MPH[3], Laura Condon PhD[3], Dimitrios Lambrou PhD[1], Gaia Sirimarco MD[4], Davide Strambo MD[4], Ashraf Eskandari MD[4], Efstathia Karagkiozi BSc[1], Anastasia Vemmou MD[5], Eleni Korompoki MD[5,6], Efstathios Manios MD[5], Konstantinos Makaritsis MD[1], Konstantinos Vemmos MD[5], Patrik Michel MD[4]

1. Department of Internal Medicine, Faculty of Medicine, School of Health Sciences, University of Thessaly, Larissa, Greece
2. NIHR School for Primary Care Research, University of Nottingham, Nottingham, United Kingdom
3. Primary Care Stratified Medicine (PRISM), Division of Primary Care, School of Medicine, University of Nottingham, Nottingham, United Kingdom
4. Stroke Center and Neurology Service, Department of Clinical Neurosciences, Centre Hospitalier Universitaire Vaudois and University of Lausanne, Lausanne, Switzerland
5. Department of Clinical Therapeutics, Medical School of Athens, Alexandra Hospital, Athens, Greece
6. Division of Brain Sciences, Department of Stroke Medicine, Imperial College, London, United Kingdom

* these authors contributed equally

Abstract/total word count:

Tables/Figures/supplemental material: 4/1/1

Corresponding author: George Ntaios MD, MSc (Stroke Medicine), PhD. Department of Internal Medicine, Faculty of Medicine, School of Health Sciences, University of Thessaly, Larissa, Greece. T: +30-241-3502888, F: +30-241-3501557, E-mail: gntaios@med.uth.gr

Short running title: Data-driven machine-learning analysis in ESUS

Abstract

Background: 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 ESUS using a data-driven, machine-learning method, and explored variation in stroke recurrence between clusters.

Methods: We used 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 67years), the optimal number of clusters was 4. 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%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 4 clusters of ESUS which were strongly associated with arterial disease, atrial cardiopathy, PFO and left ventricular disease respectively. More than half of patients were assigned to the cluster associated with arterial disease.

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

Clinical trial registration
URL: https://www.clinicaltrials.gov / Unique identifier: NCT02766205

Disclosures

George Ntaios: 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).

Kalliopi Perlepe: received travel support by Pfizer

Gaia Sirimarco: 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).

Eleni Korompoki: Speaker fees/honoraria for advisory boards from Pfizer, Amgen.

Patrik Michel: 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.

Stephen Weng: 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

Funding source

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 (ERISTA). The Swiss Cardiology Foundation supported data collection in the ASTRAL registry.

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 like atherosclerotic plaques in the carotids and the aortic arch, covert atrial fibrillation (AF), patent foramen ovale (PFO), left ventricular disease, atrial cardiopathy, cancer and cardiac valvular disease[1]. Recently, we showed that there is significant overlap of potential embolic sources (PES) in patients with ESUS: in a cohort of consecutive ESUS patients, 65.5% had ≥2 PES and 31.1% had ≥3 PES, whereas on average each patient had 2 PES[2]. In this context, it is frequently difficult to identify the actual source of embolism in an ESUS patient, when several PES co-exist[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 is 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, which need to be specified *a priori*. Each test is independent of the other tests, which results in groups, which 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] This process, therefore, is extremely advantageous for exploring the potential underlying aetiology in particularly heterogeneous diseases, like 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

Data availability statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

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 description 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 a) extracranial atherosclerosis causing ≥50% luminal stenosis in arteries supplying the area of ischemia or b) major-risk cardioembolic source or c) any other specific cause of stroke (e.g. arteritis, 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 supplemental table, which included demographics, lifestyle factors, clinical symptoms/signs during the qualifying ESUS, comorbidities, biometrics, biomarkers, vascular imaging, brain imaging, electrocardiogram and echocardiography.

In order 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 sub-cluster until we determine a small fixed number of points. Using all the baseline data provided from ESUS patients, the algorithm therefore 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 were 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 (PCA) 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. Clustering analyses and data visualisation tools were conducted using statistical software R 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 PES: 0-1 PES, 2 PES, or ≥3 PES.

Potential embolic sources were categorised as follows: atrial cardiopathy, AF, arterial disease, left ventricular 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. Left ventricular disease was diagnosed if low LV ejection fraction (<35%) or LV hypertrophy or left-sided heart failure were reported at the echocardiogram, or if LV hypertrophy was identified at the electrocardiogram (Sokolow index ≥35mm)[23]. We diagnosed cardiac valvular disease if moderate-to-severe stenosis or regurgitation of the mitral or aortic valve was reported at the echocardiogram. Atrial fibrillation 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 (ECG) performed for any reason including palpitations, irregular pulse on clinical examination, in- hospital surveillance or portable outpatient monitoring.

Clinical endpoints during follow-up

We evaluated the risk of stroke recurrence over the 10 years follow-up by cluster. Stroke during follow-up was ascertained by on-site patient visits at outpatient's clinics, contact with the patient's next of kin, or the patients' primary physician. Where possible, the outcome had been adjudicated by reviewing 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 χ^2 tests for categorical variables [24, 25]. Prior to the clustering analysis, data which 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 Institute of Health Stroke Scale (NIHSS) score at admission. The PES in each cluster were then ranked by significance and by the effect size, with 95% confidence intervals provided. In this way, we were able to "profile" each cluster and associate them to specific PES.

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.

Further, we quantified the dose-response relationship of having multiple PES 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).

Visualization of the hierarchical clustering analysis

From 30 clustering indices, it was found that the optimal number of clusters is 4 (**supplemental Figure 1**). The arrangement of the 4 clusters during the clustering process is illustrated at the dendrogram (**supplemental Figure 2**).

The principal components analysis identified that 82% of all principal components were needed to explain 100% variation of the original ESUS data (**supplemental Figure 3a**), 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 (**supplemental Figure 3b**). Cluster sizes were as follows: 44 patients (5.5%) in cluster 1, 149 patients (18.6%) in cluster 2, 430 patients (53.8%) in cluster 3, and 177 patients (22.1%) in cluster 4. There was 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, left ventricular disease, PFO and cancer. Left ventricular 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 were most prevalent in cluster 4 (100%).

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.** Left ventricular 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). Over 2,922 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 **Figure 1**).

Discussion

This data-driven machine-learning analysis of consecutive ESUS patients identified 4 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 15-20% of the overall population, were associated with atrial cardiopathy and PFO respectively; and a small cluster which included only 5% of the overall population and was associated with left ventricular disease. Atrial fibrillation was not associated with any cluster. The risk of stroke recurrence was similar across clusters.

During the recent years, there has been emerging evidence supporting an important etiological association between ESUS and atherosclerotic plaques. A recent analysis of the NAVIGATE-ESUS 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 study showed that new incident AF is less frequently detected in patients with ESUS and carotid plaques compared to those without[18]. Similar, in young adults with cryptogenic stroke, carotid plaques were associated with the absence of PFO[41]. Both latter studies show that carotid plaques act as a competing stroke etiology to other established stroke etiologies, 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 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 MRI[47] and electrocardiographic indices[48-50]. The present analysis adds to the evidence which supports an important causative role of atrial cardiopathy in ESUS and indicate that atrial cardiopathy could be the cause of stroke in 15- 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) who 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 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 atrial fibrillation, as shown in patients with implantable cardiac monitoring devices in the ASSERT[63] and TRENDS[64] studies. Finally, if the association between AF and ESUS was indeed 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 vs. aspirin reported in the AVERROES trial[65]; however, this was not the case in the NAVIGATE-ESUS and RE- SPECT ESUS 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 which 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 all 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 PES is advantageous and more informative, as there is large overlap of PES in patients with ESUS. For example, a previous analysis in the same cohort showed that left ventricular disease was present in 54.4% of the overall cohort[2]; however, the present analysis showed that the cluster which was associated with left ventricular disease included only 5% of the overall cohort. This suggests that for the majority of patients with left ventricular disease, this would represent an innocent by-stander 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.

In conclusion, this data-driven machine-learning hierarchical-clustering analysis identified 4 clusters of ESUS patients which were associated with arterial disease, atrial cardiopathy, PFO and left ventricular disease respectively. Atrial fibrillation was not associated with any cluster. The risk of stroke recurrence was not different across clusters.

References

[1]. Hart RG, Diener HC, Coutts SB*, et al.* Embolic strokes of undetermined source: the case for a new clinical construct. *Lancet Neurology*. 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. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*. 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]. Vemmos 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, Vemmos 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, Niknafs A. NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set. *Journal of Statistical Software*. 2014 **61**.

[11]. Galili T. dendextend: an R package for visualizing, adjusting and comparing trees of hierarchical clustering. *Bioinformatics*. 2015 **31:** 3718-3720.

[12]. Yaghi S, Moon YP, Mora-McLaughlin C*, et al.* Left atrial enlargement and stroke recurrence: the Northern Manhattan Stroke Study. *Stroke*. 2015 **46:** 1488-1493.

[13]. Perlepe K, Sirimarco G, Strambo D*, et al.* Left atrial diameter thresholds and new incident atrial fibrillation in embolic stroke of undetermined source. *European Journal of Internal Medicine*. 2020.

[14]. O'Neal WT, Kamel H, Kleindorfer D*, et al.* Premature Atrial Contractions on the Screening Electrocardiogram and Risk of Ischemic Stroke: The Reasons for Geographic and Racial Differences in Stroke Study. *Neuroepidemiology*. 2016 **47:** 53-58.

[15]. Ntaios G, Perlepe K, Lambrou D*, et al.* Supraventricular Extrasystoles on Standard 12-lead Electrocardiogram Predict New Incident Atrial Fibrillation after Embolic

Stroke of Undetermined Source: The AF-ESUS Study. *Journal of Stroke and Cerebrovascular Diseases*. 2020 **29:** 104626.

[16]. Bulwa Z, Gupta A. Embolic stroke of undetermined source: The role of the nonstenotic carotid plaque. *Journal of the Neurological Sciences*. 2017 **382:** 49-52.

[17]. Ntaios G, Wintermark M, Michel P. Supracardiac atherosclerosis in embolic stroke of undetermined source: the underestimated source. *European Heart Journal*. 2020.

[18]. Ntaios G, Perlepe K, Sirimarco G*, et al.* Carotid plaques and detection of atrial fibrillation in embolic stroke of undetermined source. *Neurology*. 2019 **92:** e2644-e2652.

[19]. Amarenco P, Duyckaerts C, Tzourio C, Henin D, Bousser MG, Hauw JJ. The prevalence of ulcerated plaques in the aortic arch in patients with stroke. *New England Journal of Medicine*. 1992 **326:** 221-225.

[20]. Amarenco P, Cohen A, Tzourio C*, et al.* Atherosclerotic disease of the aortic arch and the risk of ischemic stroke. *New England Journal of Medicine*. 1994 **331:** 1474-1479.

[21]. Lopes RD, Vora AN, Liaw D*, et al.* An open-Label, 2 x 2 factorial, randomized controlled trial to evaluate the safety of apixaban vs. vitamin K antagonist and aspirin vs. placebo in patients with atrial fibrillation and acute coronary syndrome and/or percutaneous coronary intervention: Rationale and design of the AUGUSTUS trial. *American Heart Journal*. 2018 **200:** 17-23.

[22]. Ntaios G, Pearce LA, Meseguer E*, et al.* Aortic Arch Atherosclerosis in Patients With Embolic Stroke of Undetermined Source: An Exploratory Analysis of the NAVIGATE ESUS Trial. *Stroke*. 2019 **50:** 3184-3190.

[23]. Ntaios G. Embolic Stroke of Undetermined Source: JACC Review Topic of the Week. *Journal of the American College of Cardiology*. 2020 **75:** 333-340.

[24]. Kruskal WH, Wallis WA. Use of Ranks in One-Criterion Variance Analysis. *Journal of the American Statistical Association*. 1952 **47:** 583-621.

[25]. McHugh ML. The Chi-square test of independence. *Biochemia Medica*. 2013 **23:** 143-149.

[26]. Royston P, White IR. Multiple Imputation by Chained Equations (MICE): Implementation in Stata. *Journal of Statistical Software*. 2011 **45**.

[27]. Ntaios G, Swaminathan B, Berkowitz SD*, et al.* Efficacy and Safety of Rivaroxaban Versus Aspirin in Embolic Stroke of Undetermined Source and Carotid Atherosclerosis. *Stroke*. 2019 **50:** 2477-2485.

[28]. Hirunagi T, Miwa S, Katsuno M. [Nonstenotic Carotid Plaque in Patients with Anterior Circulation Embolic Stroke of Undetermined Source]. *Brain Nerve*. 2018 **70:** 1295-1299.

[29]. Komatsu T, Iguchi Y, Arai A*, et al.* Large but Nonstenotic Carotid Artery Plaque in Patients With a History of Embolic Stroke of Undetermined Source. *Stroke*. 2018 **49:** 3054-3056.

[30]. Coutinho JM, Derkatch S, Potvin AR*, et al.* Nonstenotic 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: Cardiovascular Imaging*. 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. *European Journal of Nuclear Medicine and Molecular Imaging*. 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. *Journal of Stroke and Cerebrovascular Diseases*. 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. *Journal of Stroke and Cerebrovascular Diseases*. 2019**:** 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. *Journal of Stroke and Cerebrovascular Diseases*. 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: Cardiovascular Imaging*. 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. *Journal of Stroke and Cerebrovascular Diseases*. 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. *European Journal of Neurology*. 2017 **24:** 663-666.

[42]. Fitzgerald S, Dai 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. *Current Cardiology Reports*. 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. *Cerebrovascular Diseases*. 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]. Yaghi S, Liberman AL, Atalay M*, et al.* Cardiac magnetic resonance imaging: a new tool to identify cardioaortic sources in ischaemic stroke. *Journal of Neurology, Neurosurgery and 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. *Annals of Neurology*. 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. *International Journal of Stroke*. 2019 **14:** 207-214.

[52]. Ntaios 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. *International Journal of Cardiology*. 2013 **169:** 101-105.

[53]. Ntaios G, Papavasileiou V, Sagris 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. *New England Journal of Medicine*. 2014 **370:** 2478-2486.

[55]. Gladstone DJ, Spring M, Dorian P*, et al.* Atrial fibrillation in patients with cryptogenic stroke. *New England Journal of Medicine*. 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 Neurology*. 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]. Ntaios 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, Haeusler 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. *New England Journal of Medicine*. 2018 **378:** 2191-2201.

[62]. Diener HC, Sacco RL, Easton JD*, et al.* Dabigatran for Prevention of Stroke after Embolic Stroke of Undetermined Source. *New England Journal of Medicine*. 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. *New England Journal of Medicine*. 2011 **364:** 806-817.

Table 1. Baseline characteristics and outcomes of patients, stratified by cluster. IQR – Interquartile range; NIHSS – National Institute of Health Stroke Scale; n – number; % - percentage. * (per 1000 patient years).

		Cluster 1 (n=44)	**Cluster 2** (n=149)	**Cluster 3** (n=430)	**Cluster 4** (n=177)	***p*-value**
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
Antiplatelet 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) *	21.7 (7.0– 67.3)	29.3 (17.3–49.5)	29.5 (22.0–39.5)	50.1 (36.6–68.6)

Table 2. Prevalence of potential embolic sources and degree of their overlap stratified by cluster. PES – potential embolic sources; IQR – interquartile range, PFO – patent foramen ovale; n – number; % - percentage

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 valvular 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

Table 3. Multivariable logistic regression to determine association and effect size for each potential embolic source and cluster membership. 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. PES - potential embolic source; CI – confidence interval; * 100% of individuals within the cluster had the condition.

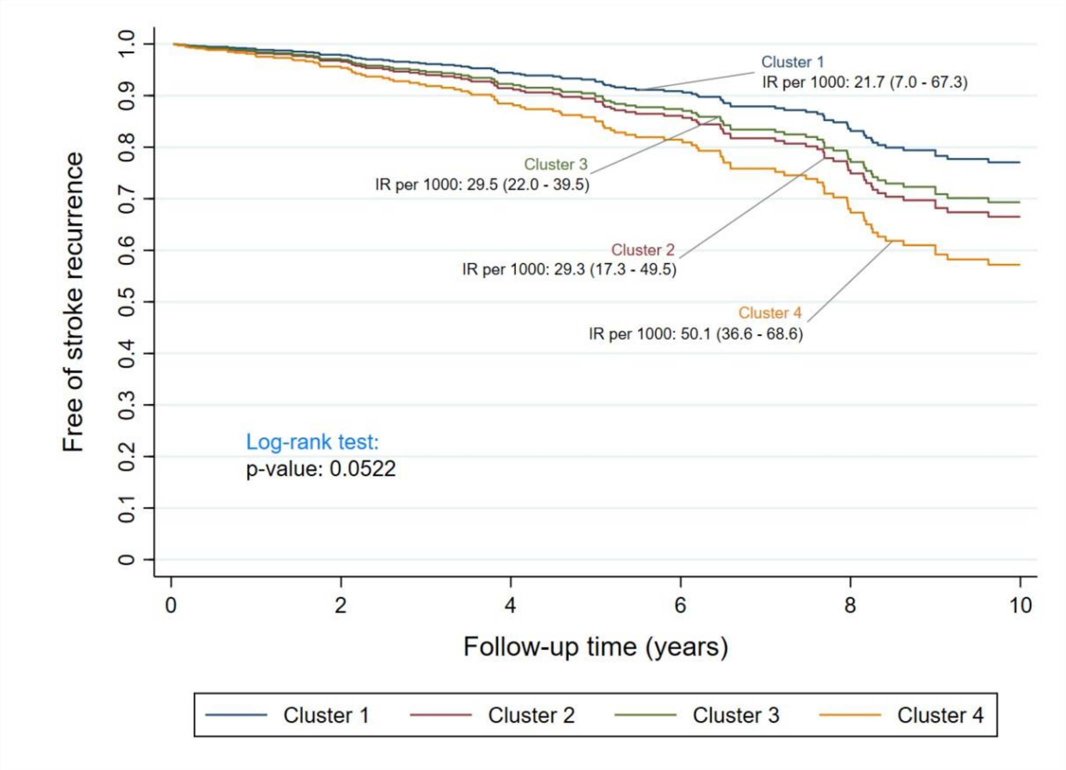
	Odds ratio	**95% CI**	**Association**
Cluster 1			
Left ventricular disease	Perfectly associated with cluster*		Positive association
Arterial disease	0.22	0.09 – 0.53	Negative association
Atrial cardiopathy	1.82	0.81 – 4.05	No association
Cardiac valvular disease	1.35	0.48 – 3.79	No association
Atrial 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 cardiomyopathy	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

Table 4. Multivariable regression analysis of the association between the phenotype clusters and stroke recurrence. 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. PES - potential embolic source; CI – confidence interval

	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

Figure 1. Ten-year survival estimates of stroke recurrence in patients with embolic stroke of undetermined source, according to the assigned phenotype clusters


Cluster 1
IR per 1000: 21.7 (7.0 - 67.3)
Cluster 3
IR per 1000: 29.5 (22.0 - 39.5)
Cluster 2
IR per 1000: 29.3 (17.3 - 49.5)
Cluster 4
IR per 1000: 50.1 (36.6 - 68.6)
Log-rank test:
p-value: 0.0522
Free of stroke recurrence
Follow-up time (years)
Cluster 1
Cluster 2
Cluster 3
Cluster 4