Seizure: European Journal of Epilepsy 111 (2023) 58–67

Contents lists available at ScienceDirect

Seizure: European Journal of Epilepsy

journal homepage: www.elsevier.com/locate/seizure

# Association of comorbid-socioeconomic clusters with mortality in late onset epilepsy derived through unsupervised machine learning

Colin B. Josephson [a,b,c,d,e,*], Arturo Gonzalez-Izquierdo [f,g], Jordan D.T. Engbers [h], Spiros Denaxas [f,g,i], Guillermo Delgado-Garcia [a], Tolulope T. Sajobi [a,b,c,d], Meng Wang [b], Mark R. Keezer [j], Samuel Wiebe [a,b,c,d,k]

[a] Department of Clinical Neurosciences, Cumming School of Medicine, University of Calgary, Calgary, AB, Canada
[b] Department of Community Health Sciences, Cumming School of Medicine, University of Calgary, AB, Canada
[c] Hotchkiss Brain Institute, University of Calgary, Calgary, AB, Canada
[d] O'Brien Institute for Public Health, University of Calgary, Calgary, AB, Canada
[e] Centre for Health Informatics, University of Calgary, Calgary, AB, Canada
[f] UCL Institute of Health Informatics, London, UK
[g] Health Data Research (HDR) UK, London, UK
[h] Desid Labs, Inc, Canada
[i] Alan Turing Institute, London, UK
[j] Department of Neurosciences, Université de Montreal, Montreal, Quebec, Canada
[k] Clinical Research Unit, Cumming School of Medicine, University of Calgary, Calgary, AB, Canada

ARTICLE INFO

*Keywords:*
Epilepsy
Cohort study
Electronic health records
Unsupervised machine learning
Elderly
Late-onset epilepsy

ABSTRACT

*Background and objectives:* Late-onset epilepsy is a heterogenous entity associated with specific aetiologies and an elevated risk of premature mortality. Specific multimorbid-socioeconomic profiles and their unique prognostic trajectories have not been described. We sought to determine if specific clusters of late onset epilepsy exist, and whether they have unique hazards of premature mortality.

*Methods:* We performed a retrospective observational cohort study linking primary and hospital-based UK electronic health records with vital statistics data (covering years 1998–2019) to identify all cases of incident late onset epilepsy (from people aged ≥65) and 1:10 age, sex, and GP practice-matched controls. We applied hierarchical agglomerative clustering using common aetiologies identified at baseline to define multimorbid-socioeconomic profiles, compare hazards of early mortality, and tabulating causes of death stratified by cluster.

*Results:* From 1,032,129 people aged ≥65, we identified 1048 cases of late onset epilepsy who were matched to 10,259 controls. Median age at epilepsy diagnosis was 68 (interquartile range: 66–72) and 474 (45%) were female. The hazard of premature mortality related to late-onset epilepsy was higher than matched controls (hazard ratio [HR] 1.73; 95% confidence interval [95%CI] 1.51–1.99). Ten unique phenotypic clusters were identified, defined by 'healthy' males and females, ischaemic stroke, intracerebral haemorrhage (ICH), ICH and alcohol misuse, dementia and anxiety, anxiety, depression in males and females, and brain tumours. Cluster-specific hazards were often similar to that derived for late-onset epilepsy as a whole. Clusters that differed significantly from the base late-onset epilepsy hazard were 'dementia and anxiety' (HR 5.36; 95%CI 3.31–8.68), 'brain tumour' (HR 4.97; 95%CI 2.89–8.56), 'ICH and alcohol misuse' (HR 2.91; 95%CI 1.76–4.81), and 'ischaemic stroke' (HR 2.83; 95%CI 1.83–4.04). These cluster-specific risks were also elevated compared to those derived for tumours, dementia, ischaemic stroke, and ICH in the whole population. Seizure-related cause of death was uncommon and restricted to the ICH, ICH and alcohol misuse, and healthy female clusters.

*Significance:* Late-onset epilepsy is an amalgam of unique phenotypic clusters that can be quantitatively defined. Late-onset epilepsy and cluster-specific comorbid profiles have complex effects on premature mortality above and beyond the base rates attributed to epilepsy and cluster-defining comorbidities alone.

* Corresponding author at: Assistant Professor of Neurology, Cumming School of Medicine, University of Calgary, Foothills Medical Centre, 1403 - 29St NW, Calgary, Alberta, Canada.

https://doi.org/10.1016/j.seizure.2023.07.016
Received 5 May 2023; Received in revised form 20 July 2023; Accepted 22 July 2023
Available online 29 July 2023

## 1. Introduction

Late-onset epilepsy is a unique entity compared to that in younger adults. After age 65, the clinicodemographic characteristics defining the condition diverge significantly from that in younger adults [1]. The aetiologies underlying late-onset epilepsy differ substantially, with strokes, neurodegenerative diseases, and brain tumours frequently reported as the primary underlying cause, where it is known [2]. Increasing longevity, coupled with the attendant rise of these associated conditions, likely accounts for the rising incidence of late-onset epilepsy in the Western world over the last 45 years [3,4]. Ostensibly, late-onset epilepsy is comparatively easy to control, with up to 75–80% achieving seizure freedom [2,5–7] compared to ~65% in clinic populations irrespective of age [8], though certain populations with post-stroke epilepsy may require polytherapy [9]. Despite this, late-onset epilepsy is associated with an increased hazard of premature mortality [10].

The reasons for premature mortality in late-onset epilepsy remain incompletely explained. Machine learning and stratified medicine approaches, such as clustering analyses, can identify unique subsets of patients at discrete risks of premature mortality, thus helping elucidate potential underlying causes and individualized targets for preventative therapy. Using conventional aetiological risk factors and conditions with bidirectional associations for epilepsy, the objective of this study was to identify unique phenotypic clusters of late-onset epilepsy and ascertain whether these groups have distinct risks and causes of death compared to each other and matched controls without epilepsy.

## 2. Materials and methods

### 2.1. Database

This study was carried out as part of the CALIBRE © resource (https://www.ucl.ac.uk/health-informatics/calibre and https://www.caliberresearch.org/). CALIBRE, led from the University College London (UCL) Institute of Health Informatics, is a research resource providing validated electronic health record phenotyping algorithms and tools for national structured data sources [11,12]. The CALIBRE resource (https://www.ucl.ac.uk/health-informatics/calibre) [11] curates United Kingdom (UK) nationally linked structured electronic health records (EHR) data from primary care (Clinical Practice Research Dataline; 'CPRD'), hospital care, and a cause-specific mortality registry up to March 31, 2019. Read codes [13] version 2 are used to code medical events, whilst the British National Formulary is used to identify prescriptions in the CPRD dataset [14,15]. The CPRD data are broadly representative of the UK population in terms of age, sex, and ethnicity [16]. The Hospital Episode Statistics (HES) database comprises information collected by audit nurses and professional clinical coders who collate secondary care and administrative data. Diagnoses in HES use the international classification of diseases (ICD-10) and the Office of Population Censuses and Surveys Classification of Interventions and Procedures terminology (OPCS-4).

### 2.2. Study design, case ascertainment and study population

This was a retrospective open cohort study of patients aged ≥65 years. We restricted the population to those aged ≥65 at epilepsy onset [1]. We used the 'Epilepsy Only' case definition from the Secure Anonymised Information Linkage (SAIL) databank (Wales, UK) to identify people with epilepsy (sensitivity of 88% and specificity of 98% following a review of clinical records of 300 patients) [17,18] (Appendix 1). The SAIL database is constructed similarly to the CALIBRE resource, with overlap in CPRD and HES datasets [19], facilitating use of this case definition which adheres to the International League Against Epilepsy's (ILAE) practical clinical definition of epilepsy. This permitted calculation of measures of sensitivity, specificity, positive predictive value, and Youden's Index [17], and is a highly reliable method of case validation [16].

To identify incident late-onset cases, we applied a 5-year washout period prior to the date of diagnosis after age 65 during which participants could not receive a code for epilepsy or seizures. People with epilepsy were matched 1:10 to controls based on year of birth (+/- 5-years), sex, and general practitioner (GP) practice. The index date for controls was that of the date of epilepsy diagnosis for the person with whom they were matched. The study period entailed January 1, 1998 (the date of CPRD linkage with HES admitted patient care data) to March 31, 2019. All patients meeting these criteria were included in the study. Datasets are linked using the NHS number, a unique ten-digit identifier assigned at first encounter with the healthcare system.

### 2.3. Exposure and outcome definitions

Unsupervised machine learning analysis, as described below, was performed using agglomerative hierarchical clustering. For survival analyses, the resultant clusters were considered the exposures. All-cause death was the primary outcome. We also examined cause-specific mortality by categorizing death as epilepsy-specific, neurological (not including epilepsy), accidental, cancer, cardiovascular (including stroke), dementia, infection, psychiatric, renal, respiratory, and 'other' using the attendant ICD-10 codes. Cause of death was ascertained from the linked United Kingdom's Office for National Statistics (ONS) database.

### 2.4. Covariates

All outcome conditions of interest were defined using the Health Data Research UK's Phenotype Library portal phenotypes (https://phenotypes.healthdatagateway.org/) [11,20]. We chose to cluster on epilepsy risk factors present at the time of epilepsy diagnosis or the index date. These specifically comprised age of diagnosis, sex, depression [21,22], anxiety [21,22], ischaemic stroke [23], intracerebral haemorrhage (ICH) [24], dementia [25], brain tumour [26], alcohol misuse (*versus* low risk drinking/'not specified [42%]') [27], poor socioeconomic status [28] (as defined by the Index of Multiple Deprivation 2015, 'IMD'; the IMD is divided into deciles with 1 being the lowest socioeconomic status and 10 being the highest [29]), Charlson comorbidity index (which contains conditions associated with epilepsy), and frailty [30]). Frailty was assessed as a continuous variable using e-Frailty index [31], which ranges from 0 to 1 and is the unweighted proportion of 36 constituent conditions with which the patient has been diagnosed. Multiple imputation ($n = 10$; averaged across iterations) was performed for the 42% with missing IMD or CCI values that were missing at random [32]. Missing data was <5% for other included variables. We omitted CNS infections and traumatic brain injuries from the clustering step due to their relative scarcity.

### 2.5. Statistical analyses

We used conditional odds ratios (categorical) and random effects panel data regression (continuous) to compare the distribution of sociodemographic and clinical features between people with incident late-onset epilepsy and the 1:10 matched controls.

For clustering, we initially normalised all data by scaling them between 0 and 1 using the MinMaxScaler from the scikit-learn package (version 0.22.1; Python 3.6.3) [21]. Briefly, each feature is scaled between 0 and 1 by subtracting the lower range bound from each sample and dividing by the range. All features described above were applied to an agglomerative hierarchical clustering algorithm using the Ward method of linkage given the ability of this algorithm to handle binary data [33,34]. The agglomerative approach initially treats individual patients as their own singleton cluster. The cluster process then begins whereby a distance matrix of all samples is created, and singleton clusters are then paired. The cluster with the smallest increase in

within-cluster variance, based on within-cluster sum of squared errors, is retained. When continuing this in an iterative process, a dendrogram is constructed and ends when a single unifying cluster remains. We used the Davies-Bouldin index, silhouette coefficient (derived using 5-fold cross validation to obtain a 95% confidence interval), and cluster stability index to determine the ideal cluster number and evaluate overall model performance. A lower Davies-Bouldin index relates to a model with better cluster separation (0 is the lowest possible score), the silhouette coefficient ranges from −1 to +1 with +1 indicating perfectly separated and non-overlapping clusters, and the stability index ranges from 0 to 1 with 1 meaning perfectly stable clusters after multiple iterations.

We used chi-squared (categorical) and Kruskal-Wallis (continuous) tests to compare features between the resulting epilepsy clusters. We used Cox proportional hazards regression modelling after ensuring proportional hazards assumptions were met [35]. For all models, the index date for the analysis was the date of epilepsy diagnosis (which was equivalent to the matching date for controls), and all participants were tracked until last follow-up or death. Follow-up was restricted to 10 years from index date given the median age of onset was 68 with few events beyond this time span.

In the first model, the hazard of death related to incident late-onset epilepsy was derived controlling for all factors included in the clustering analysis along with baseline general risk factors for premature mortality (dyslipidaemia, myocardial infarction, diabetes mellitus, chronic kidney disease, liver cirrhosis, traumatic brain injury [TBI], and CNS infection). Specifically, although CNS infections and TBI were excluded from the clustering step due to their low prevalence, we decided to include them in the mortality analyses given their strong associations with premature mortality. We also included relative defined daily dose (rDDD) of anti-seizure medications (ASMs) given its correlation with the severity of epilepsy and use of enzyme-inducing ASMs, which are associated with elevated risks of cardiovascular disease [36]. We calculated this by multiplying the prescribed pill strength by the number of prescribed pills and dividing the product by the duration of the prescription. We then performed survival analyses in which cluster assignment, age, sex, the baseline general risk factors for death and epilepsy not included in the clustering, rDDD of ASMs, and use of enzyme-inducing ASMs were used as covariates. The cluster assignment was treated as a factor variable, and all were referenced to the matched controls without epilepsy to derive the hazard ratio (HR) and 95% confidence interval (95%CI) for each group.

Finally, we tabulated causes of death for each unique cluster for those patients consenting to linkage with the Office for National Statistics, and who had the events recorded in CPRD. The primary cause of death was the immediate cause, whilst secondary cases were other significant conditions felt by the clinician to have contributed to the death.

Analyses were performed using Stata version 16.1 (StataCorp LP) [37] and Python version 3.6.3 (Python Software Foundation) [38].

### 2.6. Standard protocol approvals, registrations, and patient consents

The study was approved by the MHRA (UK) Independent Scientific Advisory Committee [17_064RA3], under Section 251 (NHS Social Care Act 2006). Patient consent is waived due to the de-identified nature of the data.

### Data availability

According to the ethics board and approval stipulations, data cannot be made available due to confidentiality and privacy concerns.

## 3. Results

### 3.1. Patient population

We identified 13,417,736 patients that had follow-up of at least one-year from inception into the database, and 1032,129 (8%) were aged ≥65. A total of 1048 cases were identified (incidence proportion 102 per 100,000) who had strictly defined incident late-onset epilepsy and were matched to 10,259 controls for a total study population of 11,307 people.

The clinical features of those with late-onset epilepsy and matched controls can be found in Table 1. Median age at epilepsy diagnosis was 68 (interquartile range, 'IQR' 66–72) and 474 (45%) were female. Those diagnosed with epilepsy had higher levels of frailty, as measured by the e-Frailty index (0.2 [95% confidence interval, '95%CI', 0.14–0.29] *versus* 0.1 [95%CI 0.05–0.2]; $p<0.001$), and significantly higher conditional odds of brain tumour (14.6; 95%CI 8.5–24.9; $p<0.001$), CNS infection (11.4; 95%CI 5.2–24.9; $p<0.001$), ischaemic stroke (9.1; 95% CI 6.4–12.9; $p<0.001$), haemorrhagic stroke (8.9; 95%CI 6.2–12.7; $p<0.001$), traumatic brain injury (8.5; 95%CI 4.5–15.8; $p<0.001$), dementia (7.4; 95%CI 5.2–10.5; $p<0.001$), alcohol misuse (1.8; 95%CI 1.2–2.3; $p<0.001$), depression (1.6; 95%CI 1.3–1.9; $p<0.001$), anxiety (1.5; 95%CI 1.2–1.8; $p<0.001$), chronic kidney injury (1.4; 95%CI 1.1–1.7; $p = 0.001$), hypertension (1.3; 95%CI 1.2–1.6; $p<0.001$; Table 1). Patients were followed for a median of 3.7 years (IQR 1.7–6.3 years; 3.2 years [IQR 1.4–6.0] for people with epilepsy *versus* 3.7 years

**Table 1**
Comparison of the 1048 incident cases of late-onset epilepsy diagnosed at ≥65 and the 1:10 age-, sex-, and GP practice-matched controls.

| Feature | Epilepsy | Matched controls | Conditional odds ratio | p-value |
|---|---|---|---|---|
| n | 1048 | 10,259 | N/A | n/a |
| Age diagnosis; median, IQR* | 68 (66–72) | 69 (66–72) | N/A | – |
| Female sex* | 474 (45%) | 4752 (46%) | N/A | – |
| IMD; median, IQR | 5 (4–6) | 5 (4–6) | N/A | <0.001 |
| Alcohol misuse | 66 (6%) | 373 (4%) | 1.75 (1.32–2.33) | <0.001 |
| Ex or current smoker | 383 (37%) | 3516 (34%) | 1.10 (0.95–1.27) | 0.16 |
| e-Frailty; median, IQR | 0.2 (0.14–0.29) | 0.1 (0.05–0.2) | – | <0.001 |
| Hypertension | 559 (53%) | 4771 (47%) | 1.38 (1.21–1.58) | <0.001 |
| Dyslipidaemia | 266 (25%) | 2345 (23%) | 1.15 (0.99–1.35) | 0.06 |
| Diabetes Mellitus | 149 (14%) | 1338 (13%) | 1.14 (0.94–1.38) | 0.16 |
| Myocardial infarction | 60 (6%) | 554 (5%) | 1.06 (0.80–1.42) | 0.65 |
| Ischaemic stroke | 78 (7%) | 91 (0.9%) | 9.13 (6.4–12.9) | <0.0001 |
| Haemorrhagic stroke | 69 (7%) | 70 (0.7%) | 8.89 (6.2–12.7) | <0.001 |
| Depression | 270 (26%) | 1884 (18%) | 1.59 (1.36–1.87) | <0.001 |
| Anxiety | 177 (17%) | 1274 (12%) | 1.49 (1.24–1.78) | <0.001 |
| Dementia | 62 (6%) | 88 (0.9%) | 7.39 (5.2–10.5) | <0.001 |
| Cirrhosis | 4 (0.4%) | 40 (0.4%) | 1.04 (0.36–3.00) | 0.93 |
| CKI | 144 (14%) | 1080 (11%) | 1.41 (1.16–1.73) | 0.001 |
| Brain tumour | 40 (4%) | 32 (0.3%) | 14.6 (8.54–24.9) | <0.001 |
| Prior TBI | 23 (2%) | 24 (0.2%) | 8.51 (4.58–15.8) | <0.001 |
| Prior CNS infection | 16 (2%) | 17 (0.2%) | 11.4 (5.22–24.9) | <0.001 |

CKI = chronic kidney injury; IMD = Index of Multiple Deprivation; IQR = interquartile range; TBI = traumatic brain injury.
* Matching variables.

[IQR 1.7–6.4] for matched controls), during which 235 (22%) people with epilepsy and 913 (9%) matched controls died with 10-years of diagnosis/matching date.

### 3.2. Clustering analysis

Agglomerative hierarchical clustering identified ten unique late-onset epilepsy clusters (Fig. 1; Table 2). Cluster validity was high, with ten clusters resulting in relative plateauing of the Davies-Bouldin index (e-Figure-1), a silhouette coefficient of 0.39 (95%CI 0.33–0.45) (e-Fig. 2), and cluster stability was 0.93. Based on these results, and reviewing the constituent distribution of variables, 10 clusters provided an ideal balance between ensuring ideal group separation and stability and avoiding an unrestrained number of small clusters that lose meaning through excessive granularity. The ten clinical clusters displayed the following distinct characteristics with reference to the dendrogram in Fig. 1:

1) Split Point A (Fig. 1) defines the first two clusters, which are the most distinct from the others. These are characterized by 'healthy males' and 'brain tumours'.
   I 'Healthy Males' ($n = 312$; 30%) is a cluster with 100% males and few other core comorbidities including no people with depression, anxiety, alcohol misuse, dementia, ischaemic strokes, ICH, or brain tumours. Sixty-eight (22%) people in this cluster died.
   II 'Brain tumours' ($n = 39$; 4%) is a cluster with 100% of people having a brain tumour. They were otherwise relatively comparable to other clusters, though no person had a myocardial infarction, dementia, TBI, or CNS infections. Fourteen (36%) people in this cluster died.
2) Split Point B divides 'Healthy females' from the remaining the clusters.
   I 'Healthy Females' ($n = 204$; 19%) is a cluster with 100% females and few other core comorbidities including no people with depression, anxiety, alcohol misuse, dementia, ischaemic strokes, ICH, or brain tumours. Thirty-four (17%) people in this cluster died.
3) Split Point C separates the clusters of patients with 'Anxiety' and 'Dementia and anxiety' (Split Point D) from the remaining clusters (Split Points E-H).
4) Spilt point D separates 'Anxiety' and 'Dementia and anxiety'.
   I 'Dementia and anxiety' ($n = 59$; 6%) is a cluster with 100% of people having dementia at baseline. Otherwise, they are comparable to other clusters apart from relatively high proportion with concomitant anxiety (23; 40%) and female sex (33; 56%). Eighteen (31%) people in this cluster died.
   II 'Anxiety' ($n = 124$; 12%) is a cluster with 100% of people having anxiety at baseline. Otherwise, they are comparable to other clusters though no people had strokes, ICH, or dementia. Twenty-five (20%) people in this cluster died.
5) Split Point E separates the Depression clusters from the Stroke clusters.
6) Split Point F separates the Depression clusters.
   I 'Male depression' ($n = 57$; 5%) is a cluster with 100% of people being male and having depression at baseline. Otherwise, the cluster is relatively free of major comorbidities at baseline. Twelve (21%) people in this cluster died.
   II 'Female depression' ($n = 67$; 6%) is a cluster with 100% of people being female and having depression at baseline. Otherwise, the cluster is relatively free of major comorbidities at baseline. Thirteen (19%) people in this cluster died.
7) Split Point G (Fig. 1) separates 'ischaemic stroke' from ICH.
   I 'Ischaemic stroke' ($n = 72$; 7%) is a cluster with 100% of people having had an ischaemic stroke at baseline. There are concurrently high proportions with hypertension (74%), and frailty (eFI 0.3; interquartile range [IQR] 0.2–0.3). Twenty-one (29%) of this cluster died.
8) The final divergence, Split Point H (Fig. 1) separates 'ICH' from 'ICH with alcohol misuse'.
   I 'ICH' ($n = 60$; 6%) is a cluster with 100% of people having had a haemorrhagic stroke at baseline. Otherwise, they are comparable to other clusters though they did require slightly higher doses of ASMs (rDDD 1; IQR 0.5–1.4) and there was a higher proportion with TBI (13%). Fourteen (23%) of this cluster died.
   II 'ICH with alcohol misuse' ($n = 60$; 6%) is a cluster with 100% of people alcohol misuse at baseline, along with the second highest proportion with ICH (6%). Fourteen (23%) of this cluster died.

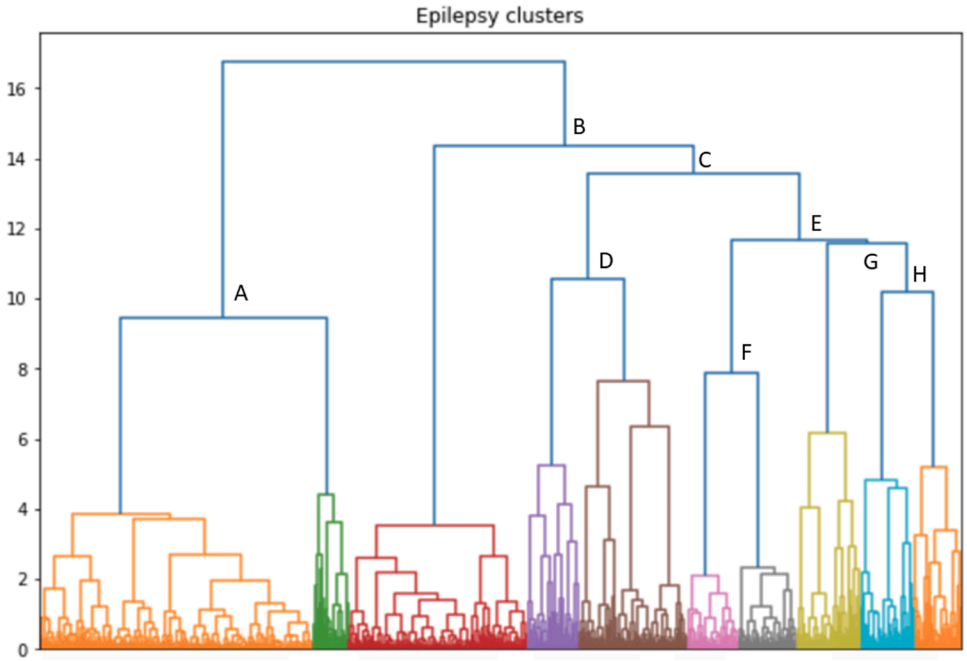


**Fig. 1.** Dendrogram displaying the linkage pattern between unique clusters of late onset epilepsy
The resultant dendrogram from agglomerative hierarchical clustering using Ward's linkage on 1048 patients with incident late-onset epilepsy using baseline aetiologies and risk factors. The cluster designations are as follows: light orange (far left): 'Healthy males', green: 'Brain tumour', red: 'Healthy female', purple: 'Dementia and anxiety', brown: 'Anxiety', pink: 'Depression male', grey: 'Depression female', yellow: 'Ischaemic stroke', blue: 'Intracerebral haemorrhage', dark orange (far right): 'Intracerebral haemorrhage and alcohol misuse'. The horizonal-axis is the starting positions of the individual patients, and the vertical-axis is Euclidean distance.

**Table 2**
Comparison of the demographic, clinical, and psychosocial variables between the ten late-onset epilepsy clusters identified through agglomerative hierarchical clustering and age-, sex-, and GP practice-matched controls. All characteristics were present at the point of diagnosis of late-onset epilepsy (which is the same as the matched date for controls).

| Feature | Healthy male | Brain tumour | Healthy female | Dementia & anxiety | Anxiety | Depression male | Depression female | Ischaemic stroke | ICH | ICH and alcohol misuse | Matched controls | p-value |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| N (%*) | 312 (30%) | 39 (4%) | 204 (19%) | 59 (6%) | 124 (12%) | 57 (5%) | 67 (6%) | 72 (7%) | 60 (6%) | 54 (5%) | 10,259 | n/a |
| Died | 68 (22%) | 14 (36%) | 34 (17%) | 18 (31%) | 25 (20%) | 12 (21%) | 13 (19%) | 21 (29%) | 14 (23%) | 16 (30%) | 913 (9%) | <0.001 |
| Current age; median, IQR | 75 (71–77) | 73 (69–76) | 76 (72–79) | 72 (69–76) | 75 (71–78) | 74 (70–77) | 74 (72–79) | 75 (71–78) | 74 (71–78) | 74 (72–79) | 75 (71–78) | <0.001 |
| Age at epilepsy diagnosis; median, IQR | 69 (66–72) | 69 (67–73) | 69 (67–73) | 69 (67–72) | 69 (66–72) | 69 (66–72) | 69 (67–73) | 70 (67–74) | 69 (66–73) | 70 (66–75) | N/A | <0.001 |
| Female sex | 0 (0%) | 16 (41%) | 204 (100%) | 33 (56%) | 80 (64%) | 0 (0%) | 67 (100%) | 26 (36%) | 30 (50%) | 18 (33%) | 4752 (46%) | <0.001 |
| IMD | 5 (4–6) | 5 (3–7) | 5 (4–6) | 6 (5–7) | 5 (4–6) | 6 (4–7) | 5 (4–6) | 5 (4–6) | 5 (4–6) | 6 (5–7) | 5 (4–6) | <0.001 |
| Alcohol misuse | 0 | 1 (3%) | 0 | 2 (3%) | 6 (4%) | 0 | 0 | 2 (3%) | 1 (2%) | 54 (100%) | 373 (4%) | <0.001 |
| Ex- or current smoker | 119 (35%) | 13 (33%) | 56 (27%) | 20 (34%) | 56 (45%) | 28 (49%) | 24 (36%) | 25 (35%) | 25 (42%) | 26 (48%) | 3516 (34%) | 0.013 |
| e-Frailty index; median, IQR | 0.1 (0.1–0.2) | 0.2 (0.1–0.2) | 0.2 (0.1–0.3) | 0.2 (0.1–0.3) | 0.2 (0.1–0.3) | 0.2 (0.1–0.3) | 0.2 (0.1–0.3) | 0.3 (0.2–0.3) | 0.2 (0.1–0.3) | 0.2 (0.1–0.3) | 0.11 (0.05–0.20) | <0.001 |
| Hypertension | 151 (48%) | 21 (54%) | 105 (51%) | 23 (40%) | 61 (49%) | 36 (63%) | 39 (58%) | 53 (74%) | 36 (60%) | 34 (63%) | 4771 (47%) | <0.001 |
| Dyslipidaemia | 66 (21%) | 5 (13%) | 64 (31%) | 15 (25%) | 28 (23%) | 21 (37%) | 17 (25%) | 22 (31%) | 16 (27%) | 12 (22%) | 2345 (23%) | 0.026 |
| Diabetes Mellitus | 49 (16%) | 5 (13%) | 25 (12%) | 8 (14%) | 15 (12%) | 13 (23%) | 9 (13%) | 12 (17%) | 5 (8%) | 8 (15%) | 2345 (23%) | 0.537 |
| Myocardial infarction | 17 (5%) | 0 (0%) | 11 (5%) | 2 (3%) | 6 (5%) | 5 (9%) | 2 (3%) | 9 (13%) | 5 (8%) | 3 (6%) | 554 (5%) | 0.233 |
| Ischaemic stroke | 0 | 1 (3%) | 0 | 1 (2%) | 0 | 0 | 0 | 72 (100%) | 1 (2%) | 3 (6%) | 91 (0.9%) | <0.001 |
| Haemorrhagic stroke | 0 | 2 (5%) | 0 | 1 (2%) | 0 | 0 | 0 | 0 | 60 (100%) | 6 (11%) | 70 (0.7%) | <0.001 |
| Depression | 0 | 10 (26%) | 0 | 22 (37%) | 58 (47%) | 57 (100%) | 67 (100%) | 26 (36%) | 14 (23%) | 16 (30%) | 1884 (18%) | <0.001 |
| Anxiety | 0 | 7 (18%) | 0 | 23 (40%) | 124 (100%) | 0 | 0 | 15 (21%) | 5 (8%) | 3 (6%) | 1274 (12%) | <0.001 |
| Dementia | 0 | 0 | 0 | 59 (100%) | 0 | 0 | 0 | 1 (1%) | 0 | 2 (4%) | 88 (0.9%) | <0.001 |
| Cirrhosis | 1 (0.3%) | 0 | 0 | 0 | 3 (2%) | 0 | 0 | 0 | 0 | 0 | 40 (0.4%) | 0.111 |
| CKI | 36 (12%) | 5 (13%) | 33 (16%) | 6 (10%) | 16 (13%) | 7 (12%) | 12 (18%) | 16 (22%) | 6 (10%) | 7 (13%) | 1080 (11%) | 0.018 |
| Prior TBI | 6 (2%) | 0 | 4 (2%) | 1 (2%) | 0 | 1 (2%) | 1 (1%) | 2 (3%) | 8 (13%) | 0 | 24 (0.2%) | <0.001 |
| Prior CNS infection | 5 (2%) | 0 | 3 (1%) | 0 | 3 (2%) | 0 | 3 (4%) | 0 | 2 (3%) | 0 | 17 (0.2%) | <0.001 |
| Brain tumour | 0 | 39 (100%) | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 (2%) | 32 (0.3%) | <0.001 |
| CCI; median, IQR, range | 2 (1–3) | 3 (2–4) | 2 (1–3) | 2 (1–3) | 2 (1–3) | 2 (1–3) | 2 (1–3) | 3 (2–4) | 2 (1–3) | 2 (1–3) | 2 (1–3) Range 1–13 | <0.001 |
| Rounded mean rDDD of ASMs, IQR | 0.7 (0.4–1.0) | 0.9 (0.5–1.3) | 0.7 (0.4–0.9) | 0.7 (0.4–1) | 0.7 (0.4–1) | 0.7 (0.4–0.9) | 0.7 (0.5–1) | 0.7 (0.5–0.9) | 1 (0.5–1.4) | 0.8 (0.4–1.13) | 0 (0–0) | <0.001 |
| Enzyme-inducing ASM | 33 (11%) | 4 (10%) | 21 (10%) | 8 (14%) | 17 (14%) | 6 (11%) | 11 (16%) | 5 (7%) | 10 (17%) | 9 (17%) | 255 (2%) | <0.001 |

ASM = antiseizure medications; CCI = Charlson comorbidity index; CKI = chronic kidney injury; IMD = Index of Multiple Deprivation; IQR = interquartile range; rDDD = relative defined daily dose rounded to the nearest 0.5; TBI = traumatic brain injury.

**p-value is an omnibus comparison across all groups including matched controls.

* percentage of people diagnosed with incident epilepsy ($n = 1048$).

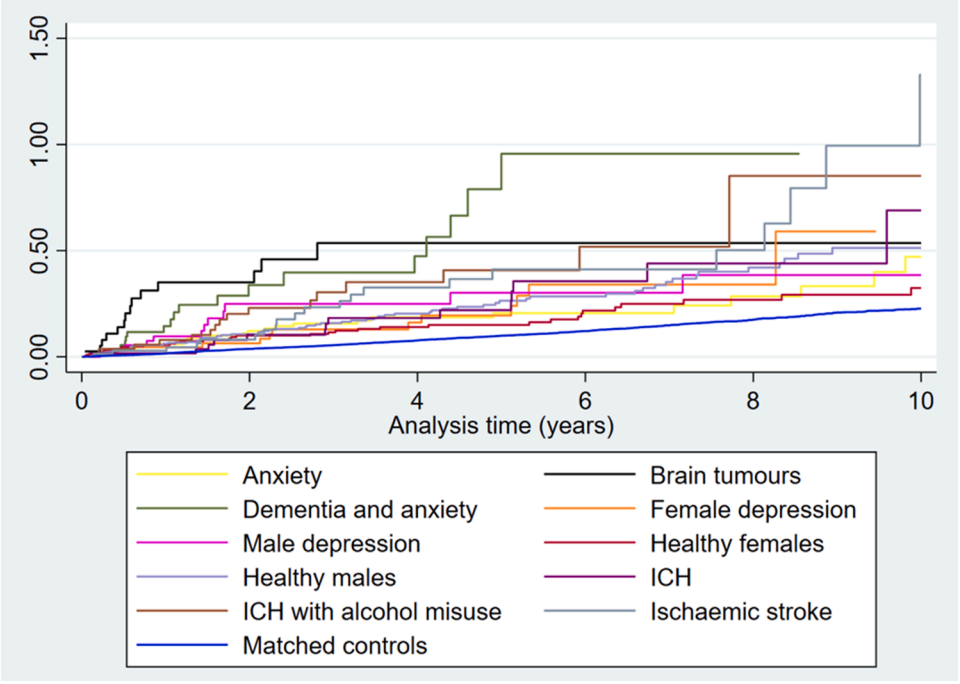


**Fig. 2.** Nelson-Aalen cumulative hazard estimator of the risk of death for each cluster
Nelson-Aalen cumulative hazard estimator of the risk of death for each cluster derived from the population of incident late-onset epilepsy and age-, sex-, and GP practice-matched controls. Analysis inception date is that on which the patient was diagnosed with late-onset epilepsy or the same date in the age-, sex-, and GP practice-matched controls. Analysis time is in years. The horizontal-axis is analysis time in years whilst the vertical-axis is cumulative hazard.

### 3.3. Hazard of death

Analysis of the hazards of death revealed multiple tiers of risk. Of 10,259 matched controls, 994 (10%) died during the 10-year follow-up period. The hazard ratio for premature mortality in late-onset epilepsy (compared to matched controls), when controlling for all clustering variables, in addition to baseline dyslipidaemia, myocardial infarction, diabetes mellitus, chronic kidney disease, cirrhosis, TBI, CNS infection, rDDD, and enzyme-inducing ASMs was 1.73 (95% confidence interval [95%CI] 1.51–1.99).

When controlling for age at diagnosis, sex, and presence of dyslipidaemia, myocardial infarction, diabetes mellitus, chronic kidney disease, cirrhosis, TBI, CNS infection, rDDD, and enzyme-inducing ASMs at the time of diagnosis/matching date, all clusters had higher hazards of death compared to matched controls (log rank p-value <0.001; Table 3; Fig. 2; e-Figure 3) with no violation of the proportional hazards assumption (global p-value = 0.07).

Individual hazard varied between clusters, with six comparable to the base hazard from late-onset epilepsy and four significantly higher. The four clusters with higher hazards compared the base rate from epilepsy were 'Dementia and anxiety' (HR 5.36; 95%CI 3.31–8.68), 'Brain tumour' (HR 4.97; 95%CI 2.89–8.56), 'ICH and alcohol misuse' (HR 2.91; 95%CI 1.76–4.81) and 'ischaemic stroke' (HR 2.83; 95%CI 1.83–4.04).

In addition to cluster designation, age of diagnosis (HR 1.08 for each 1-year increment; 95%CI 1.06–1.10), myocardial infarction (HR 1.65; 95%CI 1.34–2.02), diabetes mellitus (HR 1.39; 95%CI 1.18–1.63), cirrhosis (HR 2.97; 95%CI 1.53–5.76), CNS infection (HR 3.20; 95%CI 1.75–5.87), rDDD (HR 1.33 for each 1.0 increase in rDDD; 95%CI 1.19–1.48), and enzyme-inducing ASMs (HR 1.31; 95%CI 1.02–1.70) were also independently associated with an increased hazard of premature mortality (Table 3).

When focusing on the four clusters with significantly higher risks, the hazard of premature mortality compared to matched controls was also higher than the base hazards attributed to their cluster defining characteristics alone. The base hazard ratios in the entire population related to dementia (HR 3.54; 95%CI 2.56–4.89), brain tumours (HR 3.12; 95% CI 2.1–4.65), ICH and alcohol misuse (HR 1.20; 95%CI 0.40–3.57), and ischaemic stroke (HR 1.19; 95%CI 0.83–1.70) were all lower than the late-onset epilepsy clusters that were defined by these conditions.

**Table 3**
Results of accelerated failure time model for death adjusting for baseline sociodemographic and clinical variables, and the Charlson comorbidity index. Clusters are treated as a factor variable with each one compared to the 1:10 age-, sex-, and GP practice-matched controls.

| Characteristic | Hazard ratio | 95% confidence interval | p-value |
|---|---|---|---|
| Clusters* | | | |
| Dementia and anxiety | 5.36 | 3.31–8.68 | <0.001 |
| Brain tumour | 4.97 | 2.89–8.56 | <0.001 |
| ICH and alcohol misuse | 2.91 | 1.76–4.81 | <0.001 |
| Ischaemic stroke | 2.83 | 1.83–4.40 | <0.001 |
| Depression male | 1.99 | 1.11–3.55 | 0.019 |
| Depression female | 1.93 | 1.10–3.39 | 0.022 |
| ICH | 1.77 | 1.02–3.07 | 0.041 |
| Healthy male | 1.63 | 1.24–2.14 | <0.001 |
| Anxiety | 1.62 | 1.07–2.48 | 0.022 |
| Healthy female | 1.53 | 1.07–2.20 | 0.020 |
| Age at diagnosis | 1.07 | 1.06–1.10 | <0.001 |
| Female sex | 0.73 | 0.64–0.84 | <0.001 |
| Dyslipidaemia | 0.95 | 0.82–1.09 | 0.468 |
| Myocardial infarction | 1.65 | 1.34–2.02 | <0.001 |
| Diabetes Mellitus | 1.39 | 1.18–1.63 | <0.001 |
| Chronic kidney disease | 1.12 | 0.93–1.35 | 0.219 |
| Cirrhosis | 2.97 | 1.53–5.76 | 0.001 |
| Prior TBI | 1.16 | 0.59–2.29 | 0.647 |
| Prior CNS infection | 3.20 | 1.75–5.87 | <0.001 |
| ASM rDDD | 1.33 | 1.19–1.48 | <0.001 |
| Enzyme inducing ASM | 1.31 | 1.02–1.70 | 0.034 |

ASM = antiseizure medications; CNS = central nervous system; ICH = intracerebral haemorrhage; rDDD = relative defined daily dose; TBI = traumatic brain injury.
* Clusters are all compared to the matched controls without epilepsy.

### 3.4. Causes of death

Of the 11,307 participants, 1148 died within 10-years of epilepsy onset or the matching date. A total of 840 (73%) consented to linkage with the ONS cause of death register and their death occurred during the overlapping reporting periods between CPRD and the ONS. Cancer,

cardiovascular, and respiratory aetiologies were the most frequently reported causes of death (e-Figure 4). Only three clusters contained people for whom epilepsy and/or seizures were listed as the primary cause of death ('ICH and alcohol misuse': 1 of 6 [17%] ONS reported deaths; 'ICH': 1 of 13 [8%] ONS reported deaths; and 'Healthy females': 1 of 26 [4%] ONS reported deaths; Table 4).

Seizures or epilepsy were rarely listed as one of the 15 secondary causes of death in the ONS register accounting for only 2 of 660 (0.3%) deaths in the matched controls and 13 of the 180 (7%) deaths in which epilepsy was not listed as the primary cause in late-onset epilepsy cases. These clusters comprised 'Anxiety' (4 of 124 ONS reported deaths; 3%), 'Healthy females' (1 of 204 ONS reported deaths, 0.5%), 'Dementia and anxiety' (1 of 59 ONS reported deaths; 2%), 'Healthy males' (4 of 312 ONS reported deaths; 1%), and 'Depression male' (3 of 57 ONS reported deaths; 5%; e-Figures 5 and 6).

## 4. Discussion

Using comorbidities typically associated with late-onset epilepsy, unsupervised machine learning can identify unique phenotypic clusters associated with varying hazards of premature mortality. Based on this analysis, the base hazard of premature mortality in older adults is elevated by approximately 73%, as compared to individuals from the general population without epilepsy (HR 1.73; 95%CI 1.51–1.99). However, certain agglomerations of baseline demographics and comorbidities elevates this risk, and thus contributes to the excess mortality seen in late-onset epilepsy. For instance, depending on the combination of underlying comorbidities, this base excess hazard of 73% can be elevated to 536% ('Dementia and anxiety'), 497% ('Brain tumour'), 291% ('ICH and alcohol misuse'), and 283% ('Ischaemic stroke'), thus corroborating the fact that dementia, tumours, and vascular causes of death are common in adults with epilepsy [39,40]. Whilst the degree to which seizures contribute to premature mortality in all people with post-stroke epilepsy remains controversial [41], in this study the cohort defined by ischaemic stroke had higher hazards of death compared to the base rate in late-onset epilepsy which is consistent with what has primarily been reported in the literature [39,42,43]. The remaining 6 clusters had hazards similar to the baseline risk from late-onset epilepsy, irrespective of the fact that some had a preponderance of mental health disorders, ICH, or a distinct lack of common comorbidities at baseline. Thus, the composite of associated comorbidities may either greatly elevate or conversely exert a negligible effect on this base rate depending on the constituent conditions.

The base hazard from epilepsy is particularly pertinent given the late-onset epilepsy severity appears mild. The rDDD was low across each cluster, suggesting a low seizure burden. Interpretation of the median rDDD, though, is tempered by the fact that clinicians may be cautious about using higher doses or ASM polytherapy in older adults. Futures studies linking such data to registries containing more granular details on anatomical lateralization, localization, and types and frequency of seizures, epilepsy severity will help clarify the severity of the underling epilepsy. Corroborating the postulate of a milder epilepsy, though, is the relative scarcity of seizures listed as a primary or secondary causes of death within each cluster (though we acknowledge that seizures are not often reported as the direct cause of death in adults with epilespy [40, 44])and the fact that essentially 80% of people with late-onset epilepsy achieve complete seizure freedom [5]. What is equally pertinent is that even for clusters defined by 'healthy males' and 'healthy females', the mortality hazard remains high. Thus, even though the underlying cause cannot be readily ascertained, the very development of late-onset epilepsy confers higher mortality risks independent of common aetiologies and comorbid conditions, an issue that should be discussed with patients even in the absence of an identifiable aetiology.

Our study is novel in that it has demonstrated that unique clusters of late-onset epilepsy can be reliably identified and characterized in late-onset epilepsy. Whilst most forms of neurological disease result in elevated hazards of premature death, the heterogenous nature of late-onset epilepsy results in multiple tiers of mortality risk. The presence of late-onset epilepsy as a composite whole appears to increase the hazard by between 29 and 88%. Hence, meticulous and concerted focus on achieving seizure-freedom is prudent, even if seizures are uncommonly reported as the primary cause of death. Whilst the rDDD of ASMs was associated with greater hazards of mortality, this was likely a surrogate for more severe epilepsies. Studies evaluating the role of ASMs, and improved seizure control, are required since this is unlikely to be a panacea for completely nullifying the increased risk of premature mortality. Rather, with specific composites of comorbid disease, the hazard of premature mortality in late-onset epilepsy can rise to over 500%, therefore equally emphasising the need to identify and treat cluster-defining characteristics. What is also novel is that this study provides evidence that the known bidirectional association that depression and anxiety share with epilepsy [21,22] may persist into later adulthood. However, this could be confounded by the fact that psychiatric symptoms may also be harbingers of neurodegenerative disease [45] that

**Table 4**
Primary causes of death as listed in the United Kingdom Office for National Statistics Death Register stratified by late-onset epilepsy clusters identified through agglomerative hierarchical clustering and age-, sex-, and GP practice-matched controls. Causes were ascertained using International Classification of Disease (ICD-10) chapters and codes.

| Cause of death | Healthy male | Brain tumour | Healthy female | Dementia & anxiety | Anxiety | Depression male | Depression female | Ischaemic stroke | ICH | ICH and alcohol misuse | Matched controls |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Accidental | 0 | 1 (5%) | 1 (4%) | 0 | 1 (6%) | 0 | 0 | 0 | 0 | 0 | 9 (1.5%) |
| Chronic kidney injury | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 (0.3%) |
| Cancer | 14 (28%) | 20 (90%) | 8 (31%) | 0 | 5 (28%) | 2 (20%) | 1 (13%) | 3 (33%) | 3 (23%) | 2 (32%) | 230 (35%) |
| Cardiovascular | 20 (40%) | 0 | 6 (23%) | 5 (28%) | 3 (17%) | 4 (40%) | 2 (24%) | 3 (33%) | 7 (53%) | 1 (17%) | 181 (27%) |
| Dementia | 3 (6%) | 0 | 1 (4%) | 2 (11%) | 2 (11%) | 0 | 1 (13%) | 1 (11%) | 0 | 0 | 25 (4%) |
| Epilepsy | 0 | 0 | 1 (4%) | 0 | 0 | 0 | 0 | 0 | 1 (8%) | 1 (17%) | 0 |
| Hepatic | 1 (2%) | 1 (5%) | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 (17%) | 14 (2%) |
| Infection | 0 | 0 | 0 | 0 | 0 | 0 | 1 (13%) | 0 | 0 | 1 (17%) | 13 (2%) |
| Neurological | 3 (6%) | 0 | 1 (4%) | 9 (50%) | 1 (6%) | 0 | 0 | 0 | 1 (8%) | 0 | 34 (5%) |
| Other | 2 (4%) | 0 | 2 (7%) | 0 | 0 | 2 (20%) | 0 | 0 | 0 | 0 | 53 (8%) |
| Psychiatric | 0 | 0 | 0 | 0 | 1 (6%) | 0 | 0 | 0 | 0 | 0 | 1 (0.2%) |
| Respiratory | 7 (14%) | 0 | 6 (23%) | 2 (11%) | 5 (28%) | 2 (20%) | 3 (37%) | 2 (23%) | 1 (8%) | 0 | 98 (15%) |

*Neurological disease excludes dementia and epilepsy/seizures; cardiovascular disease includes stroke.

could lead to epilepsy, as implicated by the 'Anxiety and dementia' cluster.

This study benefits from the large, population-based cohort drawn from over 13 million people. We have robust linkage schemes between primary, secondary, and tertiary health centres, pharmacy, and national statistics data which allowed us to track patient flow across trajectories all of care. We used an epilepsy case definition with a high sensitivity and specificity [17], as well as validated case definitions for all other comorbid conditions [20]. We confirmed and quantified the performance and reliability of the clustering through the silhouette coefficient, Davies-Bouldin index, and a cluster stability index. All survival analyses were adjusted for common risk factors for premature mortality and our point estimate for the hazard related to late-onset epilepsy alone (HR 1.73; 95%CI 1.51–1.99) is remarkably consistent with that reported in other Western populations (HR 1.72; 95%CI 1.55–1.90) [10]. The overall incidence proportion for late-onset epilepsy is also consistent with what is expected in this age range over the study period [3].

Our study is not without limitations. Only 73% of participants had consented to have their ONS data linked to CPRD, meaning this information was not available for all patients. This may have led to non-differential bias since missing data are expected to be random with respect to the cause of death. Approximately 2% of the controls were exposed to enzyme-inducing ASMs. This may be due to inadvertent inclusion of false positives in matched controls, whereby a physician may have diagnosed epilepsy, prescribed an ASM, and failed to enter a Read or ICD-10 code. However, the SAIL definition does have a sensitivity of 87% for adult epilepsy, and thus whilst it is possible that the controls may have been contaminated with false positive cases, the more likely explanation is that medications such as carbamazepine are used to treat other conditions, such as trigeminal neuralgia, neuropathic pain, and psychiatric conditions such as bipolar disorder which would constitute a large portion of this 2%. Alternatively, the diagnosis of epilepsy in older adults can be challenging with seizures potentially presenting as subtle episodes of confusion or involving multiple phases of evolution [46], meaning false positive and negative diagnoses may have occurred, leading to conservative estimates due to non-differential misclassification bias. We also categorised alcohol misuse as 'misuse' *versus* 'low--risk/drinking status not specified' to provide a conservative estimate of effect. If the low-risk group were contaminated by the inclusion of higher risk drinkers, the effect estimate would be diluted. Here, despite this, alcohol misuse emerges as cluster-defining feature even despite this, indicating that even when used conservatively this is a powerful characteristic that segregates patient groups. The accuracy of the reported cause of death may be prone to misclassification since it derived from the death certificate completed by a medical practitioner on certification. Hence, not all cases have undergone autopsy confirmation and there is the possibility that seizures and epilepsy were neglected as a cause in coroner's reports in people that died of sudden death [47]. However, sudden death was rare in this population and it is reassuring that the coroner can only register deaths after investigations are concluded and when they are satisfied that each case was thoroughly investigated with a correctly certified cause of death [48]. It may also be that seizures are underreported, and that non-neurologists consider them a 'natural' part of concomitant comorbidities when in their terminal stages. Hence, they may not be as prone to document the seizures in medical records when they occur late in conditions such as stroke and dementia. However, the age of diagnosis of late-onset epilepsy in our cohort was relatively young (median age 68; IQR 66–72) meaning, if anything, epilepsy was likely diagnosed relatively early in the disease process rather than in a 'terminal phase'. Furthermore, even if those that developed seizures at older ages (close to the terminal stage of their disease) were not reported or captured in our study, the fact remains that late-onset epilepsy when diagnosed close to age 65 (when patients are unlikely to be in the late stages of dementia and stroke) is still associated with universally elevated risks of death irrespective of the underlying comorbid profile.

Unlike supervised machine learning, concepts of external validation are challenging to apply to unsupervised clustering due to the lack of 'ground truth' that prevents comparisons of predicted to observed [49]. Thus, formal external validation is not conventionally part of a clustering process given no inviolate or immutable cluster labels exist. Unlike supervised machine learning, where one knows the outcome/class label (for example, the patient is either dead or alive), no such objective cluster label is available in unsupervised machine learning, meaning it is not possible to compare predicted to observed placement. Rather, the cluster to which a person is assigned is entirely dependant on the population dynamics that surrounds them. Therefore, the same person could be placed in a unique cluster if they move to an independent population with a different distribution of clinical characteristics. Having said this, the 10 clusters derived here are independent (Davies-Bouldin index and silhouette coefficient) and remain stable when exposed to data perturbations (cluster stability index). Additionally, given the population-based nature of this study, it can be asserted that they are generalizable to the general UK population. Although the same clusters may not be derived from other populations, as their dynamics will differ from the UK, this study importantly provides a framework through which they can be derived and compared to the UK and other countries as this work is replicated.

Using large data sources and advanced analytics, we are beginning to move from counselling patients about the 'average' risk of death for a person with late-onset epilepsy to more personalised estimates based on aspects such as cluster designation. Thus, we can quickly identify these patients with disproportionately high hazards of death and attempt to intervene where feasible. The implications of this study are manifold. First, late-onset epilepsy itself portends to higher hazards of premature mortality. Studies are urgently needed to determine the relative degree to which improved seizure control, management of underlying aetiologies and comorbidities, and optimal ASM prescription patterns can reduce this risk. A holistic approach is likely required for this population, though these studies may allow us to preferentially intervene on the reversible risk factors that confer the highest risks in resource and time-constrained clinics. Second, targeted identification of those clusters, especially those at highest risk ('Dementia and anxiety', 'Brain tumour', 'ICH and alcohol misuse', and 'Ischaemic stroke'), will promote precision intervention. Future randomised controlled trials can try to treat cluster-defining characteristics to reduce mortality and identify populations most likely to benefit from directed intervention, thus increasing the signal to noise ratio. Finally, embedding such algorithms in EMR systems may help identify patients at highest risk, thus flagging them for concerted attention by the health-care team. Such personalised approaches will promote expedient management and optimization of concomitant disease, screening for occult cerebrovascular and neurodegenerative disease, and empower patients by offering more individualised estimates of their predicted disease course.

## Role of the funding sources

This analysis was in part funded by the Canadian Frailty Network (CAT2017–19) which is supported by the Government of Canada through the Networks Centres of Excellence (NCE) program, as well as through the University of Calgary. These funders did not play any role in the design, execution, analysis, or dissemination of the results for this study. The corresponding author (CBJ) has full access to all data in the study and had final responsibility for the decision to submit for publication.

## Publishing and research ethics statement

We confirm that we have read the Journal's position on issues involved in ethical publication and affirm that this report is consistent with those guidelines.

## Declaration of Competing Interest

C.B. Josephson has received grant support from Epilepsy Canada and unrestricted educational grants from UCB Pharma Inc. and Eisai Inc. for work unrelated to this project.

A. Gonzalez-Izquierdo reports no disclosures relevant to the manuscript.

J.D.T. Engbers reports no disclosures relevant to the manuscript.

S. Denaxas reports no disclosures relevant to the manuscript.

G. Delgado-Garcia reports no disclosures relevant to the manuscript.

T.T. Sajobi reports no disclosures relevant to the manuscript.

M. Wang reports no disclosures relevant to the manuscript.

M.R. Keezer has received unrestricted educational grants, as well as grants for investigator-initiated studies, from UCB Pharma Inc. and Eisai Inc. for work unrelated to this project.

S. Wiebe has received unrestricted educational grants from UCB Pharma, Eisai and Sunovion for work unrelated to this project.

## Acknowledgements

This study is based in part on data from the Clinical Practice Research Datalink (CPRD) obtained under license from the United Kingdom (UK) Medicines and Healthcare products Regulatory Agency. The data are provided by patients and collected by the NHS as part of their care and support. The interpretation and conclusions contained in this study are those of the author/s alone. Hospital Episode Statistics and Office for National Statistics (Copyright © 2019) was re-used with the permission of The Health & Social Care Information Centre. All rights reserved.

## Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.seizure.2023.07.016.

## References

[1] Josephson CB, Engbers JDT, Sajobi TT, et al. Towards a clinically informed, data-driven definition of elderly onset epilepsy. Epilepsia 2016;57(2):298–305. https://doi.org/10.1111/epi.13266.

[2] Hernández-Ronquillo L, Adams S, Ballendine S, Téllez-Zenteno JF. Epilepsy in an elderly population: classification, etiology and drug resistance. Epilepsy Res 2018; 140:90–4. https://doi.org/10.1016/j.eplepsyres.2017.12.016.

[3] Sillanpää M, Gissler M, Schmidt D. Efforts in Epilepsy Prevention in the Last 40 Years: lessons From a Large Nationwide Study. JAMA Neurol 2016;73(4):390–5. https://doi.org/10.1001/jamaneurol.2015.4515.

[4] Cloyd J, Hauser W, Towne A, et al. Epidemiological and medical aspects of epilepsy in the elderly. Epilepsy Res 2006;68:39–48. https://doi.org/10.1016/j.eplepsyres.2005.07.016.

[5] Alsfouk BAA, Hakeem H, Chen Z, Walters M, Brodie MJ, Kwan P. Characteristics and treatment outcomes of newly diagnosed epilepsy in older people: a 30-year longitudinal cohort study. Epilepsia 2020;61(12):2720–8. https://doi.org/10.1111/epi.16721.

[6] Stephen LJ, Kelly K, Mohanraj R, Brodie MJ. Pharmacological outcomes in older people with newly diagnosed epilepsy. Epilepsy Behav 2006;8(2):434–7. https://doi.org/10.1016/j.yebeh.2005.11.007.

[7] Bagshaw J, Crawford P, Chappell B. Care in people 60 years of age and over with chronic or recently diagnosed epilepsy: a note review in United Kingdom general practice. Seizure 2009;18(1):57–60. https://doi.org/10.1016/j.seizure.2008.06.003.

[8] Kwan P, Brodie MJ. Early Identification of Refractory Epilepsy. N Engl J Med 2000; 342(5):314–9. https://doi.org/10.1056/NEJM200002033420503.

[9] Winter Y, Daneshkhah N, Galland N, Kotulla I, Krüger A, Groppa S. Health-related quality of life in patients with poststroke epilepsy. Epilepsy Behav 2018;80:303–6. https://doi.org/10.1016/j.yebeh.2017.12.037.

[10] Johnson EL, Krauss GL, Kucharska-Newton A, Lam AD, Sarkis R, Gottesman RF. Mortality in patients with late-onset Epilepsy: results from the Atherosclerosis Risk in Communities Study. Neurology 2021;97(11):e1132–40. https://doi.org/10.1212/WNL.0000000000012483.

[11] Denaxas S, Gonzalez-Izquierdo A, Fitzpatrick N, Direk K, Hemingway H. Phenotyping UK Electronic Health Records from 15 million individuals for precision medicine: the CALIBER resource. Stud Health Technol Inform 2019;262: 220–3. https://doi.org/10.3233/SHTI190058.

[12] Denaxas SC, George J, Herrett E, et al. Data Resource Profile: cardiovascular disease research using linked bespoke studies and electronic health records (CALIBER). Int J Epidemiol 2012;41(6):1625–38. https://doi.org/10.1093/ije/dys188.

[13] Chisholm J. The read clinical classification. BMJ 1990;300(6732):1092. https://doi.org/10.1136/bmj.300.6732.1092.

[14] García Rodríguez LA, Pérez Gutthann S. Use of the UK General Practice Research Database for pharmacoepidemiology: GPRD and pharmacoepidemiology. Br J Clin Pharmacol 2002;45(5):419–25. https://doi.org/10.1046/j.1365-2125.1998.00701.x.

[15] British Medical Association Royal Pharmaceutical Society of Great Britain, British medial association. *BNF 57 March 2009*. BMJ Publishing RPS Publishing; 2009.

[16] Herrett E, Gallagher AM, Bhaskaran K, et al. Data resource profile: clinical practice research datalink (CPRD). Int J Epidemiol 2015;44(3):827–36. https://doi.org/10.1093/ije/dyv098.

[17] Fonferko-Shadrach B, Lacey AS, White CP, et al. Validating epilepsy diagnoses in routinely collected data. Seizure 2017;52:195–8. https://doi.org/10.1016/j.seizure.2017.10.008.

[18] Ghosh RE, Crellin E, Beatty S, Donegan K, Myles P, Williams R. How clinical practice research datalink data are used to support pharmacovigilance. Ther Adv Drug Saf 2019;10. https://doi.org/10.1177/2042098619854010.

[19] van Bodegraven B, Palin V, Mistry C, et al. Infection-related complications after common infection in association with new antibiotic prescribing in primary care: retrospective cohort study using linked electronic health records. BMJ Open 2021; 11(1):e041218. https://doi.org/10.1136/bmjopen-2020-041218.

[20] Denaxas S, Gonzalez-Izquierdo A, Direk K, et al. UK phenomics platform for developing and validating electronic health record phenotypes: CALIBER. J Am Med Inform Assoc 2019;26(12):1545–59. https://doi.org/10.1093/jamia/ocz105.

[21] Josephson CB, Lowerison M, Vallerand I, et al. Association of depression and treated depression with Epilepsy and Seizure outcomes: a multicohort analysis. JAMA Neurol 2017;74(5):533–9. https://doi.org/10.1001/jamaneurol.2016.5042.

[22] Hesdorffer DC, Ishihara L, Mynepalli L, Webb DJ, Weil J, Hauser WA. Epilepsy, suicidality, and psychiatric disorders: a bidirectional association. Ann Neurol 2012; 72(2):184–91. https://doi.org/10.1002/ana.23601.

[23] Burneo JG, Antaya TC, Allen BN, Belisle A, Shariff SZ, Saposnik G. The risk of new-onset epilepsy and refractory epilepsy in older adult stroke survivors. Neurology 2019;93(6):e568–77. https://doi.org/10.1212/WNL.0000000000007895.

[24] Lahti AM, Saloheimo P, Huhtakangas J, et al. Poststroke epilepsy in long-term survivors of primary intracerebral hemorrhage. Neurology 2017;88(23):2169–75. https://doi.org/10.1212/WNL.0000000000004009.

[25] Subota A, Pham T, Jetté N, Sauro K, Lorenzetti D, Holroyd-Leduc J. The association between dementia and epilepsy: a systematic review and meta-analysis. Epilepsia 2017;58(6):962–72. https://doi.org/10.1111/epi.13744.

[26] Loiseau J, Loiseau P, Duché B, Guyot M, Dartigues JF, Aublet B. A survey of epileptic disorders in southwest France: seizures in elderly patients. Ann Neurol 1990;27(3):232–7. https://doi.org/10.1002/ana.410270304.

[27] Samokhvalov AV, Irving H, Mohapatra S, Rehm J. Alcohol consumption, unprovoked seizures, and epilepsy: a systematic review and meta-analysis. Epilepsia 2010;51(7):1177–84. https://doi.org/10.1111/j.1528-1167.2009.02426.x.

[28] Hesdorffer DC, Tian H, Anand K, et al. Socioeconomic Status Is a Risk Factor for Epilepsy in Icelandic Adults but Not in Children. Epilepsia 2005;46(8):1297–303. https://doi.org/10.1111/j.1528-1167.2005.10705.x.

[29] Jordan H, Roderick P, Martin D. The Index of Multiple Deprivation 2000 and accessibility effects on health. J Epidemiol Community Health 2004;58(3):250–7. https://doi.org/10.1136/jech.2003.013011.

[30] Subota A, Jetté N, Josephson CB, et al. Risk factors for dementia development, frailty, and mortality in older adults with epilepsy – a population-based analysis. Epilepsy Behav 2021;120:108006. https://doi.org/10.1016/j.yebeh.2021.108006.

[31] Clegg A, Bates C, Young J, et al. Development and validation of an electronic frailty index using routine primary care electronic health record data. Age Ageing 2016; 45(3):353–60. https://doi.org/10.1093/ageing/afw039.

[32] Schafer J. Analysis of incomplete multivariate data. Chapman & Hall/CRC Press; 1997.

[33] Ward JH. Hierarchical grouping to optimize an objective function. J Am Stat Assoc 1963;58(301):236–44. https://doi.org/10.1080/01621459.1963.10500845.

[34] Raschka S, Mirjalili V. Python machine learning: machine learning and deep learning with python, scikit-learn, and tensorflow. Packt Publishing; 2017. Second Edition.

[35] Grambsch PM, Therneau TM. Proportional hazards tests and diagnostics based on weighted residuals. Biometrika 1994;81(3):515–26. https://doi.org/10.1093/biomet/81.3.515.

[36] Josephson CB, Wiebe S, Delgado-Garcia G, et al. Association of enzyme-inducing antiseizure drug use with long-term Cardiovascular disease. JAMA Neurol 2021. https://doi.org/10.1001/jamaneurol.2021.3424.

[37] StataCorp. Stata statistical software: release 16. Published online 2019.

[38] Pedregosa F, Varoquax G, Gramfort A. Scikit-learn: maching learning in Python. J Mach Learn Res 2011;12:2825–30.

[39] Hrabok M, Engbers JDT, Wiebe S, et al. Primary care electronic medical records can be used to predict risk and identify potentially modifiable factors for early and late death in adult onset epilepsy. Epilepsia 2021;62(1):51–60. https://doi.org/10.1111/epi.16738.

[40] Keezer MR, Bell GS, Neligan A, Novy J, Sander JW. Cause of death and predictors of mortality in a community-based cohort of people with epilepsy. Neurology 2016;86(8):704–12. https://doi.org/10.1212/WNL.0000000000002390.

[41] van Tuijl JH, van Raak EPM, van Oostenbrugge RJ, Aldenkamp AP, Rouhl RPW. The occurrence of seizures after ischemic stroke does not influence long-term

mortality; a 26-year follow-up study. J Neurol 2018;265(8):1780–8. https://doi.org/10.1007/s00415-018-8907-7.
[42] Hansen J, Åsberg S, Kumlien E, Zelano J. Cause of death in patients with poststroke epilepsy: results from a nationwide cohort study. PLoS ONE 2017;12(4):e0174659. https://doi.org/10.1371/journal.pone.0174659.
[43] Sinka L, Abraira L, Imbach LL, et al. Association of mortality and risk of Epilepsy with type of acute symptomatic seizure after ischemic stroke and an updated prognostic model. JAMA Neurol 2023;80(6):605–13. https://doi.org/10.1001/jamaneurol.2023.0611.
[44] Karlander M, Ljungqvist J, Sörbo A, Zelano J. Risk and cause of death in post-traumatic epilepsy: a register-based retrospective cohort study. J Neurol 2022;269(11):6014–20. https://doi.org/10.1007/s00415-022-11279-5.
[45] Byers AL, Yaffe K. Depression and risk of developing dementia. Nat Rev Neurol 2011;7(6):323–31. https://doi.org/10.1038/nrneurol.2011.60.
[46] Ghosh S, Jehi LE. New-onset epilepsy in the elderly: challenges for the internist. Cleve Clin J Med 2014;81(8):490–8. https://doi.org/10.3949/ccjm.81a.13148.
[47] Lühdorf K, Jensen LK, Plesner AM. Epilepsy in the elderly: life expectancy and causes of death. Acta Neurol Scand 1987;76(3):183–90. https://doi.org/10.1111/j.1600-0404.1987.tb03565.x.
[48] Medicines & Healthcare Products Regulatory Agency. ONS death registration data and CPRD primary care data documentation (set 21), version 2.5. Published online May 28, 2021.
[49] Tarekegn AN, Michalak K, Giacobini M. Cross-validation approach to evaluate clustering algorithms: an experimental study using multi-label datasets. SN Comput Sci 2020;1(5):263. https://doi.org/10.1007/s42979-020-00283-z.