Digital Personalized Health and Medicine
L.B. Pape-Haugaard et al. (Eds.)

doi:10.3233/SHTI200210

Using Unsupervised Learning to Identify Clinical Subtypes of Alzheimer's Disease in Electronic Health Records

Nonie ALEXANDER[ab1], Daniel C. ALEXANDER[c], Frederik BARKHOF[cdefg], and Spiros DENAXAS[abh]

[a]Institute of Health Informatics, University College London, London, UK, [b]Health Data Research UK London, [c]Centre for Medical Image Computing, Department of Medical Physics and Biomedical Engineering, University College London, London, UK, [d]UCL Institute of Neurology, University College London, London, [e]UK/Translational Imaging Group, UK/Department of Radiology and Nuclear Medicine, [f]VU University Medical Center, Amsterdam, The Netherlands, [g]National Institute for Health Research University College London Hospitals Biomedical Research Centre, London, UK, [h]Alan Turing Institute, London Uk,

Abstract. Identifying subtypes of Alzheimer's Disease (AD) can lead towards the creation of personalized interventions and potentially improve outcomes. In this study, we use UK primary care electronic health records (EHR) from the CALIBER resource to identify and characterize clinically-meaningful clusters patients using unsupervised learning approaches of MCA and K-means. We discovered and characterized five clusters with different profiles (mental health, non-typical AD, typical AD, CVD and men with cancer). The mental health cluster had faster rate of progression than all the other clusters making it a target for future research and intervention. Our results demonstrate that unsupervised learning approaches can be utilized on EHR to identify subtypes of heterogeneous conditions.

Keywords. Phenotyping, Alzheimer's disease, Electronic health records, machine learning

1. Introduction

Alzheimer's disease (AD) is a highly heterogeneous disease. Any two individuals with the disease display a different array of symptoms or progression rate[1]. Progression rate can be increased by many factors such as higher education level and comorbidities like diabetes[2]. Gaining a complete view of a patient's profile of symptoms, comorbidities and demographic factors can enable the discovery of different progression patterns and to personalise treatments for AD patients.

Almost all subtyping studies to date in AD focuses on two data types; cognitive tests[3] and brain scans[4], showing cluster(s) of hippocampal atrophy and more diffuse atrophy, corresponding to specific memory problems and more global cognitive

1 Corresponding Author, Nonie Alexander, Institute of Health Informatics, University College London, London, UK; E-mail: nonie.alexander.16@ucl.ac.uk

problems, leading to different patterns of progression[4]. Subtyping using electronic health records (EHR) has been carried out in many diseases[5][6] and are ideal data sets to apply these methods to due to the large sample size and the breadth of clinical information[7] they contain. Variables and outcomes are lifted directly from clinical data making them directly relevant to the patient's clinical management. This work aims to identify subtypes of AD using a range of clinical information on the patients.

2. Methods

2.1. Data and cohort

We used data from the Clinical Practice Research Datalink (CPRD)[8] which contains longitudinal primary care EHR from general practices in the United Kingdom[8]. Only patients from practices which have been marked up to standard are used and only data collected after the practice was found to be up to standard were used. Data were extracted and phenotypes defined using the CALIBER data resource[9,10]. We classified patients as cases if they had at least one AD Read code and no other future diagnosis indicating a different dementia subtype diagnosis[11]. Full case analysis was used. This study was approved to use CPRD data by the Independent Scientific Advisory Committee (ref. 18_111).

2.2. Variables and Outcomes

Three categories of variables were included in the analysis; symptoms, comorbidities and demographic and lifestyle factors. Symptoms and comorbidities were identified through a systematic literature review of studies identifying symptoms or finding associated comorbidities. There resulting symptoms found were memory problems, confusion, neuropsychiatric problems and motor problems. From them the disease that were identified where atrial fibrillation, anxiety, hyperglycaemia, rheumatoid arthritis, stroke, hearing loss, depression, kidney disease, heart failure, atherosclerosis and cancer. Age of onset, gender, drinking status and smoking status were all included as demographic variables in the analysis. The phenotypes for each symptom was ideally defined using previously defined and verified CALIBER phenotypes[9,10], if they were not available, using a definition from a previous studies.

We defined five clinical outcomes which were used to evaluate clusters: a) length of time to treatment discontinuation[12], b) rate of progression measured by the Mini Mental State Exam[13], c) healthcare utilization defined as number of appointments and missed appointments per year, d) all-cause mortality and time to assisted living, and e) time to assisted living. For the latter two, we created Kaplan-Meyer survival curves.

2.3. Statistical analysis

We used Multiple Correspondence Analysis (MCA) for dimensionality reduction[14]. K-means was used to identify clusters: First, *k* was decided through an elbow plots for the cluster entropy , silhouette coefficient and Bayesian Information Criterion (BIC). K-means was executed 100 times to establish the optimum solution determined by the

inter cluster variation. To evaluate the clusters silhouette score were used to measure cluster structure and cluster stability was measured through Jaccard index of cluster results based on a 100 X bootstrapped sample.

3. Results

We identified 7,913 AD patients (66.2% female, mean age 82.1, 81.3-83 95% CI). Using the first five MCA principal components, we identified five clusters: mental health cluster which has the highest prevalence of anxiety and depression comorbidities, non-typical AD which had the lowest prevalence of memory symptoms and highest of other symptoms, typical AD which had high prevalence of memory symptoms and low other symptoms, cardiovascular disease (CVD) cluster with high prevalence of cardiovascular diseases such as atrial fibrillation and associated diseases such as diabetes and a cluster of men with cancer (Table 1, Figure1) all with differing clinical outcomes (Figure 2). The clusters had a silhouette score of 0.19 (showing weak cluster structure) and mean Jaccard index of 0.78 indicating overall stability.

Table 1. Each cluster name, key variables that characterise the cluster and their outcomes. Green - highest outcome. red - lowest outcome, * significant

Cluster I.D	Demographics	Top Characteristics (figure 1)	Outcomes (figure 2)
1. Mental Health Cluster	n = 1528 80.1% female onset age = 75 Diag MMSE = 21.6	Neuropsychiatric symptoms, (83.51%) Anxiety (62.11%) Depression (69.24%)	Fastest rate of progression* (2.16 mmse points per year, 2.06-2.26 95% CI)
2. Non- typical AD Cluster	n = 1640 81.3% female onset age = 86 Diag MMSE = 20.9	Memory symptoms (44.15%) Confusion(71.04%) Neuropsychiatric symptoms (98.35%)	Shortest time till treatments discontinuation (3.09 years, 3.19-2.99 95% CI) Lowest survival probability* Shortest time till assisted living *
3. Typical AD Cluster	n = 2026 90.1% female onset age = 82 Diag MMSE = 21.7	Memory symptoms (89.09%) Confusion(0.05%) Neuropsychiatric symptoms (18.85%)	Fewest appointments per year* (16.6 apps per year, 16.15-17.1, 95% CI)
4. Cardiovascular disease Cluster	n = 686 45.9% female onset age = 81 Diag MMSE = 21.4	Diabetes (84.26%) Hypertension(81.49%) Atrial fibrillation(20.41 %)	Greatest number of appointments per year* (24.8 apps , 23.7-25.9 95% CI) and missed appointments per year* (0.96 missed apps, 0.84 - 1.08 95% CI)
5. Memory problems and Cancer Cluster	n = 1710 19.0% female onset age = 79 Diag MMSE = 22.6	Memory symptoms (90.94%) Cancer(43.27%) Male (80.99%)	Slowest progression rate (1.59 mmse points a year, 1.5-1.68 95% CI)

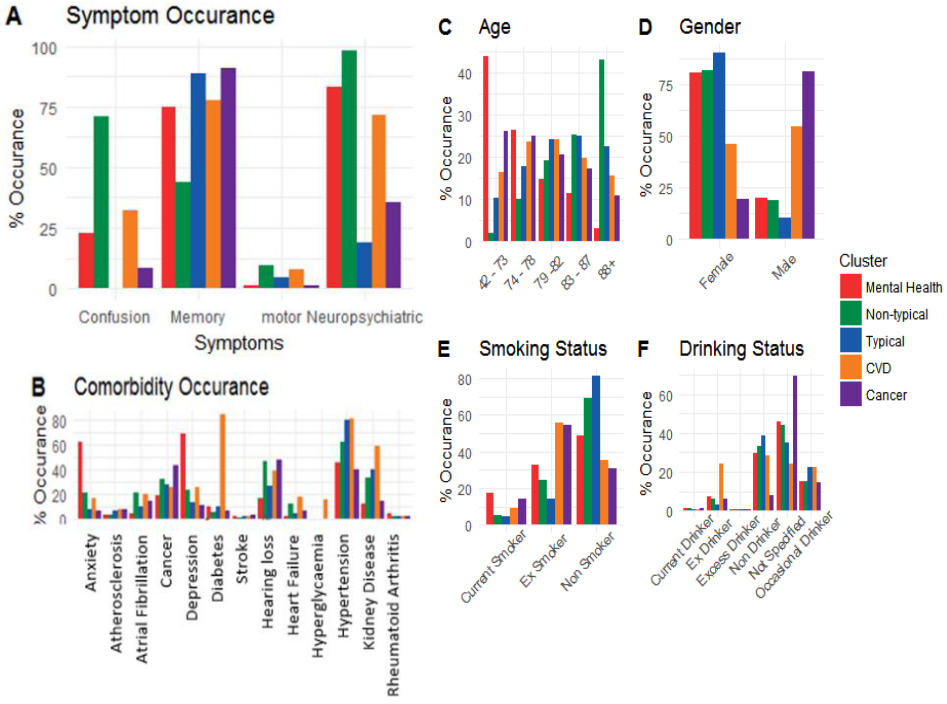

Figure 1: Distribution of the variables included in the cluster analysis for each cluser a) symptoms, b) Comorbidities, c) age, d) gender, e) clinically-recorded smoking status, f) clinically-recorded drinking status

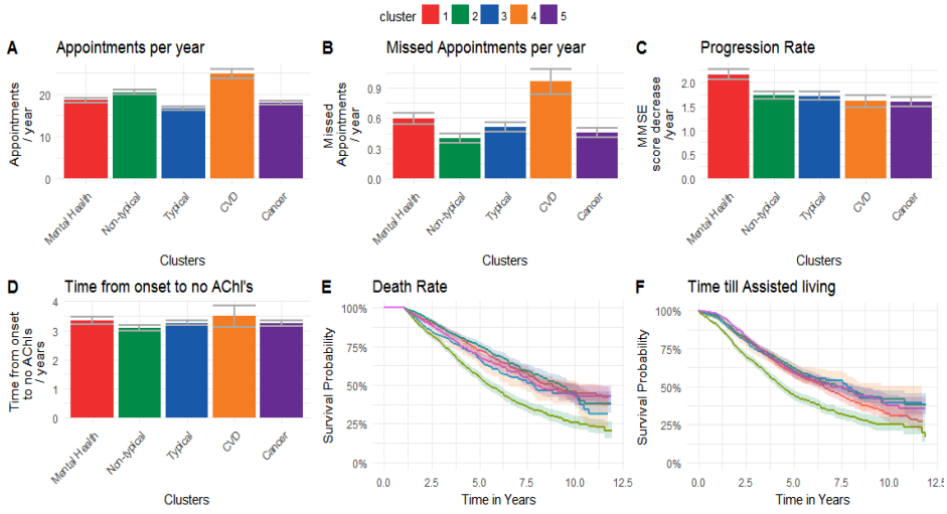

Figure 2 : Outcomes difference for variables not included in the cluster analysis per cluster: a) mean number of appointments per year with CI, b) mean missed appointments per year with 95% CI, c) mean MMSE score reduction per year with 95% CI, d) mean time to treatment discontinuation with 95% CI, e) Kaplan Meyer curve of survival probability, f) Kaplan Meyer curve on time till assisted living.

4. Discussion

Using unsupervised machine learning on a cohort of 7,913 AD patients defined through EHR, we identified five distinct clusters with different clinical profiles: a mental health cluster, a non-typical AD cluster, a typical AD cluster and clusters with mostly cardiovascular problems and a cluster with cancer. The mental health and early onset cluster had a faster rate of progression. Further research is needed to delineate the relationship between the mental health comorbidities, rate of progression and early onset. Identifying the CVD cluster can have clinical benefit as they have the highest healthcare utilization. The two clusters of typical and non-typical AD are similar to clusters found in previous research splitting patients with memory issues and patients with other cognitive issues[3]. The cluster with high prevalence of cancer, and men reported low prevalence of other symptoms aside from memory loss. There should be more investigation into whether the mechanism that leads cancer to protect against AD [2] also reduces the number of symptoms through less global atrophy. Using this data type uses a unique perspective on which to cluster patients as it offers a longitudinal and broad scope of a patient. The cluster results also have clinically relevant and impactful outcomes. There are however several weaknesses to the study such as there are some that are important but not or only partly recorded such as family history of AD recorded in EHR. Future work will look into validating these results through comparing this method with other cluster methods cluster results using imaging and cognitive test results.

References

[1] D. Mungas et.al, The effects of age on rate of progression of Alzheimer disease and dementia with associated cerebrovascular disease, *Arch. Neurol.* **58** (2001) 1243–1247.
[2] J.-Q. Li, et.al, Risk factors for predicting progression from mild cognitive impairment to Alzheimer's disease *Journal of Neurology, Neurosurgery & Psychiatry*. **87** (2016) 476–484. d
[3] N.M.E. Scheltens, et.al, Cognitive subtypes of probable Alzheimer's disease robustly identified in four cohorts, *Alzheimers. Dement.* **13** (2017) 1226–1236.
[4] A.L. Young, et.al Uncovering the heterogeneity and temporal complexity of neurodegenerative diseases with Subtype and Stage Inference, *Nat. Commun.* **9** (2018) 4273.
[5] L. Li, et.al, Identification of type 2 diabetes subgroups through topological analysis of patient similarity, *Sci. Transl. Med.* **7** (2015) 311ra174.
[6] T. Ahmad, et.al, Clinical implications of chronic heart failure phenotypes defined by cluster analysis, *J. Am. Coll. Cardiol.* **64** (2014) 1765–1774.
[7] H. Hemingwayet.al, Big data from electronic health records for early and late translational cardiovascular research: challenges and potential, *Eur. Heart J.* **39** (2018) 1481–1495.
[8] E. Herrett, A.M. Gallagher, K. Bhaskaran, H. Forbes, R. Mathur, T. van Staa, and L. Smeeth, Data Resource Profile: Clinical Practice Research Datalink (CPRD), *Int. J. Epidemiol.* **44** (2015) 827–836.
[9] S. Denaxas, et.al, UK phenomics platform for developing and validating electronic health record phenotypes: CALIBER, *J. Am. Med. Inform. Assoc.* (2019). doi:10.1093/jamia/ocz105.
[10] V. Kuan, S. Denaxas, A. Gonzalez-Izquierdo, K. Direk, O. Bhatti, S. Husain, S. Sutaria, M. Hingorani, D. Nitsch, C.A. Parisinos, R.T. Lumbers, R. Mathur, R. Sofat, J.P. Casas, I.C.K. Wong, H. Hemingway, and A.D. Hingorani, A chronological map of 308 physical and mental health conditions from 4 million individuals in the English National Health Service, *Lancet Digit Health*. **1** (2019) e63–e77.
[11] M. Pujades-Rodriguez, et.al, The diagnosis, burden and prognosis of dementia: A record-linkage cohort study in England, *PLoS One*. **13** (2018) e0199026.
[12] National Institute for Health and Care Excellence (UK), Dementia: Assessment, management and support for people living with dementia and their carers, London, 2018.
[13] S. Dubin, The Mini Mental State Exam, *The American Journal of Nursing*. **98** (1998) 16D.
[14] B. Le Roux, and H. Rouanet, Multiple Correspondence Analysis, SAGE, 2010.