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

doi:10.3233/SHTI200204

Trajectories of Disease Accumulation Using Electronic Health Records

Pere PLANELL-MORELL [a,1], Madhavi BAJEKAL [b], Spiros DENAXAS [c], Rosalind RAINE [b], and Daniel C. ALEXANDER [a]

[a] *Centre for Medical Image Computing, University College London*
[b] *Department of Applied Health Research, University College London*
[c] *Institute of Health Informatics, University College London*

Abstract. Multimorbidity is a major problem for patients and health services. However, we still do not know much about the common trajectories of disease accumulation that patients follow. We apply a data-driven method to an electronic health record dataset (CPRD) to analyse and condense the main trajectories to multimorbidity into simple networks. This analysis has never been done specifically for multimorbidity trajectories and using primary care based electronic health records. We start the analysis by evaluating temporal correlations between diseases to determine which pairs of disease appear significantly in sequence. Then, we use patient trajectories together with the temporal correlations to build networks of disease accumulation. These networks are able to represent the main pathways that patients follow to acquire multiple chronic conditions. The first network that we find contains the common diseases that multimorbid patients suffer from and shows how diseases like diabetes, COPD, cancer and osteoporosis are crucial in the disease trajectories. The results we present can help better characterize multimorbid patients and highlight common combinations helping to focus treatment to prevent or delay multimorbidity progression.

Keywords. multimorbidity, electronic health records, disease trajectory, multiple disease progression model

1. Introduction

Multimorbidity (the coexistence of two or more chronic diseases in an individual) is a growing health and healthcare challenge. There is significant evidence showing that the prevalence of patients with multiple chronic conditions has increased over time [1]. Nevertheless, we still do not know the temporal dynamics of disease accumulation over time in the development of multimorbidity. In this paper we study which are the trajectories of disease accumulation that patients follow to become multimorbid. Understanding what are the characteristics of these trajectories can be helpful to better identify and predict the disease progression of patients. We adapt the methodology proposed by Jensen et al [2] and we apply it to primary care electronic health records (EHRs) instead of secondary care EHRs. If we want to study multimorbidity we need to use primary care data since most chronic condition are treated in general practice. However, it is challenging

[1] Corresponding Author: Pere Planell-Morell; E-mail: p.morell@ucl.ac.uk

Table 1. List of chronic conditions in scope

Atrial Fibrillation (AF)	Anxiety	Alcohol problems
COPD	Coronary Heart Disease (CHD)	Chronic Kidney Disease (CKD)
Cancer	Chronic liver disease (CLD)	Depression
Diabetes	Dementia	Diverticulitis of intestine (Div)
Epilepsy	Glaucoma	Hypothyroidism
Heart Failure (HF)	Inflammatory Bowel Disease (IBD)	Learning Disabilities (LD)
Multiple Sclerosis (MS)	Motor Neurone Disease (MND)	Osteoporosis (Osteo)
Osteoarthritis	Peripheral Arterial Disease (PAD)	Parkinson's disease
Psychoactive substance misuse (Substance)	Prostate disorders	Psoriasis
Rheumatoid Arthritis (RA)	Stroke or TIA	Severe Mental Illness (SMI)

to work with primary care data because it is not standardized between countries and the disease coding system used in the UK cannot be directly mapped onto the ICD-10 coding system. We are the first to apply this methodology to find trajectories of chronic disease accumulation using this type of data.

2. Data

To perform the analysis we used the CPRD (*Clinical Practice Research Datalink*) dataset [3] that contains EHRs from general practices across the UK. Our study included 1.1 million English people aged 45 and over, followed up from 2001 to 2010. Since there is no clear definition of the chronic conditions that define multimorbid patients we combined 3 different approaches to decide which were the diseases of interest: a systematic literature review of the diseases included in multimorbidity studies, the NHS Quality and Outcomes Framework disease list and expert advice from primary-care clinicians and clinical epidemiologists. Our final list is shown in Table 1. The CALIBER platform [4] provided the data extracts and phenotyped the disease codes.

3. Methods

The method used is based on Jensen et al [2] and we adapted some aspects of the implementation to the specific circumstances of this analysis.

First, we computed the Relative Risk (RR) between all the pairs of diseases. The Relative Risk between disease A and B measures if A is a risk factor for B (i.e. a temporal correlation measure): $RR_{A\to B} = \frac{P(B_{t=T}=1|A_{t=0}=1)}{P(B_{t=T}=1|A_{t=0}=0)}$, where A_t and B_t are random variables of being diagnosed or not with diseases A and B at time t. We estimated 5-year RRs by matching each patient with $A_{t=0} = 1$ (exposed group) to N comparison groups where $A_{t=0} = 0$. The comparisons groups were matched at baseline by sex, age and deprivation index. We computed the $P-$values as the proportion of comparison groups where the count of cases with $B_{t=5} = 1$ is larger than in the exposed group. We selected all pairs $A \to B$ with $RR_{A\to B} > 1$ and P-value < 0.05. In case of selecting both $A \to B$ and $B \to A$ we checked which direction was more significant. To evaluate the directionality we performed a Binomial test with the number of cases $N_{A\to B}$ and $N_{B\to A}$ in each direction. We kept at most one direction if it had P-value < 0.05.

Once all pairs of relevant directions were obtained we went through the dataset to find patient trajectories formed only by these pairs. If a patient is diagnosed with $A \rightarrow B \rightarrow C \rightarrow D$ we considered that trajectory for our analysis if all the pairs $A \rightarrow B$, $B \rightarrow C$ and $C \rightarrow D$ are relevant. We only selected trajectories with at least 4 diagnosed conditions and with at least 5 patients going through that trajectory. Afterwards, we merged all these trajectories to form a network where each node represents a disease. Finally, we clustered the resulting network using the Markov Cluster Algorithm [5].

4. Results

We used RRs to measure the temporal correlation between the different diseases. To compute the RRs we used 1000 comparison groups for each exposed patient. In Figure 1 we show the significant 5-year RR between all pairs of diseases. From the 870 pairs of disease we find that 648 have RR>1 and P-value <0.05. There are 281 pairs of diseases (A,B) where $RR_{A\rightarrow B} > 1$ and $RR_{B\rightarrow A} > 1$ and the P-values are under 0.05. Therefore, we checked for directionality to select the most significant directions of the two. After removing the less significant directed pairs we obtained 367 relevant directed pairs of diseases.

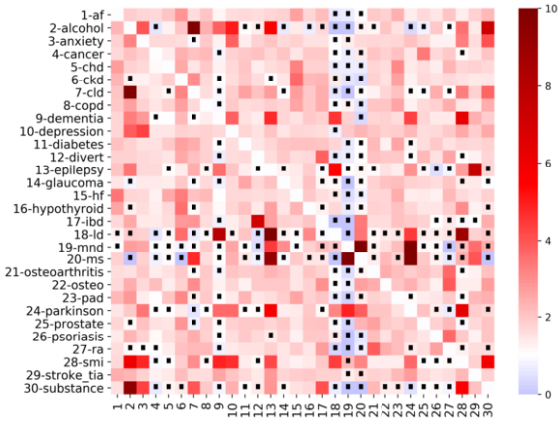

Figure 1. Relative Risks heatmap. The x-axis represents the baseline disease and the y-axis the follow-up disease. The x-axis and y-axis follow the same numerical index. The black dots indicate that RR <1 or P-value<0.05 so the relative risk is not significant. There are values with RR>10 but for visualization purposes we have limited the colorbar range.

We searched disease trajectories from patients in the CPRD dataset formed exclusively by these significant pairs of diseases and with at least 4 diagnoses in the trajectory. Then, we merged all trajectories that have at least 5 patients going through them and we obtained a network of diseases where the diseases are the nodes and the edges are the transitions in the trajectories. We applied the Markov Clustering Algorithm [5] to the network and we obtain two clusters. The main cluster is shown in Figure 2 and contains 16 diseases. To visualize the second cluster in Figure 3 we added 3 disease long trajectories. There are some diseases that do not appear in any of the two clusters since they do not appear in 3 long patient trajectories formed by *RR* significant pairs. These

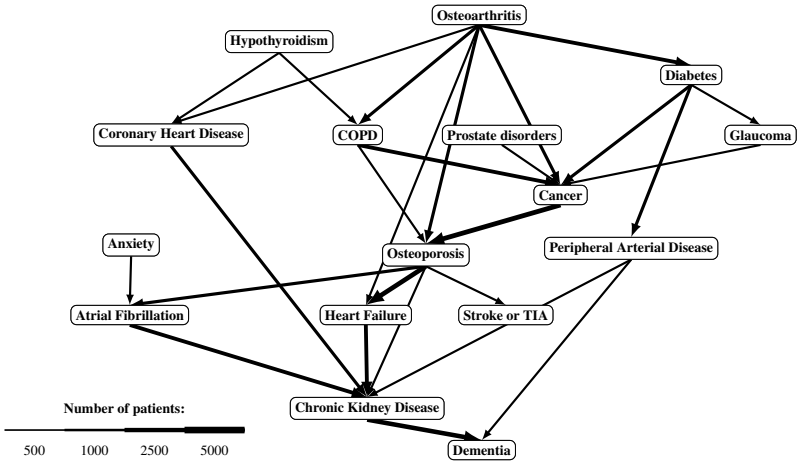

Figure 2. Main cluster of disease trajectories. The edge width represents the number of patients going through that diagnose sequence

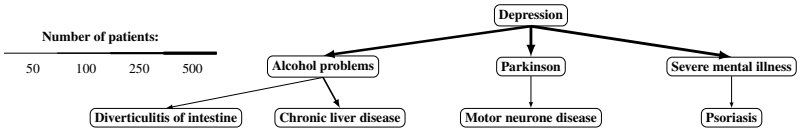

Figure 3. Second cluster of disease trajectories or Depression cluster. The edge width represents the number of patients going through that diagnose sequence

two clusters describe the most common trajectories of chronic disease accumulation. The main cluster network was build using trajectories of 4 conditions. Consequently, there are disease trajectories that start in a node with incoming edges.

5. Discussion

In this analysis we used EHRs from primary care to investigate the sequence in which patients get multiple chronic conditions. When studying multimorbidity it is not clear which are the chronic conditions that need to be taken into account. Thus, we decided to focus on a relatively small but prevalent number of diseases to understand what are the most common trajectories that multimorbid patients follow. We started by computing the RRs between diseases to evaluate the temporal correlation between any pair of conditions. Previous studies have focused on studying RRs of specific pairs of diseases. However, we were able to directly estimate 648 RRs thanks to the large dataset used. Most RRs related to learning disabilities, multiple sclerosis and motor neurone disease were found not significant due to the small amount of diagnosed cases. It is not possible to compare directly these RRs to previous studies because the time frame and disease definition can vary. Nevertheless, we found similar estimates for multiple pairs like CKD and Dementia [6] or PAD and CHD [7].

The main cluster shown in Figure 2 condenses the most common trajectories to multimorbidity into a single network. We can see how diseases like cancer, COPD, dia-

betes and osteoporosis are central in the progression. We also can observe clinically well known trajectories like *Diabetes* → *PAD* → *Stroke* [7]. The second cluster in Figure 3 shows reasonably expected alcohol problems related trajectories with depression as the initiating disease. However, there are many other trajectories that the two clusters do not cover. A dataset with a longer follow up period would help improve the results since we would be able to observe more complete trajectories. Previous work have applied this methodology to Secondary Care data and ICD-10 codes [2]. However, secondary care data does not capture most of the long lasting chronic conditions that multimorbid patients suffer from.

We know that sex and deprivation index are key factors to understand multimorbid patterns [8]. Consequently, in the future we should focus on finding specific trajectory networks for different population subgroups.

6. Conclusion

We used a data driven method to obtain networks that describe the main trajectories that patients follow to become multimorbid. The application of this method to a primary care dataset like CPRD confirms some of the already know temporal correlations between diseases and detects longer less known disease sequences. However, this method does not capture all the complex patterns that multimorbid patients follow. A main limitation of the model is that it only uses a directionality between the pairs of diseases. In the future we should focus on detecting specific trajectory patterns for different populations subgroups.

This work was partially supported by the National Institute for Health Research Collaboration for Leadership in Applied Health Research and Care North Thames at Bart's Health NHS Trust. The views expressed in this article are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care. The NIHR UCLH Biomedical Research Centre and EPSRC grant M020533/1 also supported this work.

References

[1] MacMahon, S. Multimorbidity: A priority for global health research. *The Academy of Medical Sciences: London, UK* (2018).

[2] Jensen, AB et al. Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients. *Nature communications* **5** (2014), 4022.

[3] Herrett, E et al. Data resource profile: clinical practice research datalink (CPRD).*International journal of epidemiology* **44** (2015), 827-836.

[4] Denaxas SC et al. Data Resource Profile: Cardiovascular disease research using linked bespoke studies and electronic health records (CALIBER). *Int J Epidemiol* **41** (2012), 1625-1638.

[5] Van Dongen, SM. Graph clustering by flow simulation. *Diss* (2000).

[6] Helmer, C et al. Chronic kidney disease, cognitive decline, and incident dementia: the 3C Study. *Neurology* **77** (2011), 2043-2051.

[7] Wattanakit, K et al. Risk factors for peripheral arterial disease incidence in persons with diabetes: the Atherosclerosis Risk in Communities (ARIC) Study. *Atherosclerosis* **180** (2005), 389-397.

[8] Chan, MS. Socio-economic inequalities in life expectancy of older adults with and without multimorbidity: a record linkage study of 1.1 million people in England. *International journal of epidemiology* **48** (2019), 1340-1351.