Early Cluster Prediction for Patients with Multi-Morbidities

NEWCASTLE UNIVERSITY

School of Computing

Arushi Nautiyal

Under the Supervision of: ***Paolo Missier***

01. Introduction

- Multiple long-term conditions, known as multimorbidity (MLTC-M), refer to the presence of two or more chronic diseases in an individual, significantly impacting their quality of life and posing challenges to the healthcare system.[1]
- **In England, 6.75 million people live with more than one long-term disease**, highlighting the scale of the issue [3].
- Research areas focus on understanding MLTCs by identifying disease patterns, predicting disease clusters and studying disease trajectory similarities [2]. Better understanding of these patterns can lead to improved clinical decision-making and patient management.

- This **project focuses on early disease cluster prediction, where clusters are formed based on observed co-occurrence in patients' medical histories and associating patients with those clusters.**
- Patients may transition between clusters, making them stable or unstable based on their attraction to different clusters during their medical trajectory.
- The patient trajectory comprises age at diagnosis, health condition, gender, and demographic data.

02. Methodology Used

- **Data**: This project uses primary care data for approximately 143,000 individuals in the UK, obtained through the UK Biobank [1].
- **Clustering Technique:** Latent Dirichlet Allocation (LDA) is a topic modelling method whose aim is to find topics a document belongs to, based on the words in it and then probability is calculated for words' association to different topics(cluster here).
- **Patient Trajectories:** Using Trend analysis and term relative weight of patients' diseases to find if a patient prominently belongs to a particular cluster.
- **Early Cluster Prediction:** Prediction is done in 2 ways:

1. Prediction at Stage 0(1st stage) is solely done on the basis of higher relative term weight(TRW), that is, if a patient's disease X has a higher TRW value for cluster Y as compared to that for other clusters, the patient is assigned to cluster Y.
2. For further stages, these TRW values for each disease are used to form a vector with magnitude R & angle θ which decides if a patient falls into a specific cluster or not.

Related literature

This project extends the work carried out by Kremer et. al.[1] in their project for tracking MLTCs, using the concept of stability to characterise clusters rather than the patients themselves. adding novelty to this research domain.

03. Analysis

Out of 143,000 patients, peoples have an average of **7 MTLCs**, and the aim of this project is to find out how early a dominant cluster can be predicted along with precision and if some diseases can co-occur more frequently in one population than the other. Below is a example diagram of how this is planned and calculated:

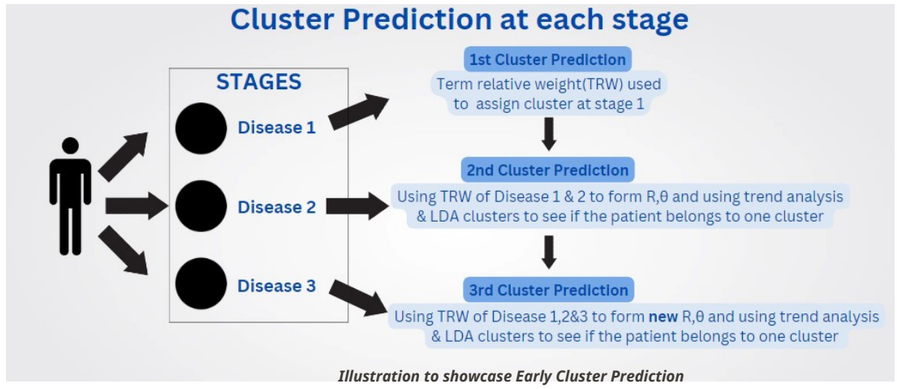
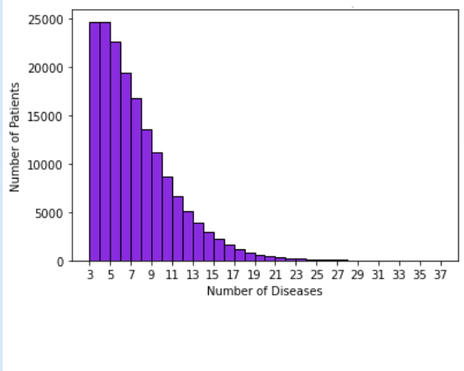


Illustration to showcase Early Cluster Prediction

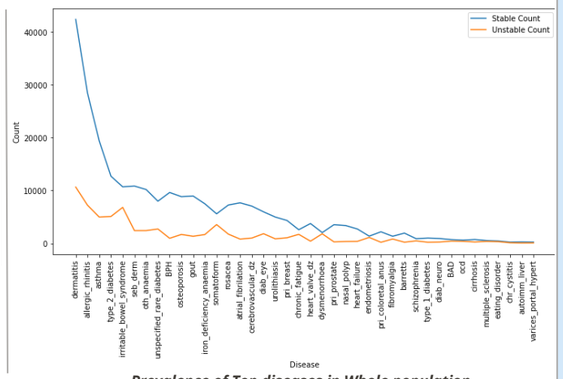
THINGS TO NOTE:

- *Patients who belonged to a cluster in the end are labelled Stable, while those associated with more than 1 cluster are labelled Unstable. This analysis has considered both stable and unstable population in order to make an accurate prediction in early stages.*
- *For a disease to have a high relative term weight(TRW) in one cluster/topic means it strongly associates within a particular topic as compared to the other remaining topics, which in turn means that diseases with higher TRW for a cluster may be a higher contributor in making a patient more stable.*

Below are 2 graphs: the first one shows distribution of MTLC count across population, and the second one is plotted for the top 10 diseases based on their relative term weight for 4 clusters (where cluster count=4) and the disease prevalence in stable and unstable population.



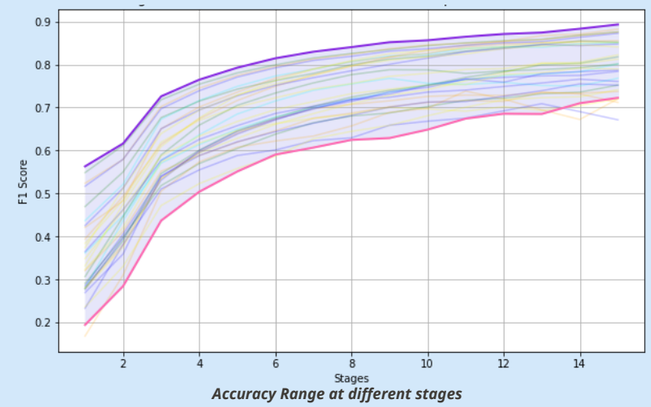
Distribution of number of MTLCs per patient



Prevalence of Top diseases in Whole population.

04. Results/Findings

Considering the previous research paper as the ground truth, final cluster prediction is taken from there as a true value. This true value is compared against the predicted value at each stage in this project and then the F1-score(a measure of accuracy) is calculated. Below is a plot depicting the accuracy range for a given stage for the taken cluster counts.



Accuracy Range at different stages

The F1 score increases as the stage progresses. This is plotted for various cluster counts (4,5,6,7 & 8 clusters) and surprisingly a similar pattern is found in each case.

Please note: While there are patients with MLTC count as high as 37, the F1 score analysis considers first 15 stages as the aim here is to predict with precision as early as possible.

Conclusion: This project backs up the previous research and extends the pre-existing method and breaks it down into stages, adding precision per stage as a new dimension. However, this research can be further improvised by considering different clustering methods or considering path similarities. Also, only considering weighted scores do not account for multimorbidity by chance[5] and may suffer generalizability problems[4].

References

[1] R.Kremer et al.,"Tracking trajectories of multiple long-term conditions using dynamic patient-cluster associations".

[2] A. Bisquera et al.,"Identifying longitudinal clusters of multimorbidity in an urban setting: A population-based cross-sectional study".

[3] UK Department of Health,2012 data . Available at: <https://www.gov.uk/government/organisations/department-of-health-and-social-care>

[4] Shu Kay Ng et al., "Patterns of multimorbid health conditions: a systematic review of analytical methods and comparison analysis"

[5] John R et al., "Patterns and impact of co morbidity and multimorbidity among community-resident American Indian elders".