International Journal of Population Data Science

Journal Website: www.ijpds.org

Identification of falls subgroups through semantic similarity analysis

Almohaimeed, Muhannad^{1*}, Brass, Andy¹, and Ba-Dhfari, Thamer¹

¹The University of Manchester

Objectives

The information contained within medical data is often used to make new medical discoveries. However, the most common way to use such data has been to query the data to answer very specific questions. For example, does having diabetes cause some patients to experience falls? If researchers have good questions, then the data can provide good answers. But are there any other equally important questions that could be asked of the data that people haven't yet thought to ask?

We are exploring a new strategy that we have developed to look for unusual and interesting patterns about falls in the elderly at subgroups level to see the different risks associated with different groups. Some of these risks will be associated with questions that are already well-known, but some should point to new and important questions that have not yet been asked. This opens up a better opportunity to identify patients at risk of falls, helping guide policy so as to reduce falls.

Approach

We mapped patient records into a low dimensional space using the notions of semantic similarity (Resnik node-based) and machine learning (principal component analysis) to provide a good representation of the data. This representation was used for clustering and visualisation through the DBSCAN algorithm. To look for enrichment in the resultant clusters, we analysed each cluster separately and look at the sets of patients defined in these clusters. Then, classic data mining techniques were used in order to generate hypotheses. The associations found were then be tested using more traditional comorbidity measures such as relative risk (RR) and its confidence intervals.

Results

We demonstrated the methodology on 589,169 older adults from clinical practice research datalink (CPRD). We successfully identified

six distinct subgroups of falls from the elderly population who are identified with different risks. Some of the associations found are well defined in the literature; for example, depression and musculoskeletal conditions are significantly associated with falls. However, a number of associations are not reported in the clinical literature. Such hypotheses need further exploration by epidemiologists.

Conclusion

Future work will focus on incorporating temporal dimension which might provide useful insights into missed opportunities detection and risk modelling and understanding of a disease. Last, this methodology holds promises for the study of other complex diseases using any source of data which are described using terms from taxonomies or ontologies.

*Corresponding Author:

Email Address: muhannad.almohaimeed@postgrad.manchester.ac.uk (M. Almohaimeed)