Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220427

Tracking Temporal Clusters from Patient Networks

Judith LAMBERT [a,b,1], Anne-Louise LEUTENEGGER[e], Anne-Sophie JANNOT[a,c] and Anaïs BAUDOT[b,d]
[a] *INSERM, CRC, Team 22, UMR1138, Paris, France* ; [b] *INSERM, CNRS, Aix Marseille Univ, MMG, UMR1251, Marseille, France* ; [c] *Department of Statistics, Medical Informatic and Public Health, HEGP, AP-HP* ; [d] *Barcelona Supercomputing Center, Barcelona, Spain* ; [e] *INSERM, Université de Paris, NeuroDiderot, UMR1141, Paris, France*

Abstract. Creating homogeneous groups (clusters) of patients from medico-administrative databases provides a better understanding of health determinants. But in these databases, patients have truncated care pathways. We developed an approach based on patient networks to construct care trajectories from such truncated data. We tested this approach on antithrombotic treatments prescribed from 2008 to 2018 contained in the échantillon généraliste des bénéficiaires (EGB). We constructed a patient network for each patients' age (years from birth). We then applied the Markov clustering algorithm in each network. The care trajectories were finally constructed by matching clusters identified in two consecutive networks. We calculated the silhouette score to assess the performance of this network approach compared to three existing approaches. We identified 12 care trajectories that we were able to associate with pathologies. The best silhouette score was obtained for the network approach. Our approach allowed to highlight care trajectories taking into account the longitudinal, multidimensional and truncated nature of data from medico-administrative databases.

Keywords. Longitudinal clustering, Patient networks, Care trajectories

1. Introduction

Finding homogeneous groups (clusters) of patients in medico-administrative databases helps to better understand health determinants and improve patient prognosis. But the longitudinal and multidimensional data of each patient are available at different times of their life in these databases, thus forming truncated care pathways. Clustering such data is therefore challenging. There are three types of approaches for clustering longitudinal data[1]. But the number of clusters must be specified, missing values are not handled or must be imputed and only one longitudinal factor at a time can be considered.

We developed an approach to construct care trajectories based on patient networks. Patient networks have the advantage of handling heterogeneous and missing data, preserving patient privacy and not requiring to prespecify the number of clusters[2]. As a use case, we analyzed drug prescriptions contained in the échantillon généraliste des bénéficiares (EGB).

[1] Corresponding Author, Judith LAMBERT; E-mail: judith.lambert@crc.jussieu.fr

2. Material and Methods

2.1 Echantillon généraliste des bénéficiaires (EGB)

The EGB is a random sample from the French health insurance database recording healthcare reimbursements from approximately 660,000 individuals followed for 11 years. We extracted all prescriptions for antithrombotic agents from 2008 to 2018 among patients aged 60 to 70 resulting in a sample of 30,111 patients.

2.2 A patient clustering approach based on patient networks

A patient network is a graph defined with nodes representing patients and edges representing the similarity between patient nodes. We constructed a patient network for each patients' age by computing the similarity using the Cosine similarity[3]. We applied the Markov clustering algorithm[4] on each patient network to identify patient clusters. We then tracked these clusters based on the number of patients they have in common between each consecutive age.

2.3 Competitive models

We compared the above approach with three existing approaches. As these approaches imply to set a priori the number of clusters, we chose the number of trajectories identified with the network approach. Their clustering performance was compared by calculating the silhouette score.

3. Results

We identified with the network approach 12 care trajectories characterized by a specific predominant drug and composed of 100 patients at least. The best silhouette score was obtained with the network approach (0.44).

4. Discussion

Among the 12 care trajectories identified with the network approach, we were able to associate three of them with stroke, infarction and arrhythmia. The clustering performance was better with the network approach compared to the three existing approaches. The use of patient networks therefore allowed to take into account the longitudinal, multidimensional and truncated nature of data.

References

[1] Liao TW. Clustering of time series data—a survey. Pattern recognition, 2005, vol. 38, no 11, p. 1857-1874.

[2] Pai S, Bader GD. Patient similarity networks for precision medicine. Journal of molecular biology, 2018, vol. 430, no 18, p. 2924-2938. Singhal A. Modern information retrieval: A brief overview. IEEE Data Eng. Bull., 2001, vol. 24, no 4, p. 35-43.