Topological Data Analysis to identify subgroups of type-2 Diabetes Mellitus

Introduction

There are several factors that contribute to the development of Type-2 Diabetes Mellitus (T2DM) (1); some evidence suggests that it is not only influenced by a deficiency in the pancreatic functions, but from a more complex path-way of the disease (2,3).

Unsupervised machine learning algorithms have been applied to gain a better understanding of the T2DM (4,5,6). Topological Data Analysis (TDA) has been recently implemented in the analysis of health datasets (7,8,9,10) to get insight from large and complex sets of data by studying its shape. TDA permits to find clusters of data that other unsupervised methods cannot identify (11).

TDA has been applied to identify subtypes of T2DM patients in the American population (12). There is not much information regarding subtypes of this disease and the implementation of TDA. The aim of this research was to perform a TDA in T2DM English patients using linked datasets to identify clusters of the disease.

Methods

CALIBER database links patient information from all healthcare settings in England. Subjects with one year of T2DM diagnosis during the 1998-2010 period, and not missing data were selected. The final population included 6,851 subjects.

The analysis included variables regarding sociodemographic information, lifestyles, nutritional status, and mental, endocrine, and circulatory diseases. Categorical variables were classified as whether the event was present in the 6-previous year of the T2DM diagnosis; mean was estimated for numeric features in the same period. Numeric data was standardized, and categorical variables were transformed into dummies when needed.

TDA was applied to identify clusters of data. This algorithm (figure 1), creates a distance matrix to evaluate similarity, and applies filter functions to project the data. The filters can be any function that converts the data into a single number. Two parameters are also applied to determine the later connections of the network. Finally, data points are grouped using a clustering algorithm. A node represents a group of subjects, and the edges the connections (similarity). A multinomial logistic regression was performed to identify statistical differences among groups.

Results

The sample population comprised 54% of women. Approximately, 44% were not smokers, and 84% did not consume alcohol regularly. Likewise, 7% have had a myocardial infarction, 60% had hypertension and 13% mental diseases. The TDA showed three groups of T2DM (Figure 2). The first group (70%) have an equitable distribution of clinical and socio-behavioural characteristics; the second group (12%), the most deprived cluster and mainly composed by women, have subjects with several undesired clinical complications and lifestyles behaviours; the third group (18%), mainly composed by men, have few clinical complications and acceptable lifestyles. Multinomial logistic regression showed a statically significant difference among groups.

Conclusions

TDA is an useful algorithm to perform exploratory and unsupervised analysis. The results suggest the existence of subtypes of T2DM. Further analyses are needed to confirm different subtypes of T2DM, and the utility of TDA to identify clusters in data; this study sums up to such purposes.

Acknowledgments

AstraZeneca, CONACyT, FIDERH, SEP.

Figures

Figure 1 presents the TDA mapper algorithm process, an x-coordinate filter function is applied to a circular distance matrix, and five intervals are established.

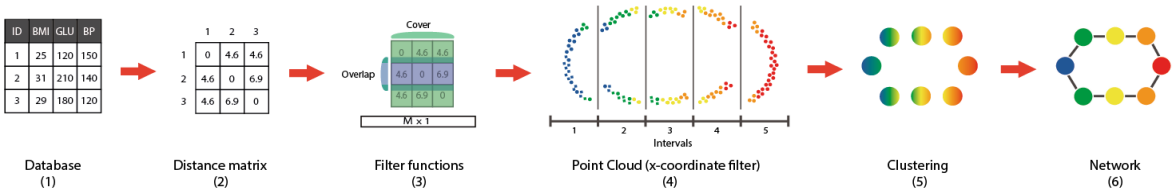
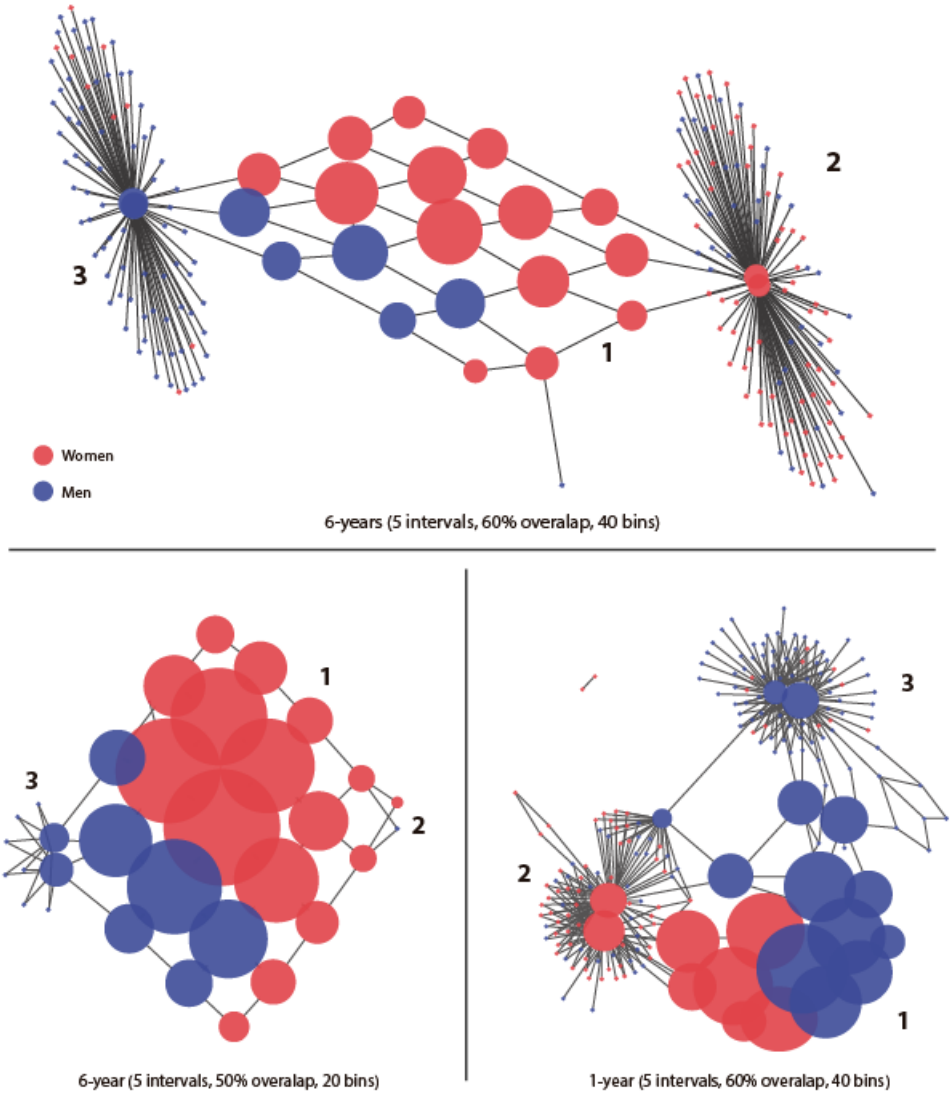


Figure 2 shows the TDA mapper output coloured by gender. Two additional graphs below are also plotted with different parameters; however, the shape remains similar.



References

1- Inzucchi, S.; Bergenstal, R.; Buse, J. Diamant, M.; Ferrannini, E.; Nauck, M.; Peters, A.; Tsapas, A.; Wender, R.; Matthew, D. (2012). Management of hyperglycaemia in type 2 diabetes: a patient-centered approach. Position statement of the American Diabetes Association (ADA) and the European Association for the Study of Diabetes (EASD). Diabetologia; 55(6):1577-96.

2- Færch, K.; Witte, D.; Tabák, A.; Perreault, L.; Herder, C.; Brunner, E.; Kivimäki, M.; Vistisen, D. (2013). Trajectories of cardiometabolic risk factors before diagnosis of three subtypes of type 2 diabetes: a post-hoc analysis of the longitudinal Whitehall II cohort study. The Lancet, Diabetes and Endocrinology; 1(1): 43-51.

3- Wild, S.; Byrne, C. (2013). Towards a personalised diagnosis of type 2 diabetes. The LANCET; 1(1):6-7.

4- Nithya, R.; Manikandan, P.; Ramyachitra, D. (2015). Analysis of clustering technique for the diabetes dataset using the training set parameter. International Journal of Advanced Research in Computer and Communication Engineering; 4(9): 166-169.

5- Kothainayaki, M.; Thangaraj, P. (2013). Clustering and classifying diabetic data sets using K-means algorithm. Journal of Applied Information Science; 1(1): 23-27.

6- Pala, T.; Camurcu, A. (2014). Evaluation of data mining classification and clustering techniques for diabetes. Malaysian Journal of Computing; 2(1):37-45.

7- Hinks, T.; Zhou, X., Staples, K.; Dimitrov, B.; Manta, A.; Petrossian, T.; Lum, P.; Smith, C.; Ward, J.; Howarth, P.; Walls, A.; Gadola, SD.; Djukanović, R. (2015). Multidimensional endotypes of asthma: topological data analysis of cross-sectional clinical, pathological, and immunological data. The LANCET; 385(Supp 1): S42.

8- Nielson J, Cooper S, Yue J, Sorani M, Inoue T, Yuh E, et al. (2017). Uncovering precision phenotype-biomarker associations in traumatic brain injury using topological data analysis. PLoS ONE 12(3): e0169490.

9- Nicolau, M.; Levine, A.; Carlsson, G. (2011). Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival. PNAS; 108(17): 7265-70.

10- Mehrdad, Y.; Smarr, L.; Knight, R. (2016). Using Topological Data Analysis to find discrimination between microbial states in human microbiome data. International Conference on Machine Learning (ICML).

11- Lum, P.; Singh, G.; Lehman, A.; IshkanoV, T.; Vejdemo-Johansson, M.; Alagappan, M.; Carlsson, J.; Carlsson, G. (2013). Extracting insights from the shape of complex data using topology. Sci Rep; 3: 1236.

12- Li, L.; Cheng, W. Y.; Glicksberg, B. S.; Gottesman, O.; Tamler, R.; Chen, R.; BottingeR, E. P. & Dudley, J. T. (2015). Identification of type 2 diabetes subgroups through topological analysis of patient similarity. Sci Transl Med; 7: 311ra174.