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

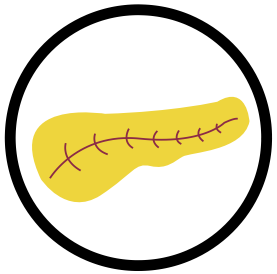
Juan Vallarta

MSc Health Data Science

David Prieto-Merino

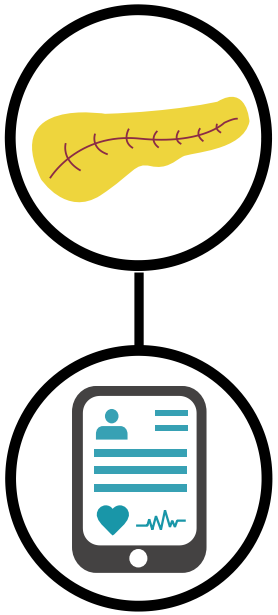
International Chair of Statistical Analysis and Big Data

Introduction



Some evidence suggests that the pathogenesis of the Type-2 Diabetes Mellitus (T2DM) it is not only influenced by a deficiency in the pancreatic functions, but from a more complex pathway of the disease.

Introduction



Some evidence suggests that the pathogenesis of the Type-2 Diabetes Mellitus (T2DM) it is not only influenced by a deficiency in the pancreatic functions, but from a more complex pathway of the disease.

The use of electronic medical records and the implementation of new analytical techniques, such as machine learning algorithms, can provide a better understanding of diseases.

Introduction

Some evidence suggests that the pathogenesis of the Type-2 Diabetes Mellitus (T2DM) it is not only influenced by a deficiency in the pancreatic functions, but from a more complex pathway of the disease.

The use of electronic medical records and the implementation of new analytical techniques, such as machine learning algorithms, can provide a better understanding of diseases.

Topological Data Analysis (TDA) is an unsupervised algorithm which main characteristic is to study the shape of data. This technique has been previously used to identify subtypes of T2DM in the American population. However, there is not much information regarding subtypes of this disease and the implementation of TDA.

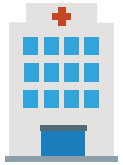
Objective

To perform a TDA using Electronic Medical Records (CALIBER dataset) to identify unique clusters of T2DM patients.

Methods

Database and Study Population

CPRD (primary care)
Life-styles
Nutritional Status
Lab tests
Diagnostics
Prescriptions
Procedures



HES (secondary care)
Sociodemographic
Administrative
Lab tests
Diagnostics
Prescriptions
Procedures

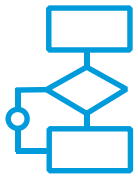
linkage

CALIBER
England
1998-2010

filtered

algorithm needs
server limitations

101,514 patients
with T2DM



6,851 patients
with T2DM

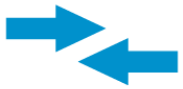
ONS (Dep. statistics)
Mortality
Deprivation

MINAP
Cardiovascular
diseases

Data Preprocessing

Numeric features

ID	Year1	Year2	Year3	Year4	Year5	Year6	BMI μ	BMI
1	25	26	28	27	30	32	28	1.2
2	32	30	31	32	30	32	31	1.8
3	24	22	23	25	25	25	24	0.6



HDL μ	HDL
100	0.2
180	2.1
165	1.4

Mean was obtained for numeric variables
(6 years previous to the T2DM diagnosis)

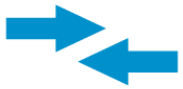
Numeric features
were standardised

Data Preprocessing

Numeric features

ID	Year1	Year2	Year3	Year4	Year5	Year6	BMI μ	BMI
1	25	26	28	27	30	32	28	1.2
2	32	30	31	32	30	32	31	1.8
3	24	22	23	25	25	25	24	0.6

Mean was obtained for numeric variables
(6 years previous to the T2DM diagnosis)



HDL μ	HDL
100	0.2
180	2.1
165	1.4

Numeric features were standardised

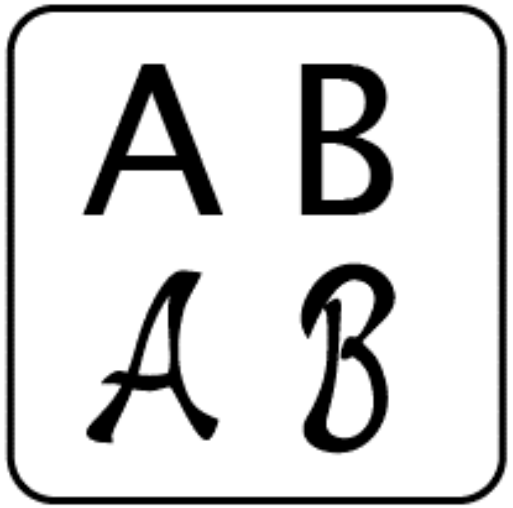
Categorical features

ID	Year1	Year2	Year3	Year4	Year5	Year6	MI
1	Yes	No	No	No	No	Yes	Yes
2	No	No	No	No	No	No	No
3	No	No	No	Yes	No	No	Yes

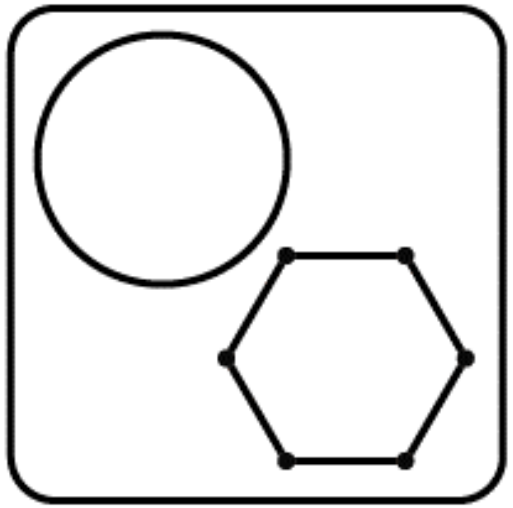
Categorical variables were transformed into dummies
(6 years previous to the T2DM diagnosis)

Topological Data Analysis

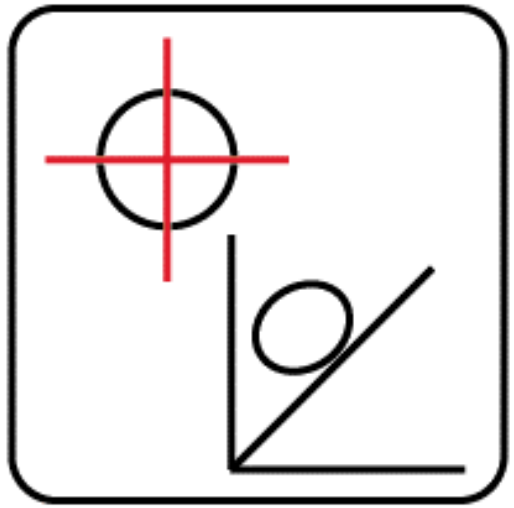
TDA Properties



Deformation
Invariance



Compressed
Representation



Coordinate
Freeness

Data Analysis (TDA Mapper)

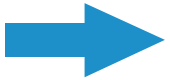
ID	IMC	GLU	HP
1	25	120	150
2	31	210	140
3	29	180	120

Database

Data Analysis (TDA Mapper)

ID	IMC	GLU	HP
1	25	120	150
2	31	210	140
3	29	180	120

Database



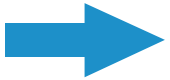
	1	2	3
1	0	4.6	4.6
2	4.6	0	6.9
3	4.6	6.9	0

Distance Matrix

Data Analysis (TDA Mapper)

ID	IMC	GLU	HP
1	25	120	150
2	31	210	140
3	29	180	120

Database



	1	2	3
1	0	4.6	4.6
2	4.6	0	6.9
3	4.6	6.9	0

Distance Matrix

$A=U\Sigma V^T$

$f(x)=\max d(x,y)$

f (x)

Filter functions
Definition of parameters
(Intervals, Overlap)

Data Analysis (TDA Mapper)

ID	IMC	GLU	HP
1	25	120	150
2	31	210	140
3	29	180	120

Database

	1	2	3
1	0	4.6	4.6
2	4.6	0	6.9
3	4.6	6.9	0

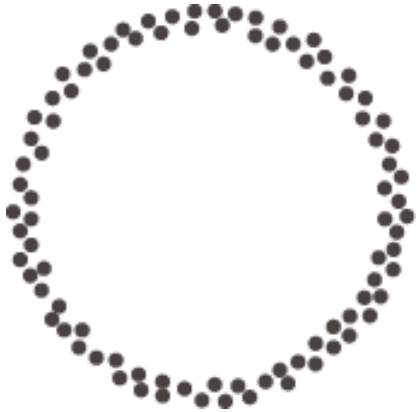
Distance Matrix

$A=U\Sigma V^T$

$f(x)=\max d(x,y)$

f (x)

Filter functions
Definition of parameters
(Intervals, Overlap)



Sphere projection

Data Analysis (TDA Mapper)

ID	IMC	GLU	HP
1	25	120	150
2	31	210	140
3	29	180	120

Database

	1	2	3
1	0	4.6	4.6
2	4.6	0	6.9
3	4.6	6.9	0

Distance Matrix

$A=U\Sigma V^T$

$f(x)=\max d(x,y)$

f (x)

Filter functions
Definition of parameters
(Intervals, Overlap)

Sphere projection

Data Analysis (TDA Mapper)

ID	IMC	GLU	HP
1	25	120	150
2	31	210	140
3	29	180	120

Database

	1	2	3
1	0	4.6	4.6
2	4.6	0	6.9
3	4.6	6.9	0

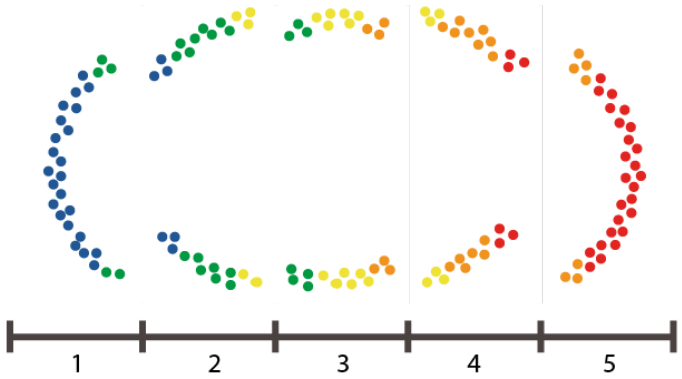
Distance Matrix

$A=U\Sigma V^T$

$f(x)=\max d(x,y)$

$f(x)$

Filter functions
Definition of parameters
(Intervals, Overlap)



Projection at 5 intervals
(x-coordinates filter)

Data Analysis (TDA Mapper)

ID	IMC	GLU	HP
1	25	120	150
2	31	210	140
3	29	180	120

Database

	1	2	3
1	0	4.6	4.6
2	4.6	0	6.9
3	4.6	6.9	0

Distance Matrix

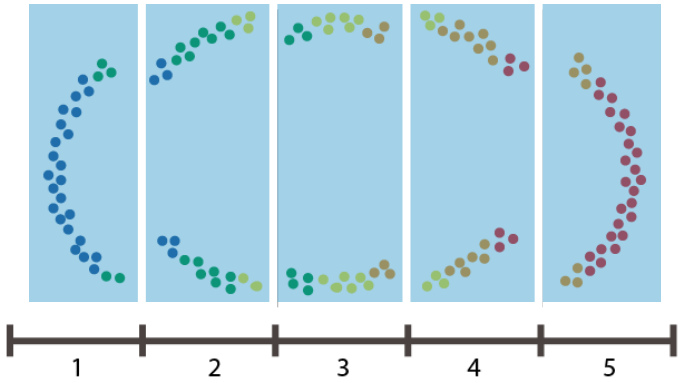
$A=U\Sigma V^T$

$f(x)=\max d(x,y)$

f (x)

Filter functions
Definition of parameters
(Intervals, Overlap)

Intervals



Projection at 5 intervals
(x-coordinates filter)

Data Analysis (TDA Mapper)

ID	IMC	GLU	HP
1	25	120	150
2	31	210	140
3	29	180	120

Database

	1	2	3
1	0	4.6	4.6
2	4.6	0	6.9
3	4.6	6.9	0

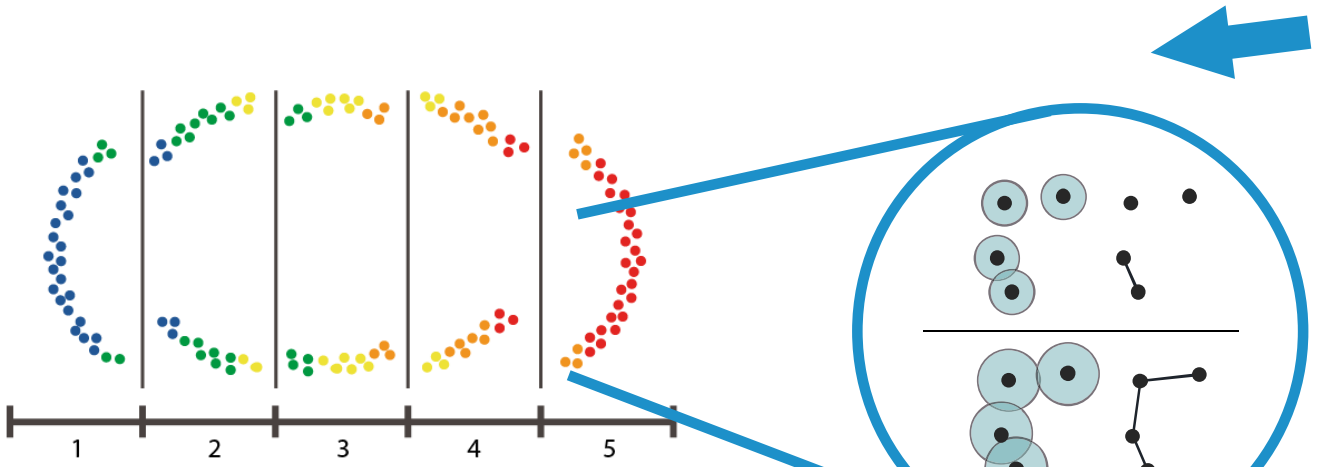
Distance Matrix

$A=U\Sigma V^T$

$f(x)=\max d(x,y)$

$f(x)$

Filter functions
Definition of parameters
(Intervals, Overlap)



Projection at 5 intervals
(x-coordinates filter)

Overlap

Data Analysis (TDA Mapper)

ID	IMC	GLU	HP
1	25	120	150
2	31	210	140
3	29	180	120

Database

	1	2	3
1	0	4.6	4.6
2	4.6	0	6.9
3	4.6	6.9	0

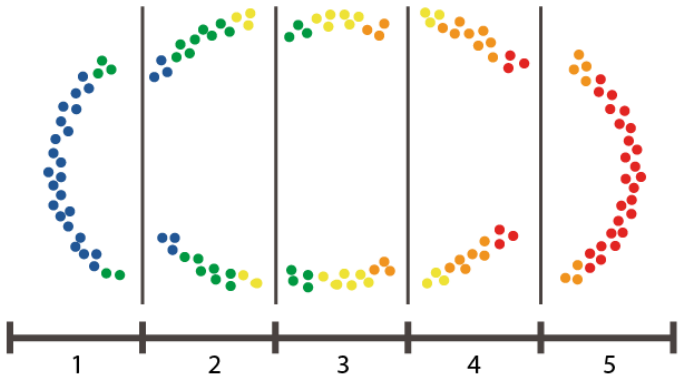
Distance Matrix

$A=U\Sigma V^T$

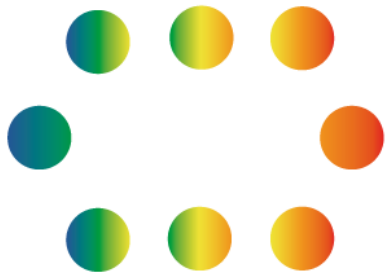
$f(x)=\max d(x,y)$

$f(x)$

Filter functions
Definition of parameters
(Intervals, Overlap)



Projection at 5 intervals
(x-coordinates filter)



Clustering

Data Analysis (TDA Mapper)

ID	IMC	GLU	HP
1	25	120	150
2	31	210	140
3	29	180	120

Database

	1	2	3
1	0	4.6	4.6
2	4.6	0	6.9
3	4.6	6.9	0

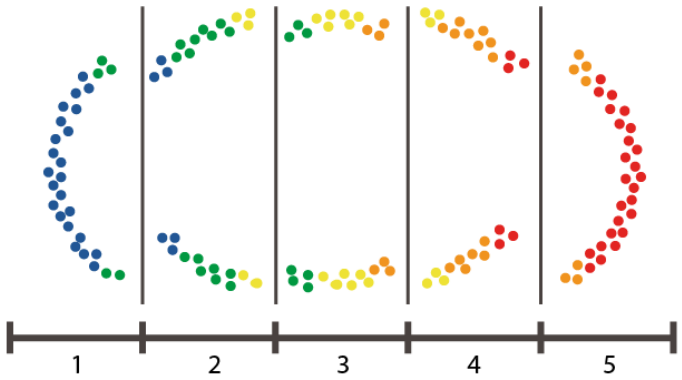
Distance Matrix

$A=U\Sigma V^T$

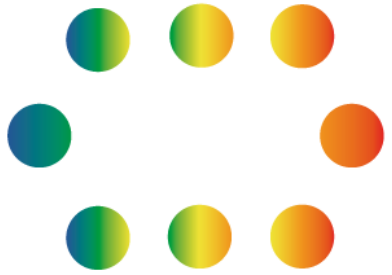
$f(x)=\max d(x,y)$

$f(x)$

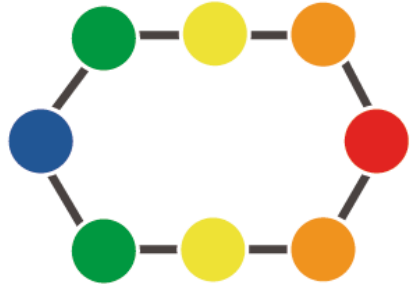
Filter functions
Definition of parameters
(Intervals, Overlap)



Projection at 5 intervals
(x-coordinates filter)



Clustering

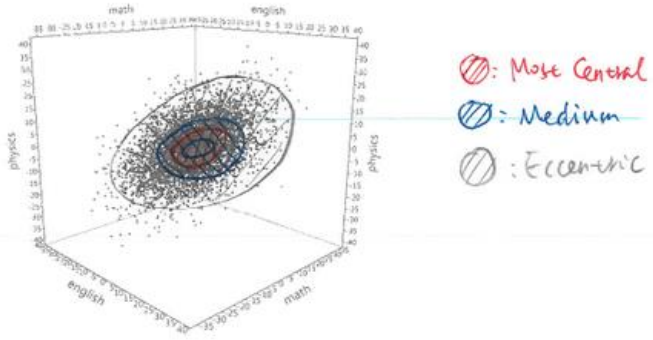


TDA

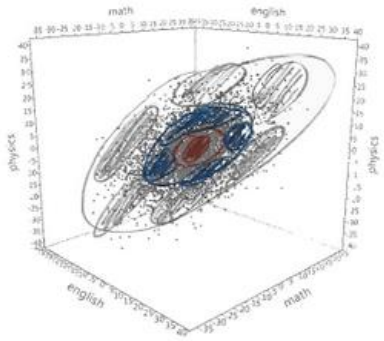
Filter Examples

Filter 1: L2-Centrality

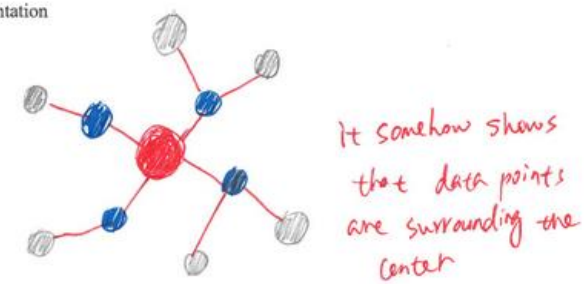
Step 1: Filtration



Step 2: Clustering

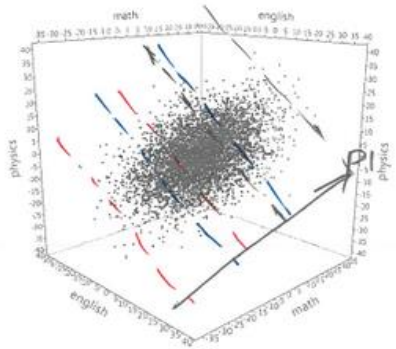


Step 3: Topological Representation

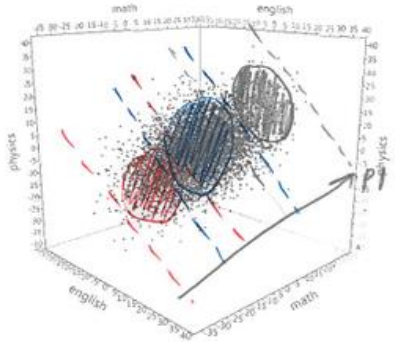


Filter 2: PCA-P1

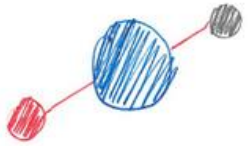
Step 1: Filtration



Step 2: Clustering



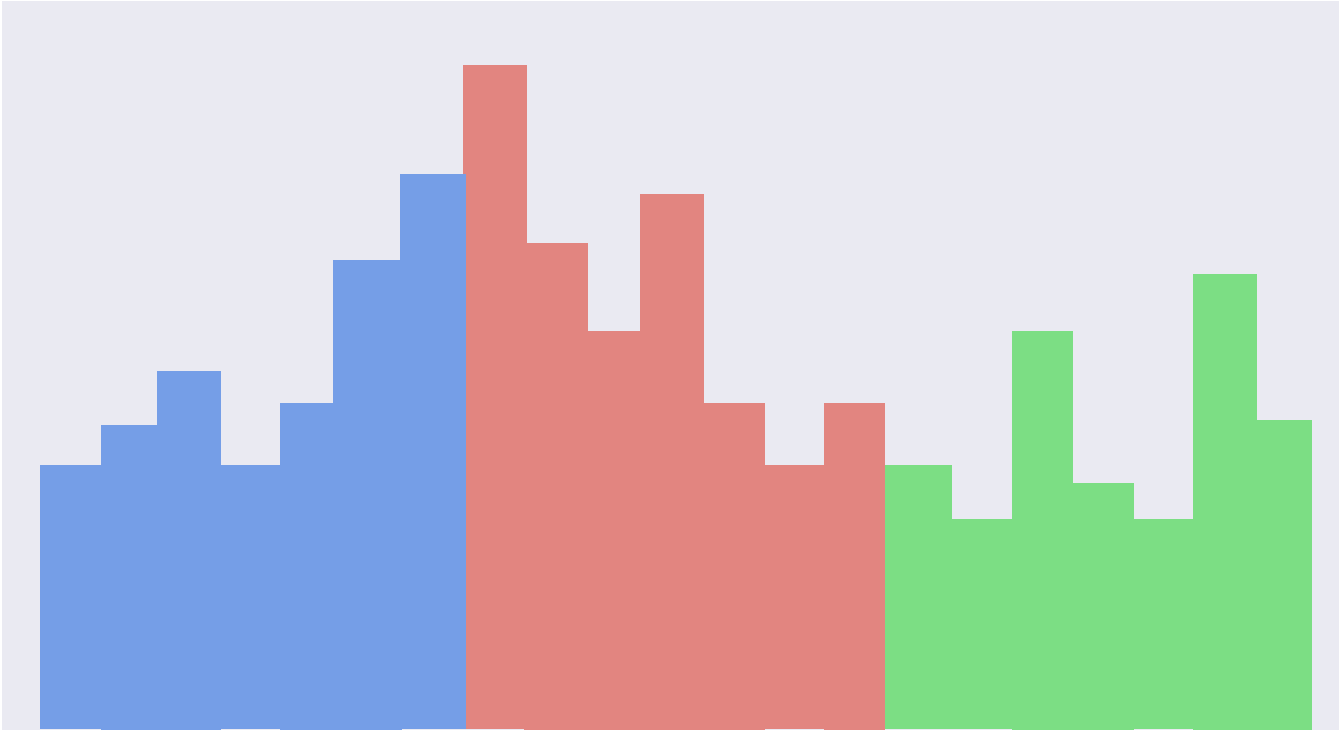
Step 3: Topological Representation



Jing E., 2015

Statistical Analysis

Multinomial Logistic Regression using a backward approach.



Results

Descriptive Results

54% women.

75% diet to lose weight.

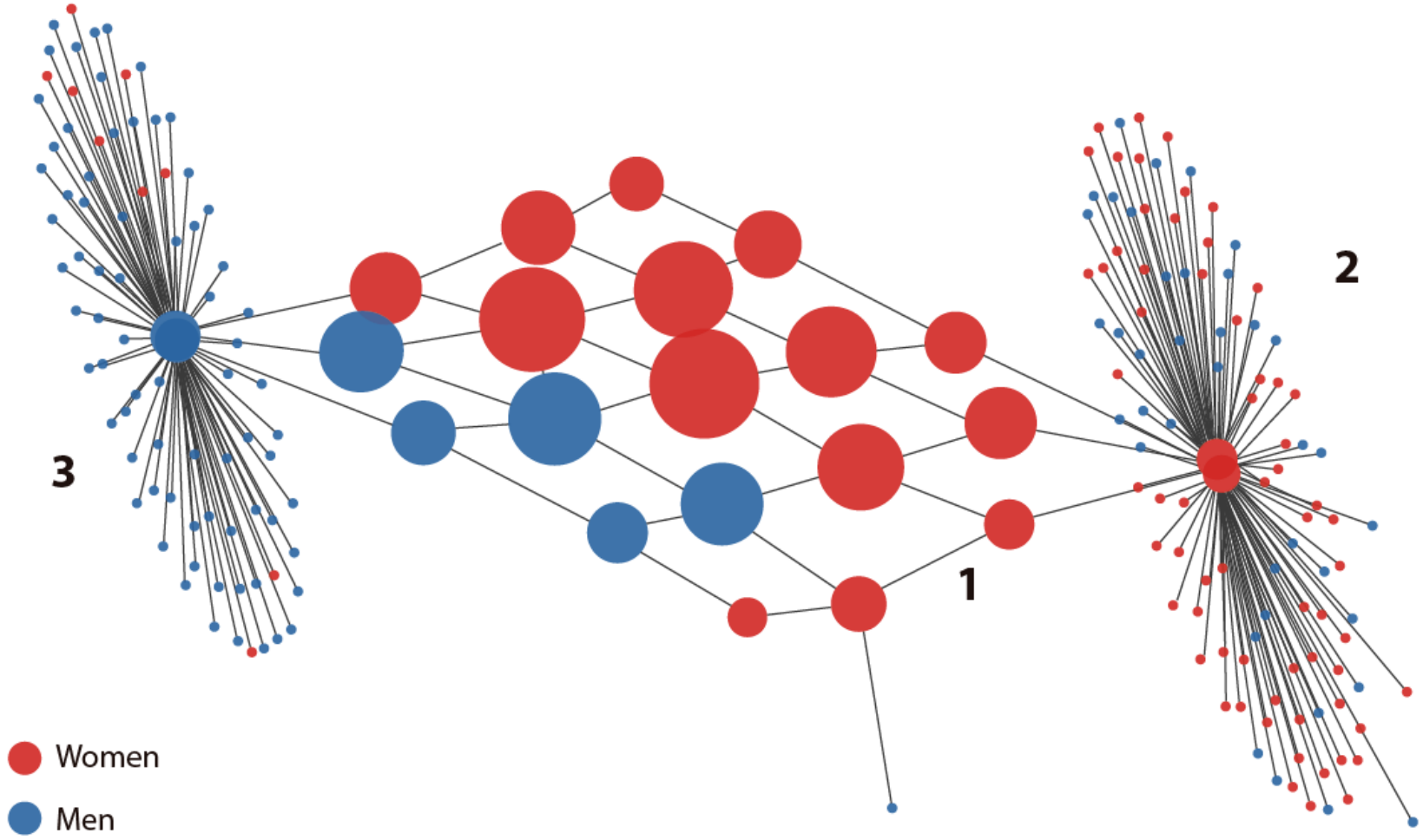
12% depression.

44% non-smokers and 30% ex-smokers.

84% non-alcoholic.

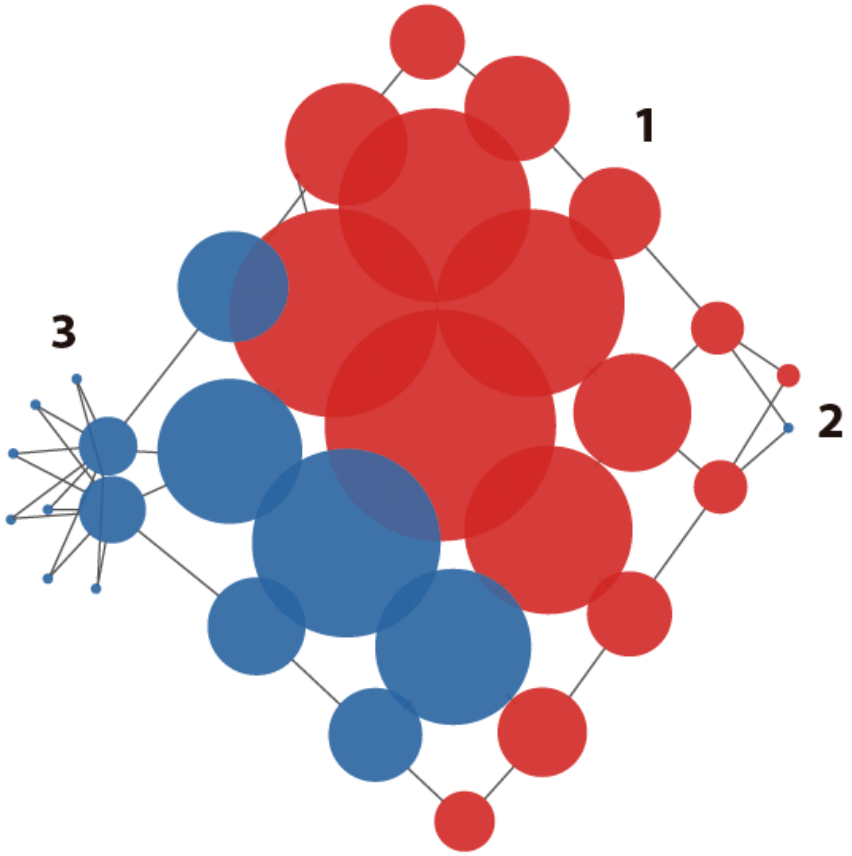
7% episode of heart attack.

TDA Output

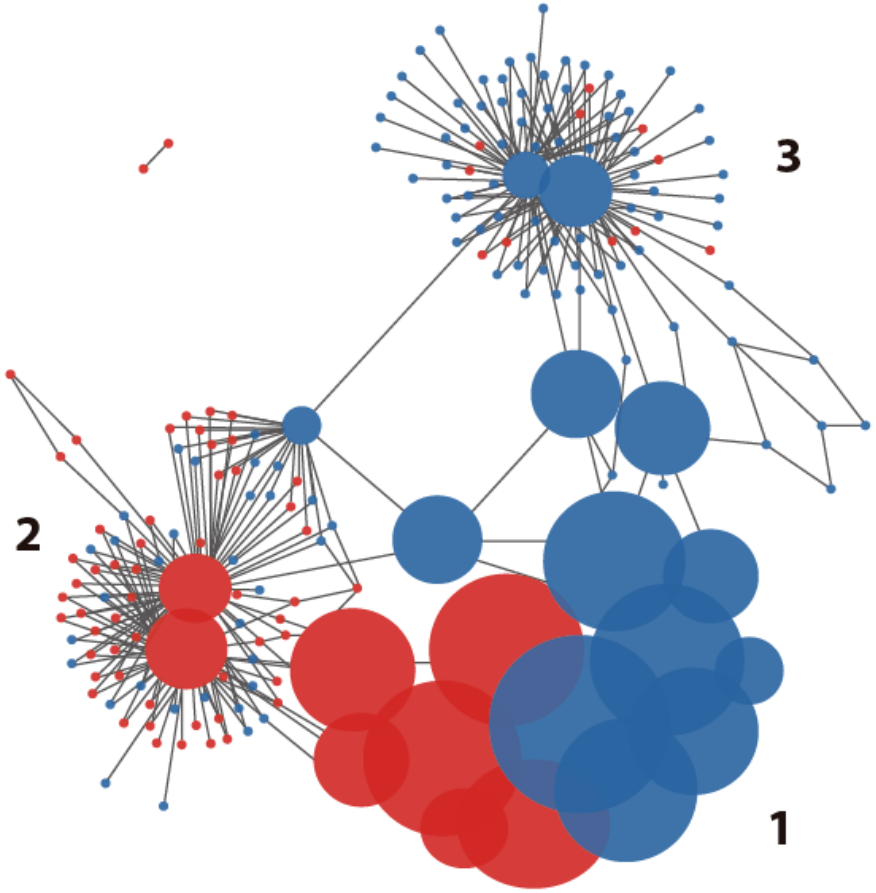


6-years (5 intervals, 60% overlap, 40 bins when clustering)

TDA alternative outputs

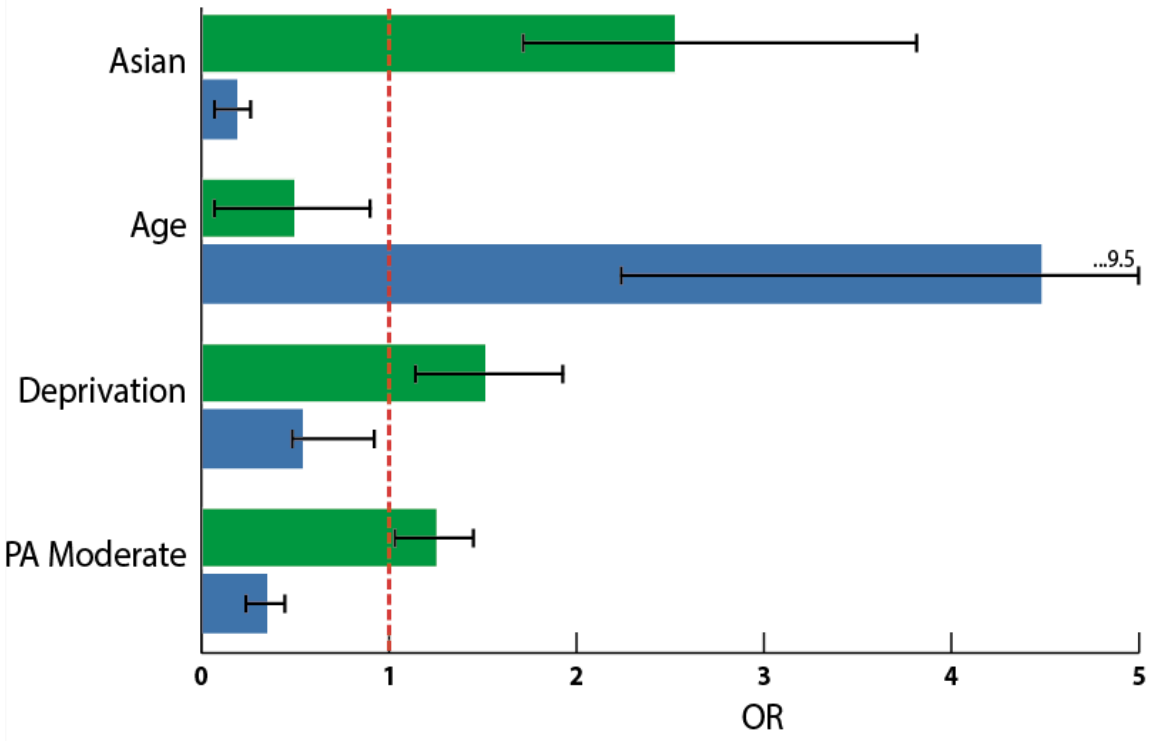


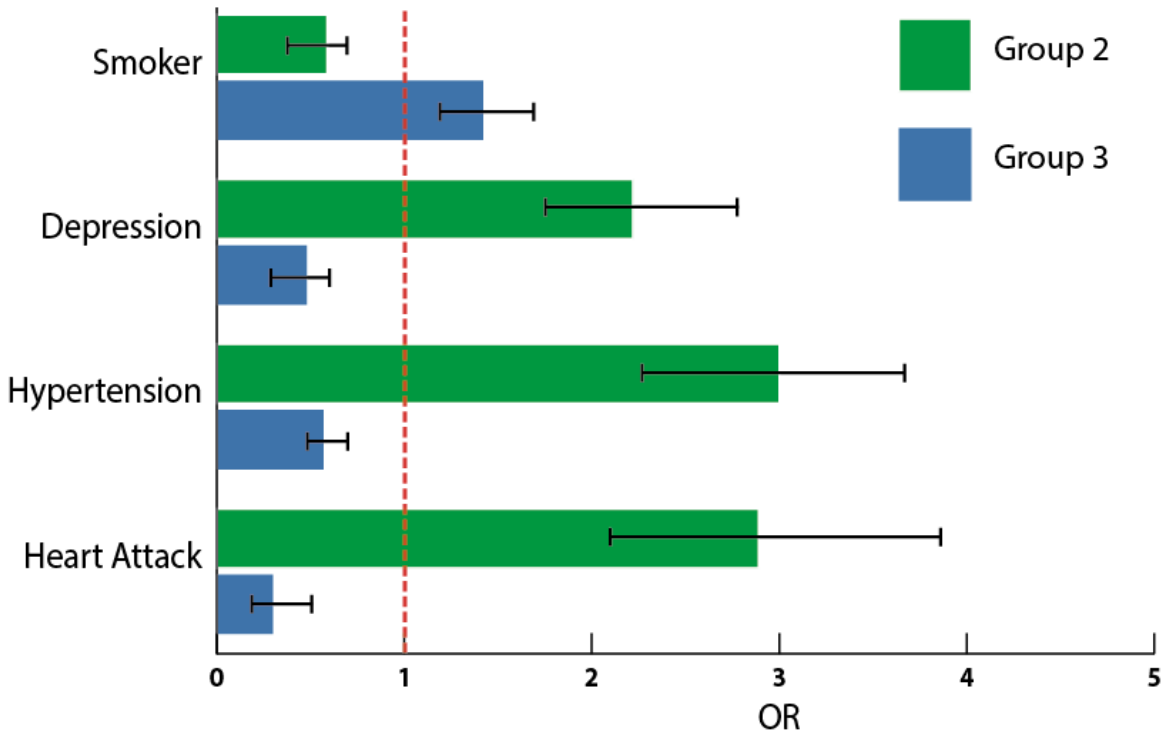
6-years (5 intervals, 50% overlap, 20 bins when clustering)



1-year (5 intervals, 60% overlap, 40 bins when clustering)

Multinomial Logistic Regression





Conclusions

Conclusions

- TDA is an useful algorithm to visualize and understand high dimensional datasets, and to find clusters in data.

- The results suggest the existence of subgroups of T2DM patients with unique clinical, sociodemographic, and behavioural characteristics. This can be useful to target different type of treatments.

Many thanks Sponsors!!

References

- Wild, S.; Byrne, C. (2013). Towards a personalised diagnosis of type 2 diabetes. The LANCET; 1(1):6-7.
- Kavakiotis, I.; Tsave, O.; Salifoglou, A.; Maglaveras, N.; Vlahavas, I.; Chouvarda, I. (2017). Machine Learning and Data Mining Methods in Diabetes Research. Computational and Structural Biotechnology Journal; 15:104–116.
- 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.
- Jing, E. (2015). Topological Data Analysis (TDA) - Visualizing High Dimensional Data. https://sites.google.com/site/icictamu/blog/topologicaldataanalysistda-visualizinghighdimensionaldata [Accessed: 18 July 2017].