CAMDA 2024: Synthetic Clinical Health Records Challenge

Unlocking Insights in Diabetes Pathologies Using Synthetic Data

UNAM

amphora HEALTH

Diabetes is a **global** concern with clinical and economic impacts

Type 2 Diabetes Mellitus (T2D) is a prevalent metabolic disorder characterized by hyperglycemia due to defects in insulin secretion or action.
T2D often leads to severe complications, including **cardiovascular diseases**, **nephropathy**, **retinopathy**, and **neuropathy**.

Image generated with DALL-E

537 million adults are living with diabetes

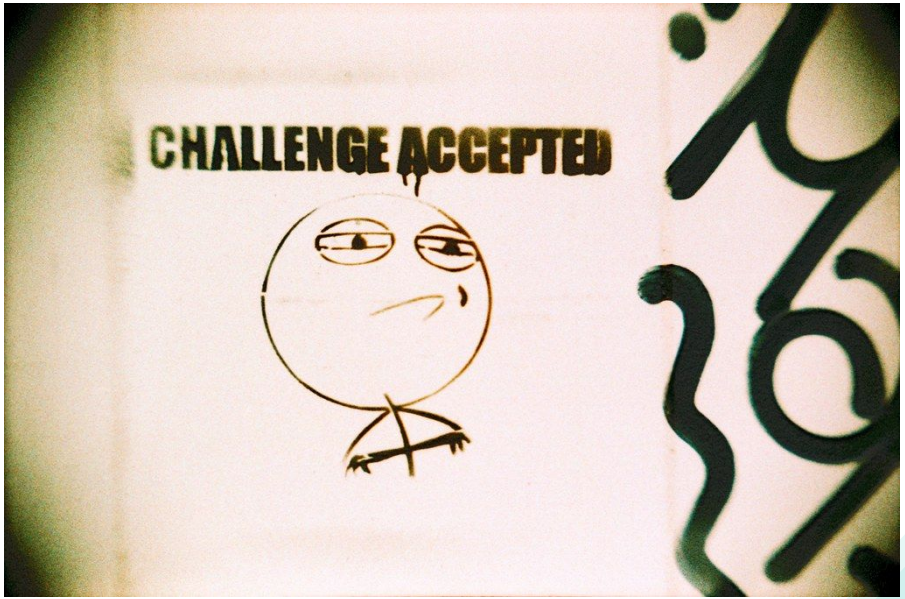
3 in 4 adults with diabetes live in low- and middle-income countries

6.7 million deaths due to diabetes in 2021

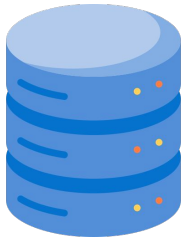
Statistics from https://diabetesatlas.org/

The CAMDA challenge aims to develop predictive models and original analyses using synthetic patient data.

- **Primary Objective:**
  - **Predictive Analysis:**
    - **Identify relationships between pathologies associated with diabetes.**
    - **Predict specific pathologies before they are diagnosed.**
  - **Disease Trajectory:**
    - **Predict the progression of diseases in patients.**
- **Open to Other Approaches:**
  - **The development of any other study approach is encouraged.**



Data Sources



Synthetic database with 999,936 patients generated from 984,414 real diabetes patients from Andalucía, Spain

DiabetIA

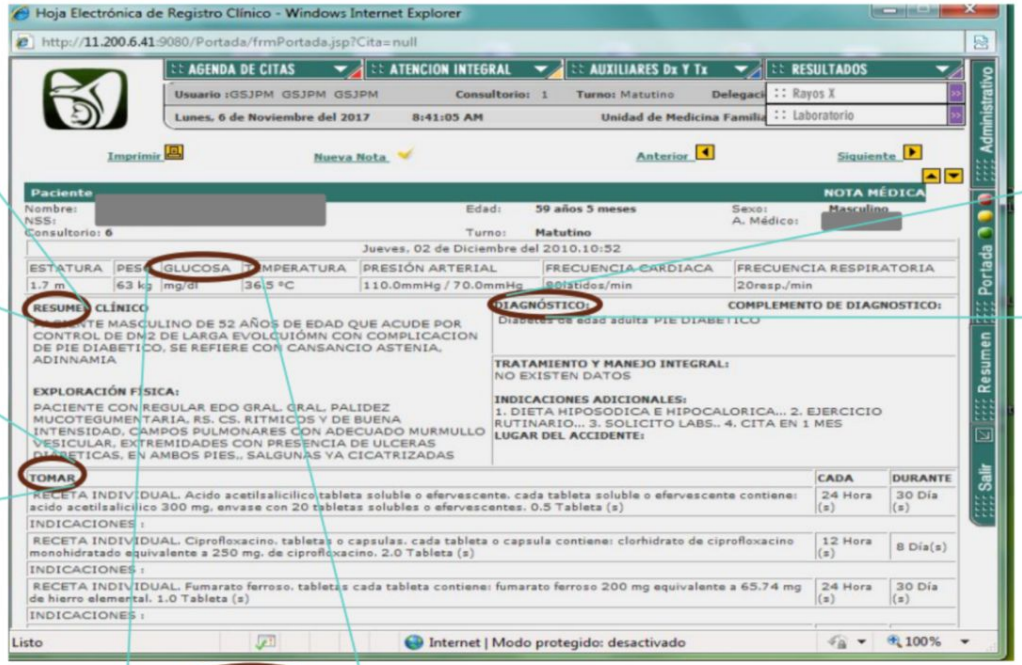
Research database with 136,000 patients from primary care units from Michoacan, Mexico

Cleaning processes

Electronic Medical Record at UMFs

Medical Note
Signs, Symptoms, sometimes lab results

Treatment
Aspirin, Ciprofloxacin, Iron(II) fumarate



Diagnosis
Diabetes
Foot ulcers

Tracking
Ht, Wt, Glucose, BP, HR, RR

Source data was delivered on JSON format

Unknown gender

multiple ages

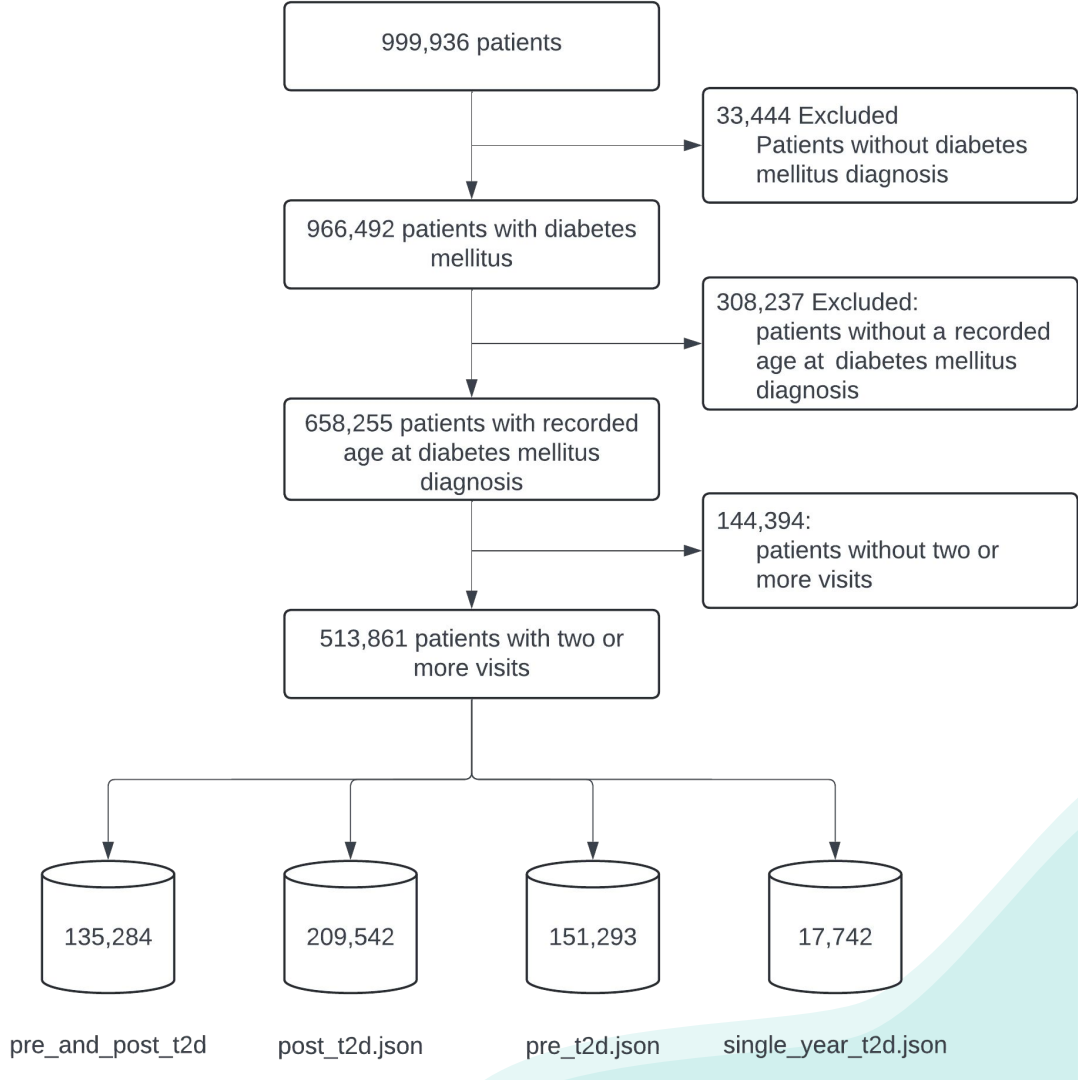
{"0": [[], ["9063", "1001", "9058"], ["506", "9071", "604"], ["9071", "1302"], ["9072", "9073"], ["9073"], ["9073", "9077", "507"], ["9077"], ["9078"], ["9078", "401", "403"], ["9081", "1101"], ["9083"], ["9083"], ["9083"]], "1": [["1111"], ["914", "9069", "9088", "9072"], ["9076", "401"], ["9076", "9082", "9084"], ["9085", "913", "9076", "909"], ["9085", "1001", "909"], ["507", "9085", "502"], ["1302", "9086", "911"], ["9086"], ["9087", "404", "901"], ["9087", "404", "910"], ["9090", "1401", "1201"], ["9090"], ["9090"]], "2": [["1111"], ["9052", "9051", "9047"], ["913", "9057", "9053"], ["507", "401", "1302"], ["9061", "604"], ["906", "9061", "604", "502"], ["9062", "604", "500"], ["402", "9062", "604", "500"], ["403", "9062", "604", "500"], ["604", "603", "703", "9063"], ["9063", "500", "602"], ["9063", "603", "905", "1102"], ["1102", "603"], ["501", "905"]], "3": [["1111"], ["507"], ["1102", "9055", "402"], ["9056"], ["9056", "502", "1102"], ["9056", "1102", "501"], ["9061", "1102", "403", "501"], ["9061", "403"], ["9063", "206", "1102", "703", "1101"], ["1102", "501", "9066"], ["9066", "1102", "401", "501"], ["9066", "401"], ["9070", "404"]], "4": [["1111"], ["913", "9058", "9062"], ["9064"], ["9069", "507"], ["9070", "507"], ["401", "1302"], ["9074"], ["9074"], ["9074"], ["9074"], ["9074", "914"], ["9076", "906", "914", "204"], ["501", "9081"]], "5": [["1111"], ["910", "913"], ["401", "9066"], ["9068"], ["9068", "503"], ["9069", "402"], ["9070"], ["9070"], ["1302", "9071"], ["9071"], ["9071"], ["9072", "914"], ["9077"]], "6": [["1111"], ["9079"], ["9080"], ["9081", "401"], ["9082"], ["9083"], ["9083", "507", "1001", "1302"], ["9083", "502"], ["9084", "906", "903", "604"], ["914"], ["9086"], ["9088"]], "7": [["2222"], ["913"], ["507", "9061"], ["1302", "9062"], ["910", "402", "9071"], ["9074", "401"], ["9074"], ["9074", "604", "404"], ["9075", "604"], ["9075", "500", "604"], ["501", "9076", "603"], ["604"]], "8": [["2222"], ["913"], ["9066"], ["1302", "9070"], ["9071", "914"], ["9071", "401", "914"], ["9071", "1401", "909"], ["402", "9077"], ["9078"], ["9079", "910"], ["910", "503", "601", "9079", "702"], ["910", "1402", "500"]], "9": [["1111"], ["503", "9061", "9060"], ["501", "9064", "1002", "1304", "502"], ["9064"], ["401", "1401", "9066"], ["9069"], ["9069", "913", "502"], ["507", "402", "9071"], ["9072", "914"], ["9072", "403"], ["9075", "1103"], ["1103"]], "10": [["1111"], ["913"], ["402", "9064"], ["9067"], ["9067", "1302"], ["9068"], ["1401", "9070"], ["9073", "401"], ["9075"], ["901", "9077", "1001"], []], "11": [["2222"], ["1302"], [], ["507", "9074"], ["9075"], ["9077", "910"], ["401", "9078", "403"], ["9078", "403"], ["9079", "914"], ["9079"], ["9079", "604", "500"]], "12": [["2222"], ["9073"], ["9080", "913", "507"], ["507", "9080"], ["1302", "9080"], [], ["9084", "401"], ["9084"], ["9086"], ["9090", "501"], ["9092"]], "13": [[], ["9063"], ["9073"], ["9073"], ["9073", "1401"], ["9077", "403"], ["9077", "403"], ["9077"], ["401"], ["9081"], ["9086", "9083"]], "14": [["2222"], ["9065"], ["9072", "913"], ["9072", "1001"], ["9074", "401", "1001"], ["9074", "1302"], ["9074"], ["9078", "9074"], ["914", "9079"], ["9079", "500", "402"], ["9085"]], "15": [["1111"], ["913"], ["9046", "404"], ["9048"], ["914", "1401"], ["9053", "507"], ["9054", "910"], ["9056", "1302", "1402"], ["9057", "1301"], ["9057", "401"], ["9058"]], "16": [["2222"], ["913"], ["1302", "402", "9060"], ["507", "9062"], ["9064", "401"], ["9064", "404"], ["9065"], ["9068", "910"], ["9068", "901"], ["1201", "9074", "901"], ["604"]], "17": [["1111"], ["507", "9050"], ["9053"], ["1302", "9057"], ["913", "401", "9057"], ["9060"], ["9063"], ["9063"], ["9065"], ["9066"], ["9070"]], "18": [["1111"], ["402"], ["401", "9072"], ["913", "507", "9072"], ["9078"], ["403", "9079"], ["9079"], ["9080"], ["9080", "702", "1402"], ["9080", "503", "203", "1301"], ["9080", "914", "1001"]],

Unknown age of onset

Unknown disease

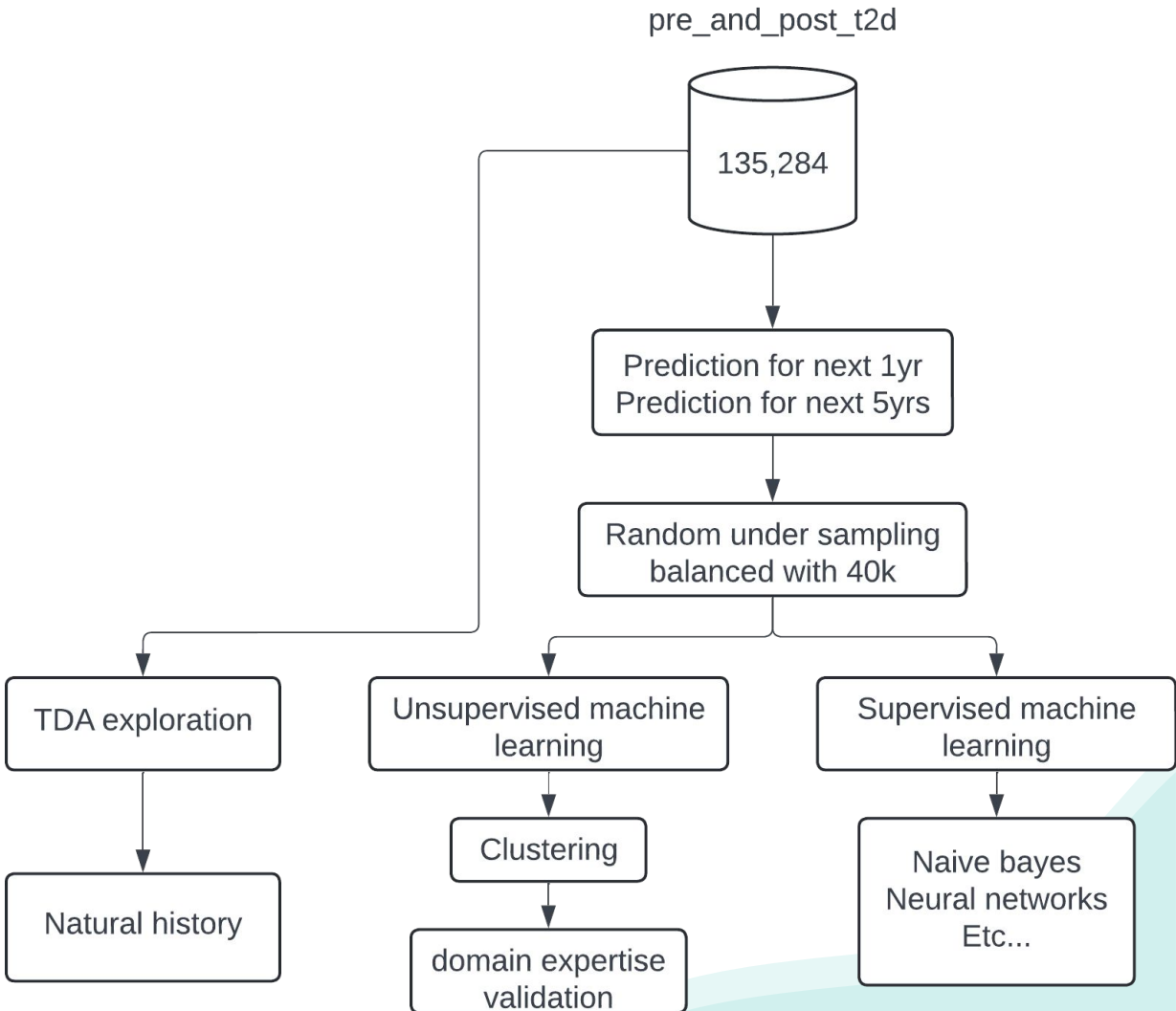
Preprocessing pipeline

Before the Hackathon we established diverse cleaning criteria to understand the data and to clean errors.



Workflow

Along the hackathon we established some work roles to distribute the possible tasks.

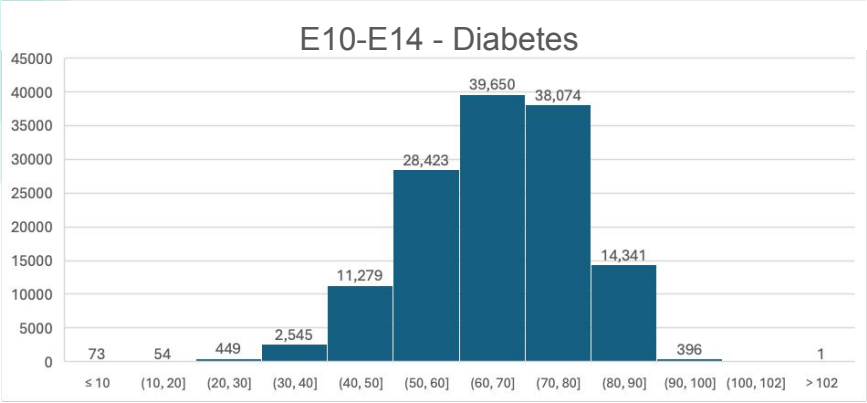


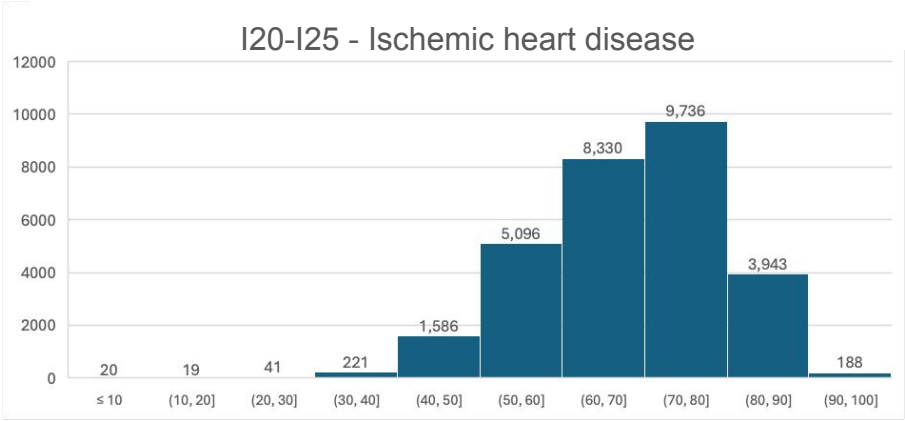
Patient journey

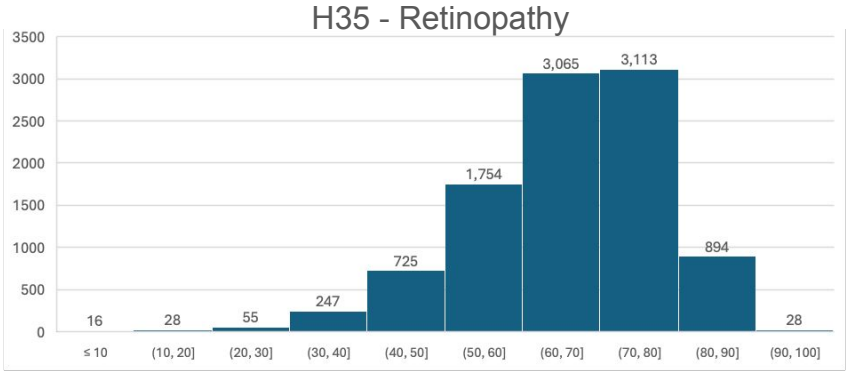
	CAMDA-All	CAMDA-DM	CAMDA-No DM*	DiabetIA
N	860,983	658,255	202,728	136,674
Sex				
Male	433,808 (50.39%)	327,240 (49.71%)	106,568 (52.57%)	79,098
Female	387,793 (45.04%)	300,290 (45.62%)	87,503 (43.16%)	57,576
Unknown	39,382 (4.57%)	30,725 (4.67%)	8,657 (4.27%)	n.a.
Age at DM				
0-18	5,761 (0.67%)	5,761 (0.88%)	n.a.	n.a.
18-44	61,439 (7.14%)	61,439 (9.33%)	n.a.	11,575
45-64	264,350 (30.70%)	264,350 (40.16%)	n.a.	12,466
60>	326,705 (37.95%)	326,705 (49.63%)	n.a.	3,256
Diagnosis				
Diabetes	658,255 (76.45%)	658,255 (100.00%)	n.a.	36,348
Hypertension	427,077 (49.60%)	341,244 (51.84%)	85,833 (42.34%)	62,856
Hyperlipidaemia	365,827 (42.49%)	285,180 (43.32%)	80,647 (39.78%)	44,448
Arthrosis	300,028 (34.85%)	237,748 (36.12%)	62,280 (30.72%)	41,635
Anxiety disorder	183,966 (21.37%)	144,079 (21.89%)	39,887 (19.68%)	36,267
Heart failure	153,809 (17.86%)	123,359 (18.74%)	30,450 (15.02%)	1,674
Obesity	129,938 (15.09%)	100,269 (15.23%)	29,669 (14.63%)	42,206
Ischemic heart dis.	103,882 (12.07%)	82,692 (12.56%)	21,190 (10.45%)	6,031
COPD	102,507 (11.91%)	82,808 (12.58%)	19,699 (9.72%)	8,760
Tobacco depend.	94,255 (10.95%)	72,237 (10.97%)	22,018 (10.86%)	20

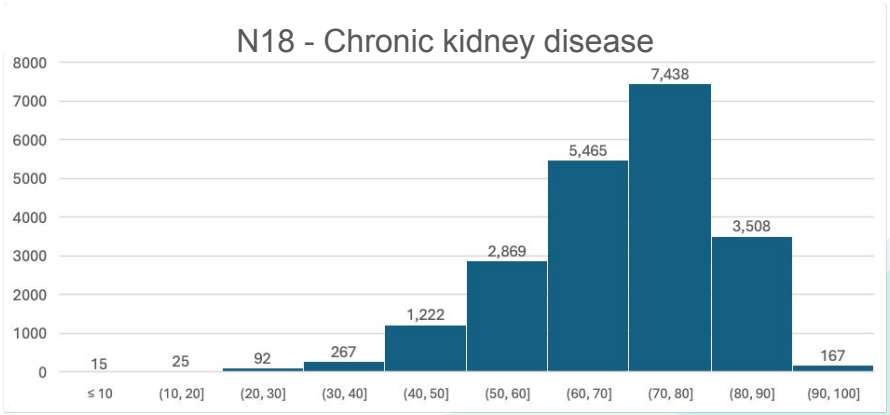
Table 1. Baseline characteristics

Exploratory data analysis



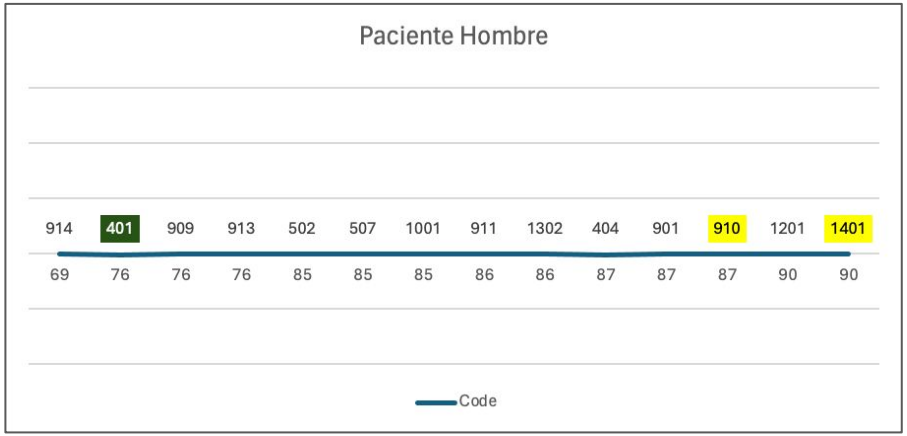




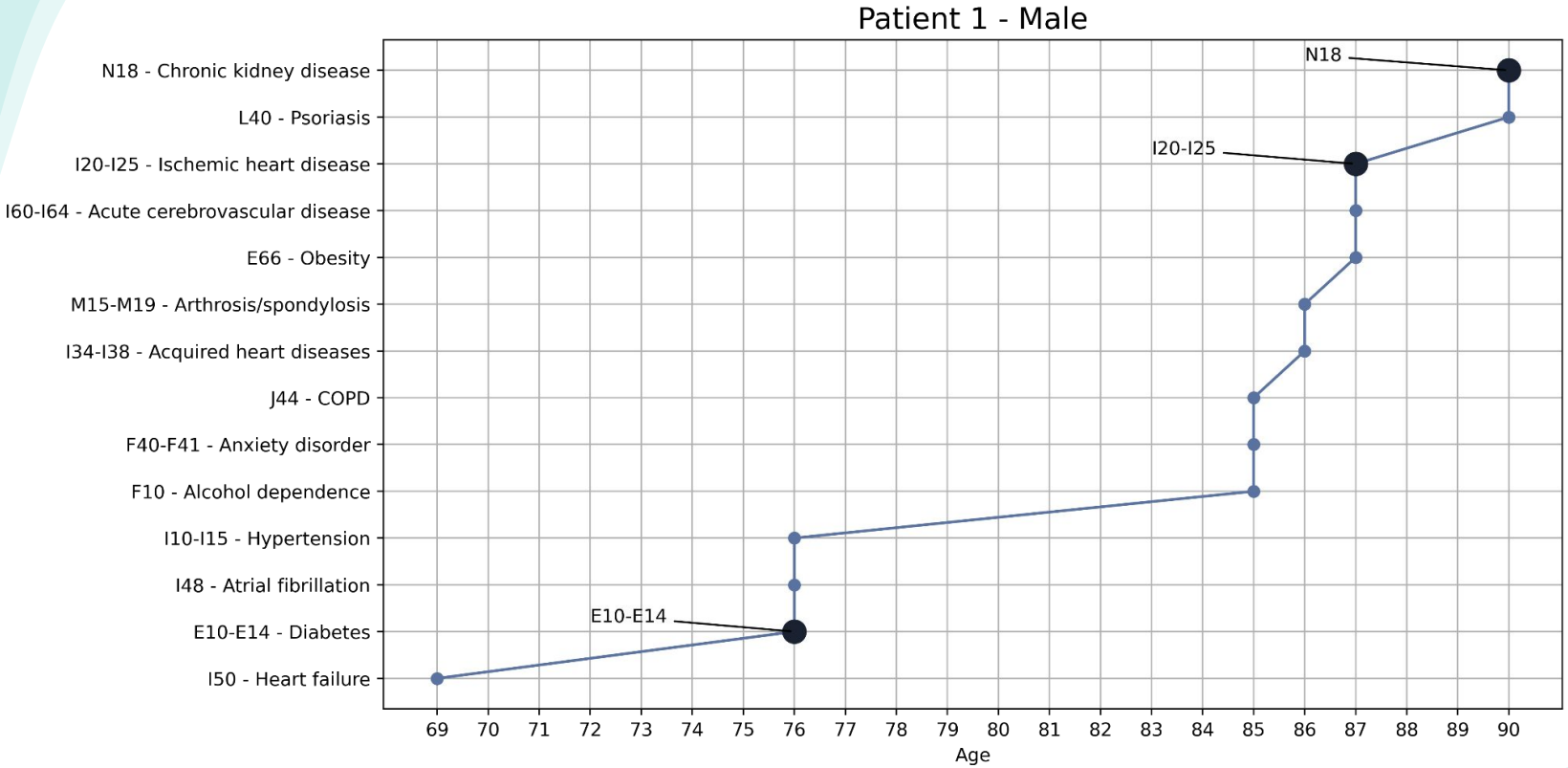


Historia de un paciente

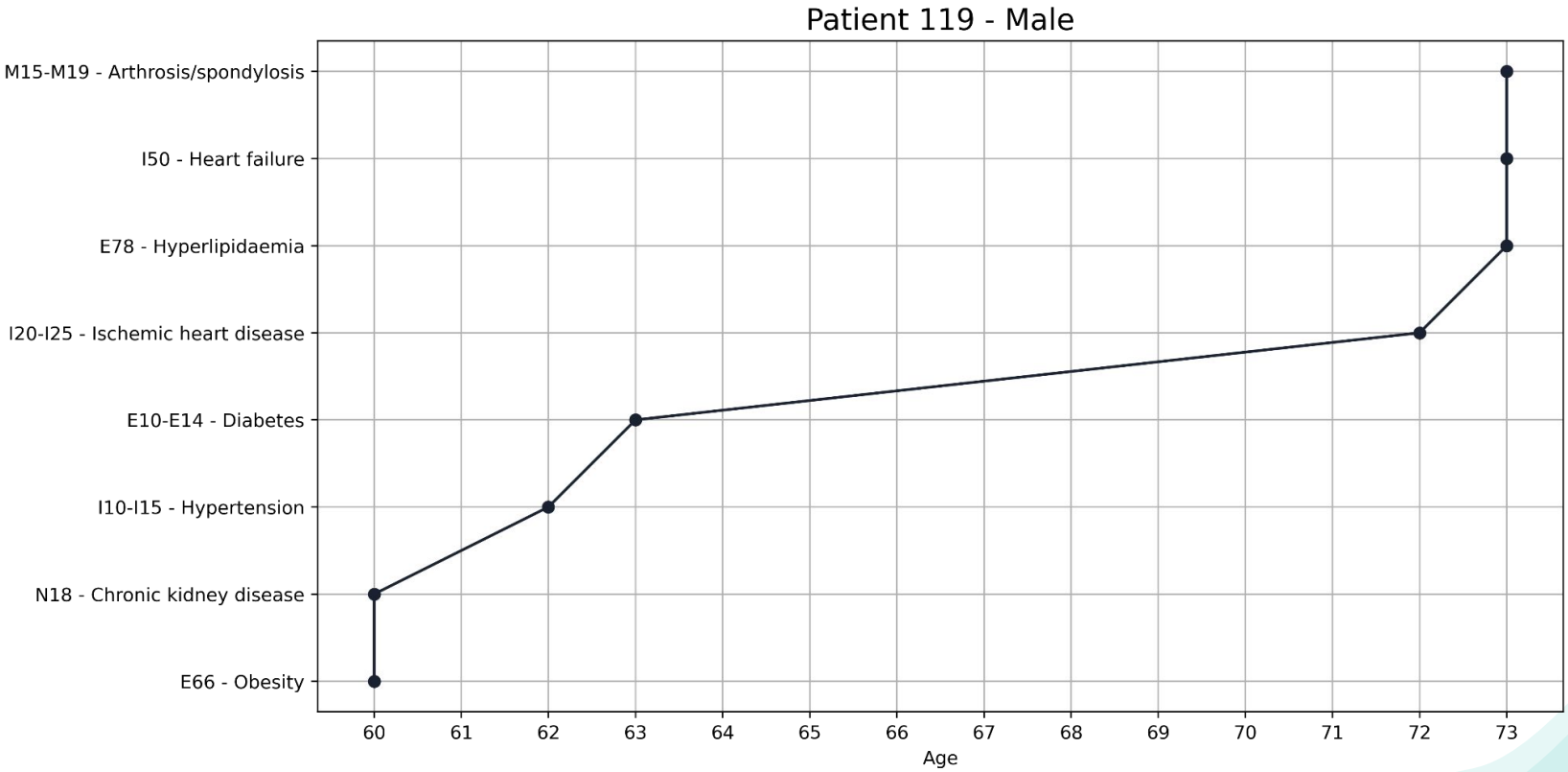
Code	Age	cie	cie name
914	69	I50	Insuficiencia cardiaca
401	76	E10-E14	Diabetes
909	76	I48	Fibrilación auricular
913	76	I10-I15	Hipertensión
502	85	F10	Dependencia alcohol
507	85	F40-F41	Trastorno de ansiedad
1001	85	J44	EPOC
911	86	I34-I38	Enfermedad valvular adquirida
1302	86	M15-M19	Artrosis/espondilosis
404	87	E66	Obesidad
901	87	I60-I64	Enfermedad cerebrovascular aguda
910	87	I20-I25	Cardiopatía isquémica
1201	90	L40	Psoriasis
1401	90	N18	Insuficiencia renal crónica



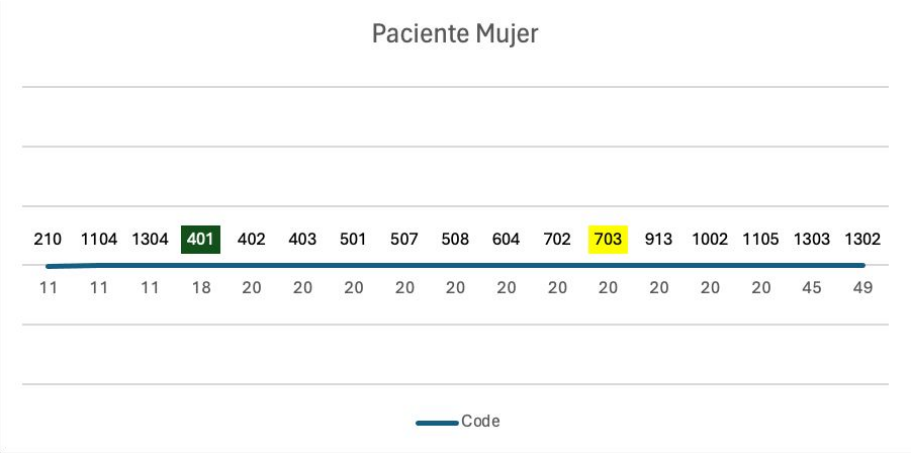
Patient journey



Historia de un paciente



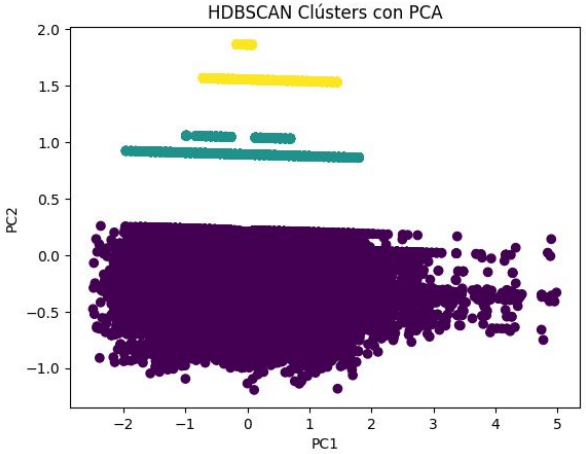
Code	Age	cie	cie name
210	11	C50	Cáncer de mama
1104	11	K21	Enfermedad por reflujo gastroesofágico
1304	11	M13	Otra artropatía
401	18	E10-E14	Diabetes
402	20	E78	Dislipemia
403	20	E03	Hipotiroidismo
501	20	F09	Otro trastorno mental orgánico
507	20	F40-F41	Trastorno de ansiedad
508	20	F50	Trastorno conducta alimentaria
604	20	G80-G83	Enfermedad neurológica con déficit motor no ACV
702	20	H40-H42	Glaucoma
703	20	H35	Retinopatía
913	20	I10-I15	Hipertensión
1002	20	J45	Asma
1105	20	K50-K51	Enteritis regional y colitis ulcerosa
1303	45	M10	Gota y otras artropatías por cristales
1302	49	M15-M19	Artrosis/espondilosis

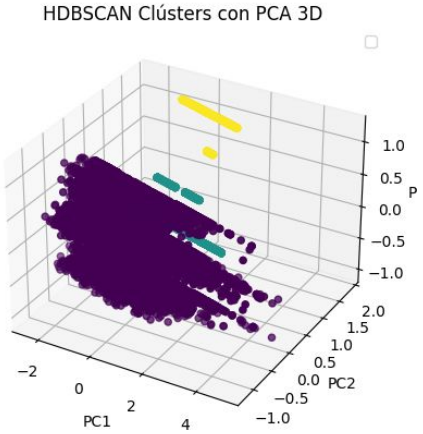


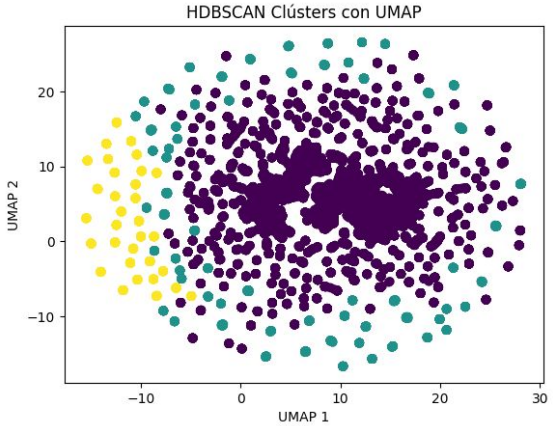
Machine Learning - Unsupervised

Pred1y_E10-E14

```
Silhouette Score: 0.5070738010551737
Calinski-Harabasz Index: 6305.610658387975
Davies-Bouldin Index: 0.7837843341507784
```

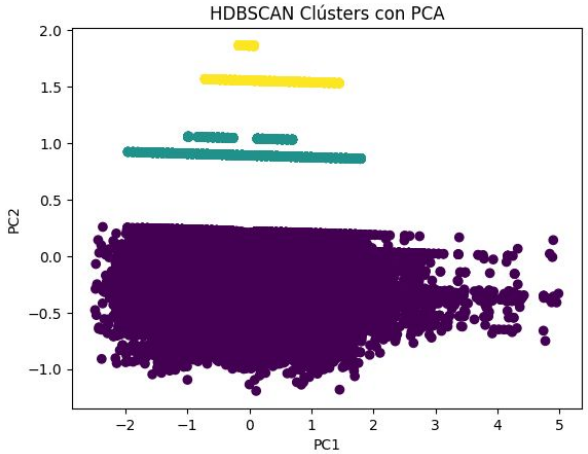






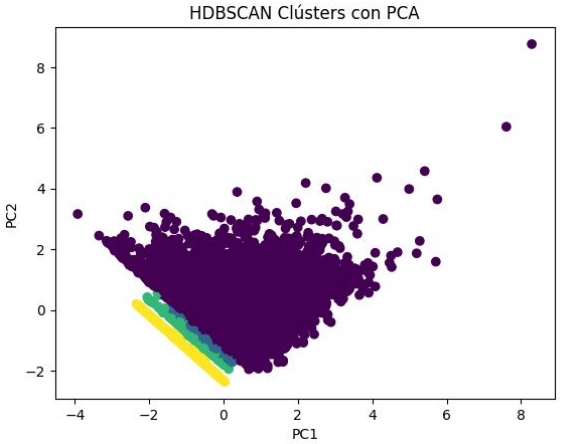
Pred1y_E10-E14: cluster statistics

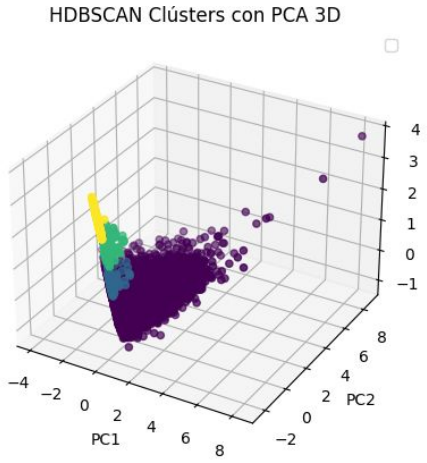
```
Cluster cluster_0:
  Count: 6507
  E78: 7.91%
  I10-I15: 100.00%
  pred1y_E10-E14: 38.88%

Cluster cluster_1:
  Count: 2005
  E78: 100.00%
  I10-I15: 10.07%
  pred1y E10-E14: 44.34%
  cluster: 100.00%
```

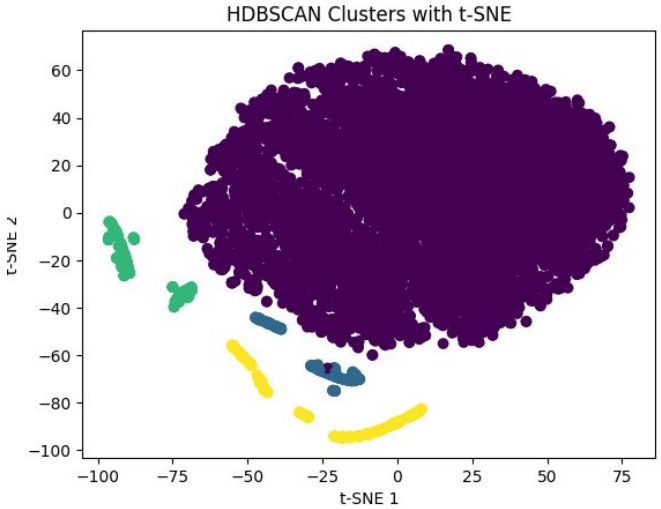


Pred1y N18

```
Silhouette Score: 0.4327961276787492
Calinski-Harabasz Index: 522.2822723900165
Davies-Bouldin Index: 1.0892402140023372
```

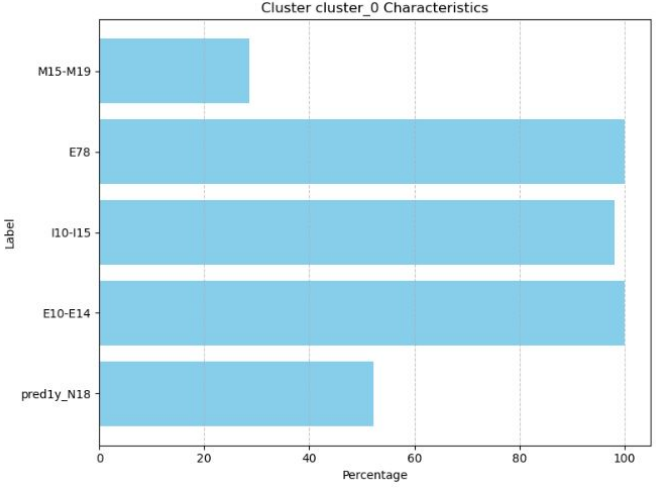


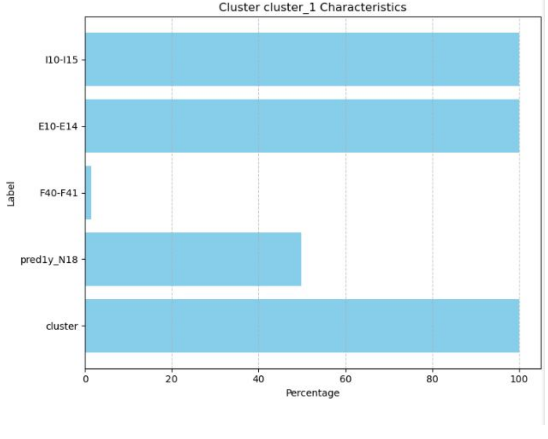




Pred1y N18 cluster statistics

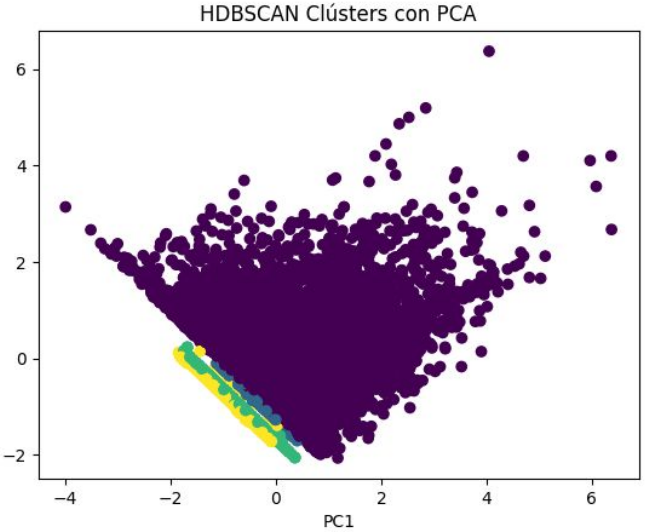
```
Cluster cluster_0:
  Count: 207
  M15-M19: 28.50%
  E78: 100.00%
  I10-I15: 98.07%
  E10-E14: 100.00%
  pred1y_N18: 52.17%

Cluster cluster_1:
  Count: 297
  I10-I15: 100.00%
  E10-E14: 100.00%
  F40-F41: 1.35%
  pred1y_N18: 49.83%
  cluster: 100.00%
```

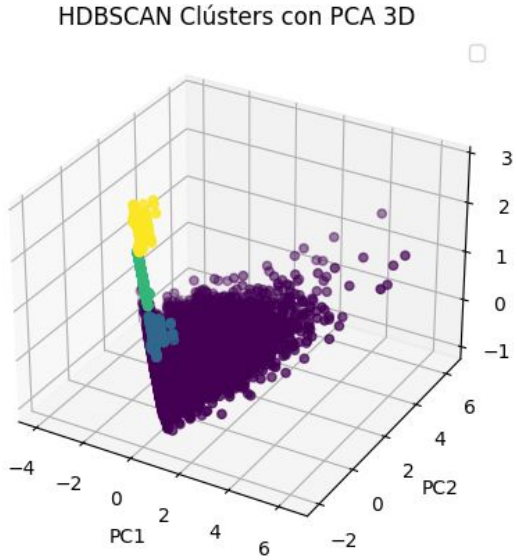


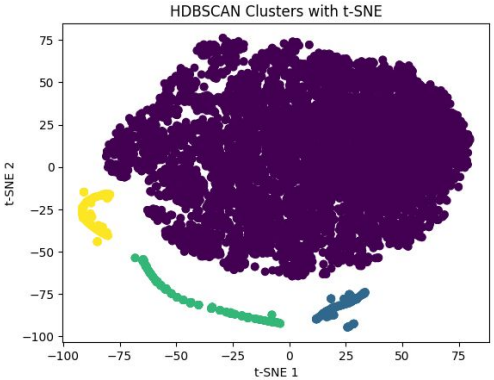


Pred1y I10-I25

```
Silhouette Score: 0.48596086787203774
Calinski-Harabasz Index: 684.7594413744539
Davies-Bouldin Index: 0.9084053095507406
```

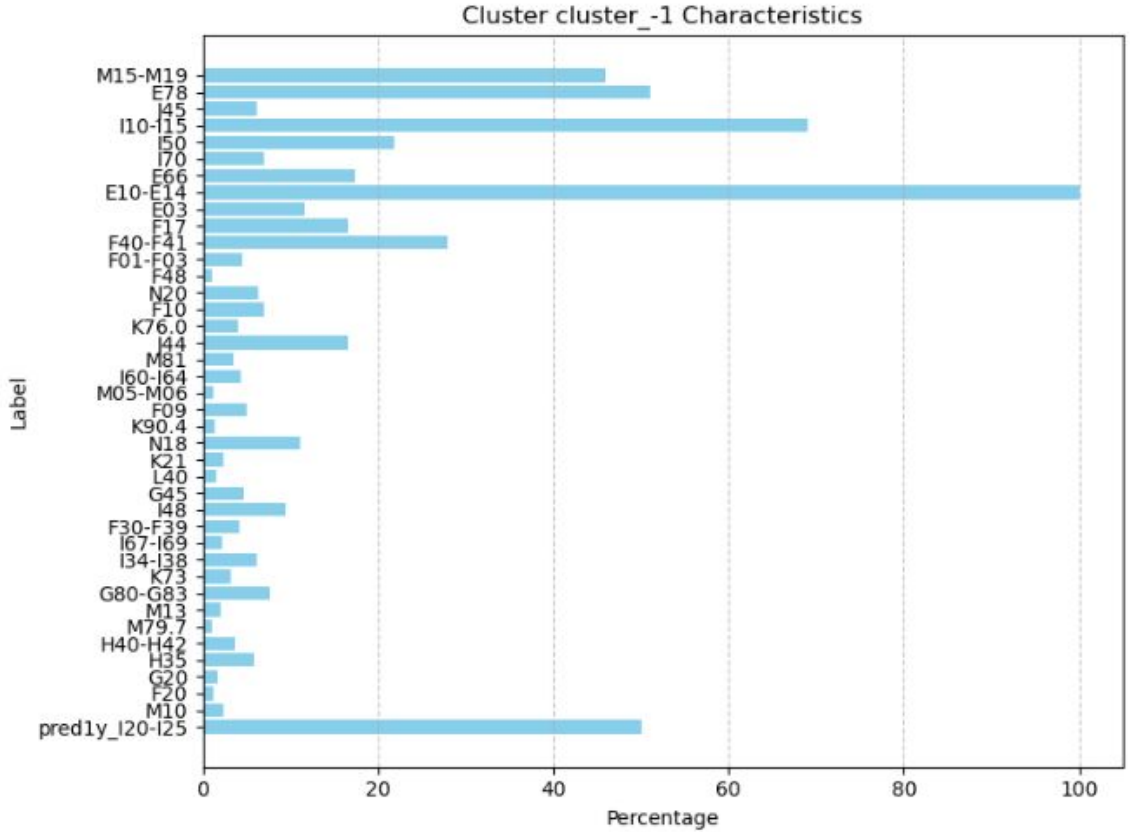






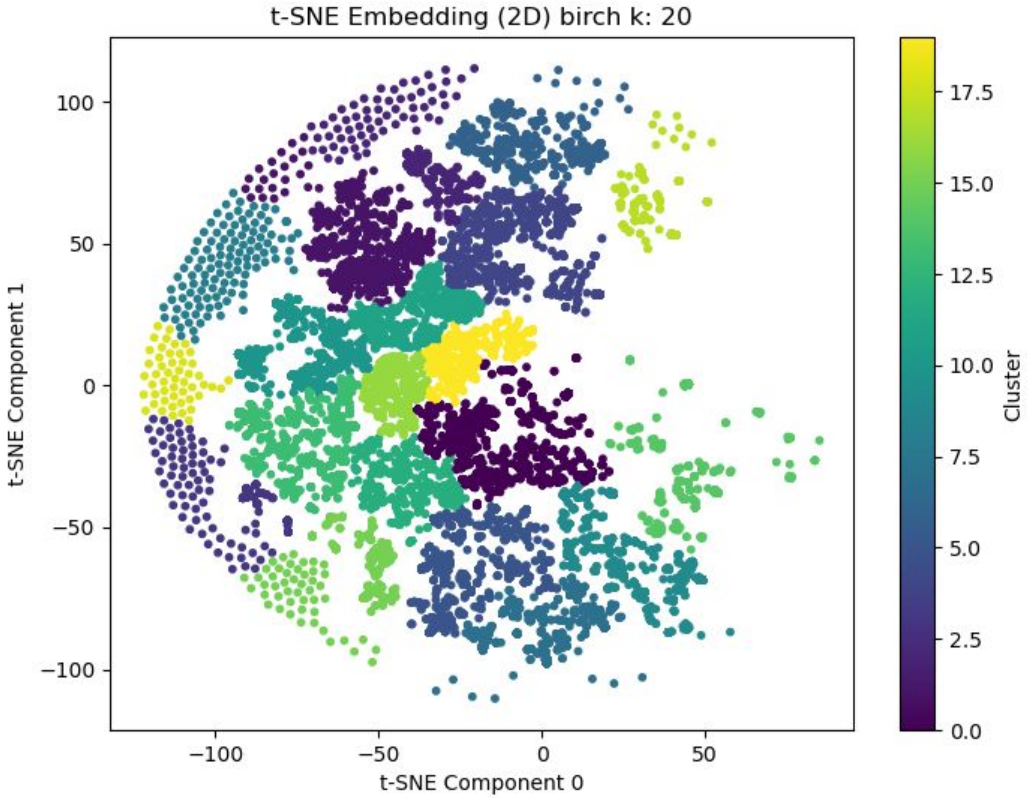
Pred1y I10-125: cluster statistics

```
Cluster cluster_0:
  Count: 214
  M15-M19: 100.00%
  E78: 12.15%
  I10-I15: 96.26%
  E10-E14: 100.00%
  pred1y_I20-I25: 56.07%

Cluster cluster_1:
  Count: 366
  E78: 1.91%
  I10-I15: 100.00%
  E10-E14: 100.00%
  pred1y_I20-I25: 62.30%

Cluster cluster_2:
  Count: 247
  E78: 100.00%
  I10-I15: 90.69%
  E10-E14: 100.00%
  pred1y_I20-I25: 61.54%
```



The optimal number of clusters using K-means was 20

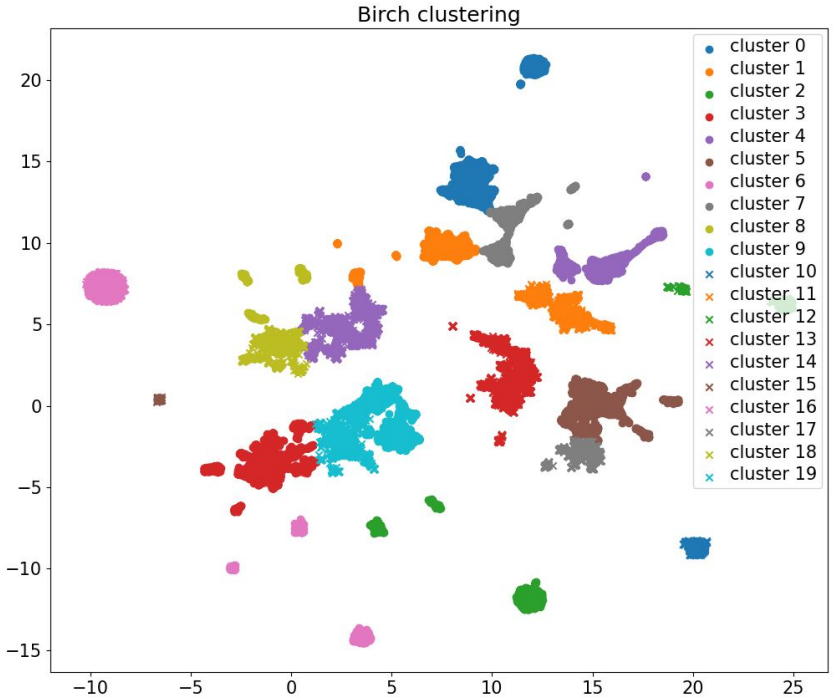
numero de k	davie_bouldin
2	0.59947153
3	0.576883319
4	0.598755167
5	0.661523996
6	0.695442286
7	0.736037083
8	0.726506325
9	0.759551464
10	0.838508708
11	0.810525694
12	0.918350226
13	0.940305845
14	0.918301493
15	0.965000859
16	1.103668931
17	1.052953553
18	1.176387112
19	1.286954245
20	1.260468491



Applying this knowledge

- projection using umap 2D
- considering only diseases
- grouping in 20 clusters using birch

- validated by Amphora's medical team as feasible comorbidity groups
- no predictive capability



Machine Learning - supervised

SVM RBF (ROC AUC)

E10-E14, 1 year	H35, 1 year	N18, 1 year	Z89, 1 year	I20-I25, 1 year
0.588 tsne2d	0.621 umap2d	0.638	0.693	0.635
0.581	0.621	0.618 pca9d	0.652 umap2d	0.633 pca9d
0.578 umap2d	0.594 pca9d	0.603 tsne2d	0.615 tsne2d	0.611 umap2d
0.577 pca9d	0.584 tsne2d	0.598 umap2d	0.614 pca9d	0.592 tsne2d
5 years	**5 years**	**5 years**	**5 years**	**5 years**
0.614 pca9d	0.652	0.640	0.758 pca9d	0.641
0.611	0.637 umap2d	0.629 pca9d	0.703	0.632 pca9d
0.608 umap2d	0.631 tsne2d	0.617 tsne2d	0.682 tsne2d	0.625 tsne2d
0.602 tsne2d	0.626 pca9d	0.607 umap2d	0.613 umap2d	0.621 umap2d

Classifying using Naïve Bayes

	DB	Accuracy	Recall	specificity	Precision	AUC	Tamaño
5 años	Z89 pca9d	0,71	0,71	0,76	0,712	0,71	500
	N18 pca9d	0,618	0,618	0,553	0,619	0,617	27256
	I20-I25 pca9d	0,629	0,629	0,510	0,634	0,627	30308
	H35 umap2d	0,595	0,595	0,380	0,620	0,597	12872
	E10-E14 tsne2d	0,590	0,590	0,545	0,591	0,590	40000
1 año	Z89 umap9d	0,708	0,708	0,733	0,716	0,700	116
	N18 umap9d	0,592	0,592	0,522	0,592	0,590	8626
	I20-I25 pca9d	0,603	0,603	0,470	0,616	0,607	9880
	H35 umap2d	0,578	0,578	0,375	0,607	0,587	4124
	E10-E14 umap2d	0,576	0,576	0,518	0,578	0,577	40000

NN (ROC AUC)

With optimal hyperparameters

E10-E14	H35	N18	Z89	I20-I25
1 year	**1 year**	**1 year**	**1 year**	**1 year**
0.583	0.590 umap2d	0.631	0.606	0.607
0.577 pca9d	0.574 pca9d	0.621 pca9d	0.500 umap2d	0.596 pca9d
0.570 umap2d	0.572	0.571 tsne2d	0.462 tsne2d	0.595 umap2d
0.567 tsne2d	0.530 tsne2d	0.531 umap2d	0.455 pca9d	0.589 tsne2d
5 years	**5 years**	**5 years**	**5 years**	**5 years**
0.613 pca9d	0.623	0.635	0.664 tsne2d	0.646
0.611	0.613 umap2d	0.609 pca9d	0.634	0.636 umap2d
0.604 tsne2d	0.611 pca9d	0.593 umap2d	0.555 umap2d	0.630 tsne2d
0.572 umap2d	0.587 tsne2d	0.569 tsne2d	0.497 pca9d	0.625 pca9d

XGBOOST


XGBoost Performance Metrics
db_name
5y_Z89.csv
5y_I20-I25.csv
1y_I20-I25.csv
5y_H35.csv
1y_H35.csv
5y_N18.csv
1y_N18.csv
5y_E10-E14.csv
1y_E10-E14.csv
1y_Z89.csv
0.86
0.9
0.68
0.74
0.71
0.83
0.47
0.62
0.71
0.84
0.44
0.58
0.71
0.84
0.45
0.6
0.71
0.85
0.44
0.58
0.71
0.83
0.44
0.6
0.69
0.83
0.42
0.6
0.68
0.82
0.39
0.57
0.65
0.84
0.33
0.55
0.61
0.89
0.2
0.4
auc
sensitivity
specificity
precision

With this approach we conducted a validation study

- Using XGBoost we trained different models
- **no_split** means full data usage, **cluster** means data enrichment using clustering, and **diabetia** means validation data.
- each model was trained using CAMDA database and evaluated using diabetIA database

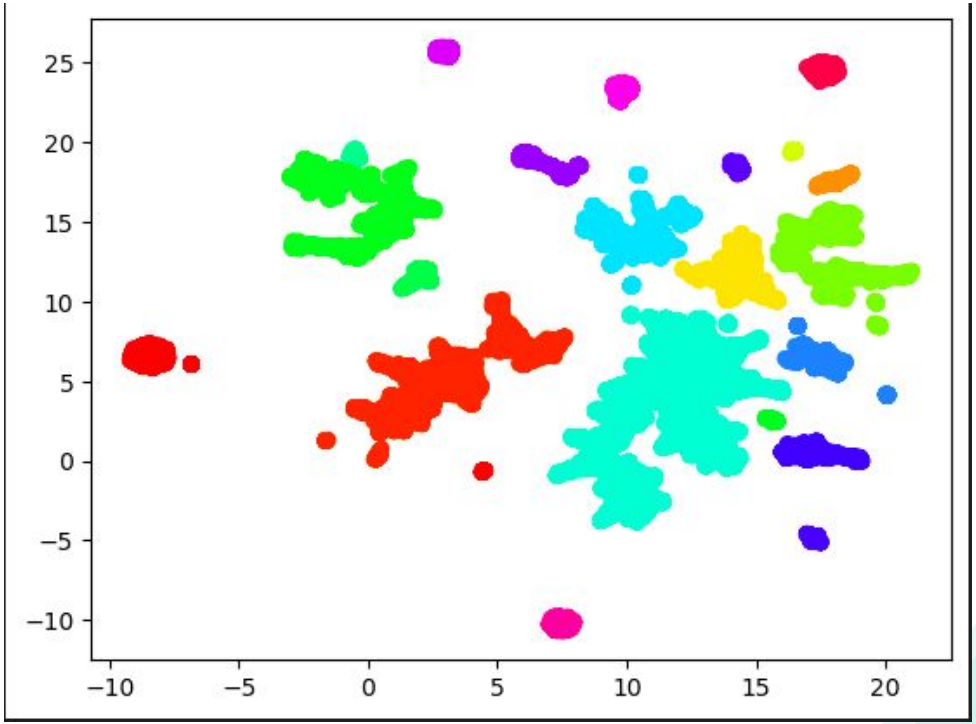
condition	# score	database	method
N18	0.57	complete	no_split + cluster + diabetia
E10-E14	0.56	complete	no_split + diabetia
E10-E14	0.56	complete	no_split + cluster + diabetia
I20-I25	0.55	complete	no_split + cluster + diabetia
I20-I25	0.55	complete	no_split + cluster + diabetia
E10-E14	0.55	complete	no_split + cluster + diabetia
I20-I25	0.55	complete	no_split + diabetia
I20-I25	0.55	complete	no_split + diabetia
E10-E14	0.55	complete	no_split + diabetia
N18	0.55	complete	no_split + diabetia
I20-I25	0.54	complete	no_split + diabetia
I20-I25	0.53	complete	no_split + cluster + diabetia
H35	0.51	complete	no_split + cluster + diabetia
H35	0.51	complete	no_split + cluster + diabetia
H35	0.51	complete	no_split + diabetia
H35	0.49	complete	no_split + diabetia

Other analyzes

We explored the use of TDA in data segmentation

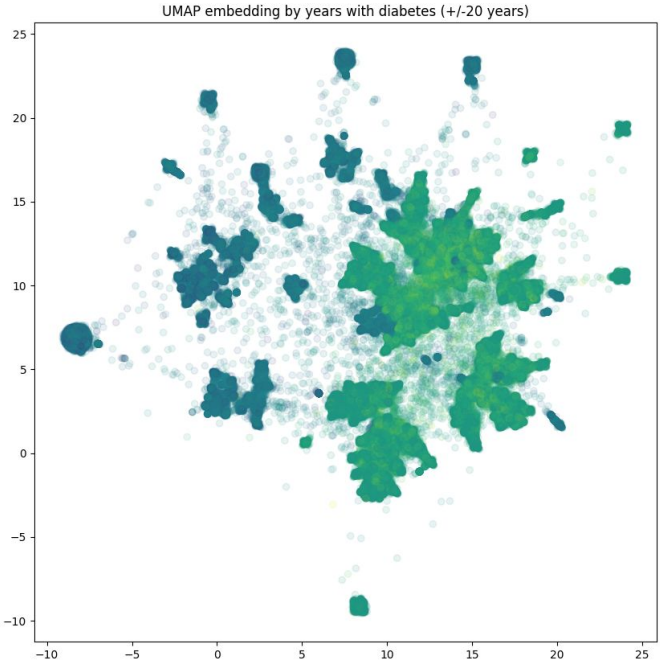
TDA makes better use of geometric data structure in the data segmentation.

The problem is that TDA doent's scale very good with massive data.



Inspired by TDA, some alternatives were explored to understand the progression of diseases

In this approach we can observe differences before and after diabetes but further exploration is required.



How does our results compares with the literature.

Published ML models for diabetic complications

T2D complication	N	Model	Classification Problem	AUC	Source
Kidney Disease	3,816,329 Russian patients	XGB	6-month	0.79	(Derevitskii, 2020)
	380 EHR Chinese patients	GBM	1-year	0.88	(Zou, 2022)
	11,789 Clinical Trials participants	NN	1-year	0.82	(Nagaraj, 2020)
	135,284 CAMDA patients	XGB	1-year	0.68	CAMDA Team
Diabetic Foot	326,853 patients USA	DT	All time	0.88	(Stefanopoulos, 2022)
	2,559 inpatients Singapore	XGB	All time	0.82	(Oei, 2024)
	135,284 CAMDA patients	XGB	1-year	0.61	CAMDA Team
Ischemic heart disease	3,816,329 Russian patients	XGB	6-month	0.84	(Derevitskii, 2020)
	135,284 CAMDA patients	XGB	1-year	0.65	CAMDA Team
Retinopathy	3,816,329 Russian patients	XGB	6-month	0.79	(Derevitskii, 2020)
	135,284 CAMDA patients	XGB	1-year	0.71	CAMDA Team

Discussion

1. The synthetic CAMDA dataset yields interesting patient journeys, yet they need validation from our clinical team.
2. Our clustering and TDA techniques allows for the identification of subgroups among the diabetic patients.
3. Our supervised methods are performing well, given the presence of diagnosis data only, but could benefit from additional variables such as laboratory values, weight, height, BMI, and other clinical findings.

Future work (before paper submission)

1. Improve the DiabetIA database integration to the pipeline as a validation set.
2. Inverse the training and validation sets to test the transfer learning challenge.
3. Use LSTM to estimate the most probable next step during the patient's trajectories

Thank you to the CAMDA Diabetes subgroup!

1. Universidad Nacional Autónoma de México, Centro de Ciencias Matemáticas
2. Universidad Nacional Autónoma de México, Escuela Nacional de Estudios Superiores Unidad Morelia
3. Universidad Michoacana de San Nicolás de Hidalgo, Facultad de Ciencias Físico Matemáticas
4. Amphora Health