A-proof report

COVID-19 rehabilitation patterns

1. data-mining
2. time-series modeling

Bruna A. Guedes

Overview

Medical team

- Research Problem
  - medical team : data-mining
- Data statistics
  - Within admission
  - After discharge
  - Tool for data analysis

Machine Learning ADM prediction

- Research Problem
  - AI: time-series modeling
- Feature Engineering
- Modelling and Results
- Discussion and Outlook

For more information please check this Github repository
And this folder on google docs with progress over time

Medical team

Medical team Research questions

- What is the mean or median level of functioning on the different ICF domains at hospital admission, at hospital discharge and at the 6 weeks and 3 months outpatient visits?
- What is the mean course in the level of functioning of these domains during hospital stay (from admission to discharge)?
- What is the frequency of notes per ICF domain/level?
- What different patterns in recovery of functioning can be distinguished?

Filters applied for Data analysis only

Initial dataset: 110781 instances

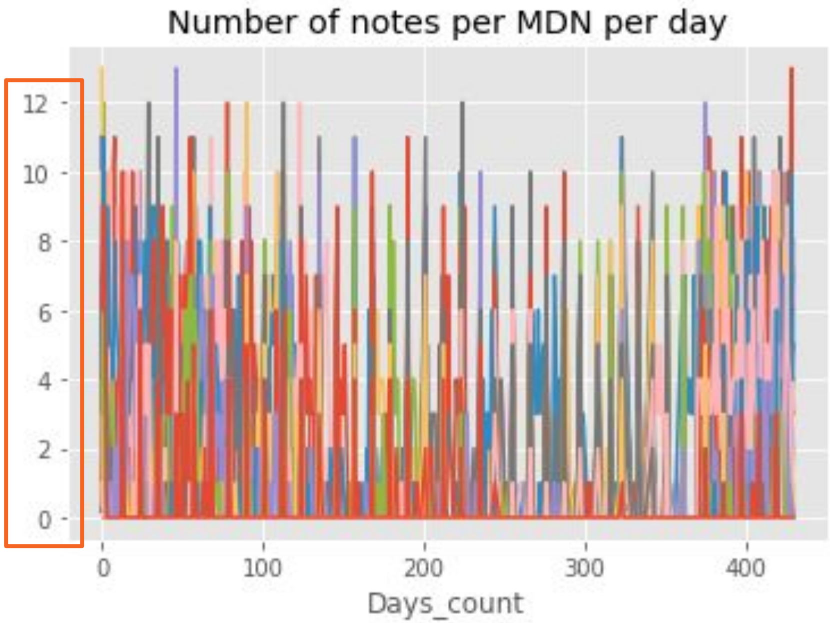
- Drop instances with no evaluation for any domain studied: 71287 instances and 1289 unique patients
- Max days difference between note and previous note: 428.0
- Filter considering discharge if more than 2 days with no annotations:58093 instances and still 1289 patients (they all have a day 0)
  - n unique days count: 386
  - min days count: 0
  - mean days count: 199
  - max days count: 429
    - obs: it doesn't mean there are notes every day, it's from the first admission to the last day of note from the patient

Data statistics

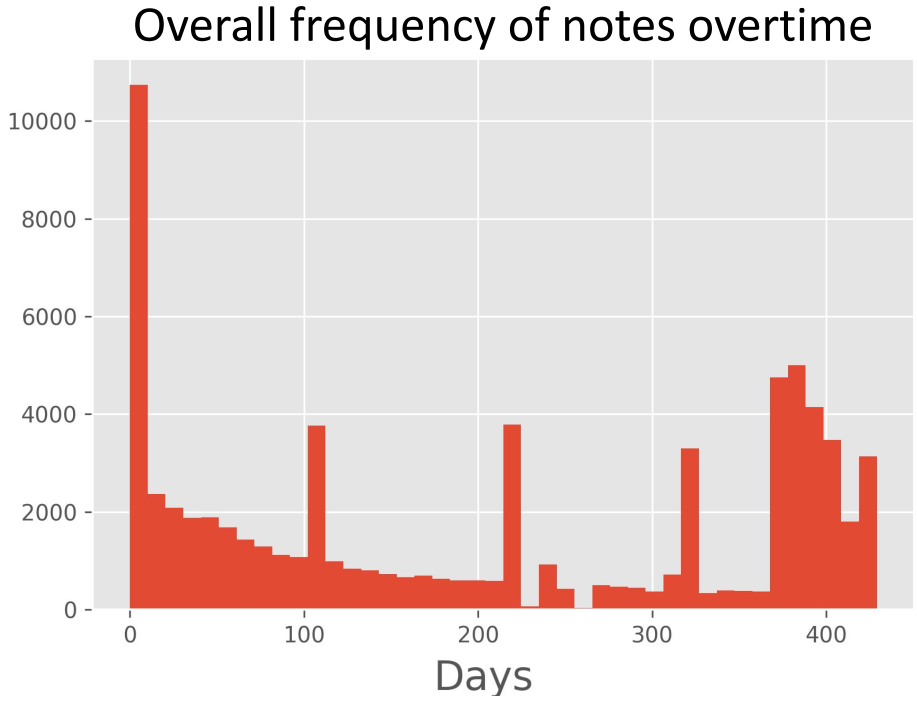
Dataset description

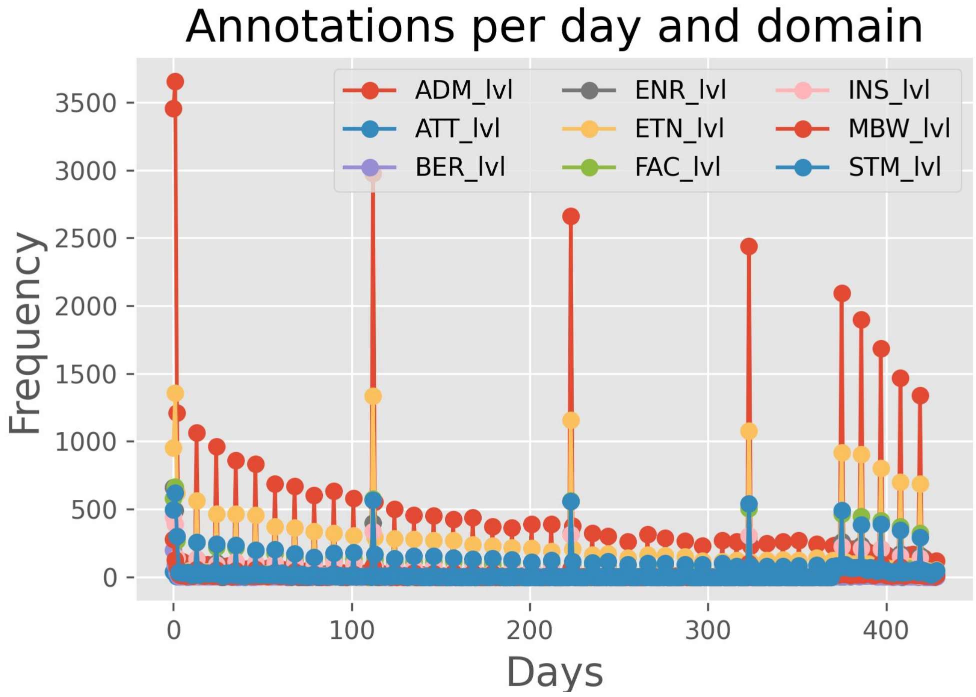
- Dataset from 1/1/2020 until 31/03/2021
- Number of instances with info of at least one domain: 71287
- Initial number of patients: 1289
- Max days difference between note and previous note: 428.0
- N unique days count: 386
- Min days count: 0
- Mean days count: 199 days
- Max days count: 429

	uniqueID	total notes
>50 per ID	443	51933
>100 per ID	192	34400
>500 per ID	2	1105

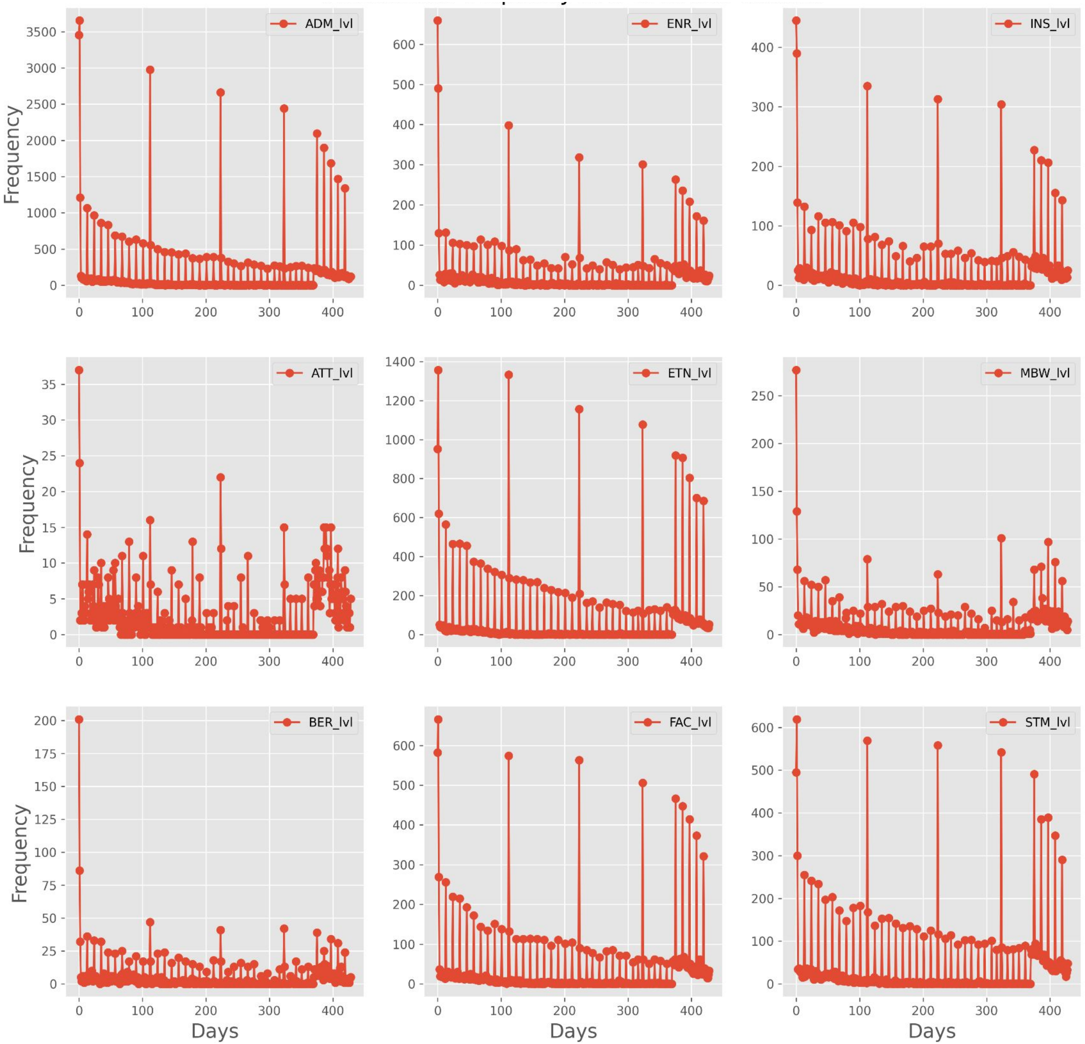


Distribution of frequency of notes over time

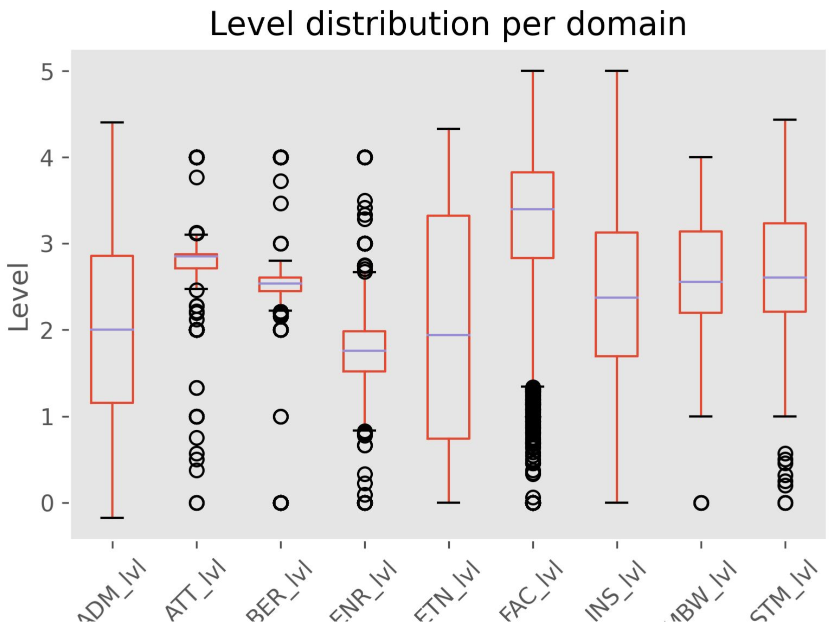




Annotations frequency over time and domain



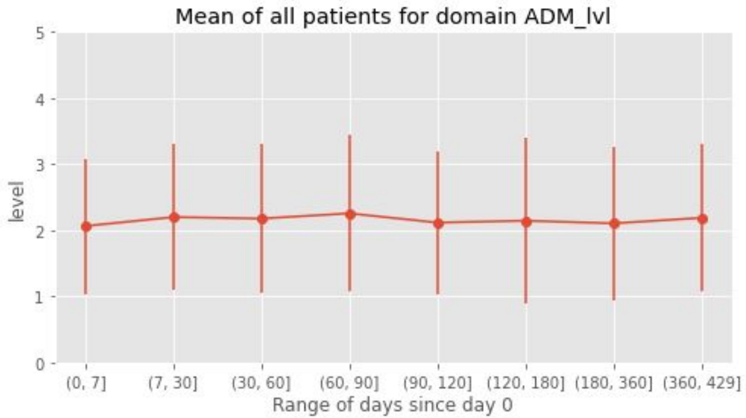
Outliers and missing values

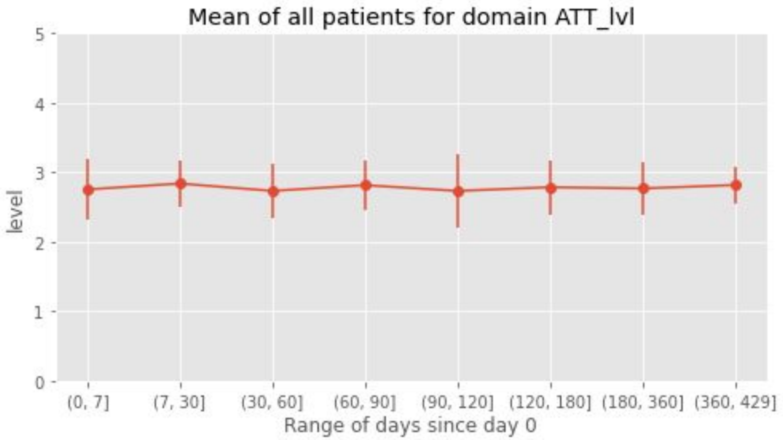


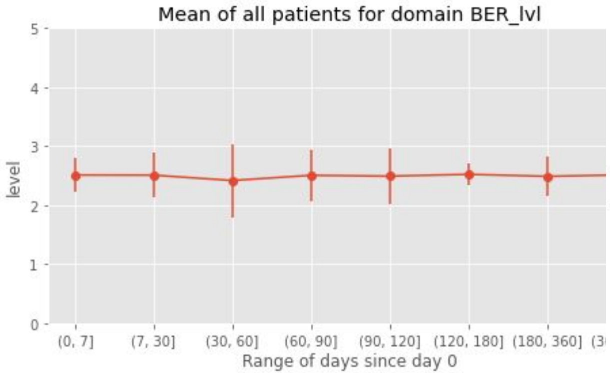
- Missing values
  - Interpolation of ADM_lvl and removal of other domains for modelling

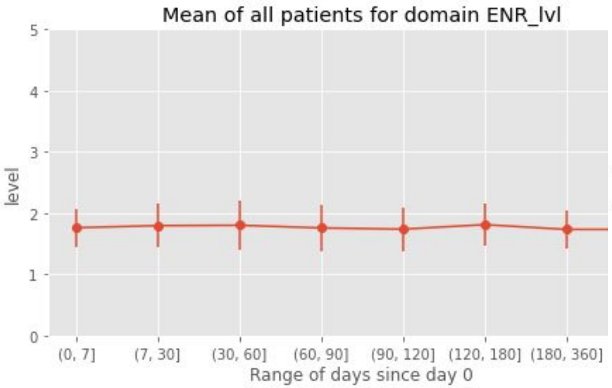
ADM_lvl	4777
ATT_lvl	17746
BER_lvl	17542
ENR_lvl	15966
ETN_lvl	10443
FAC_lvl	15019
INS_lvl	16257
MBW_lvl	17285
STM_lvl	14146

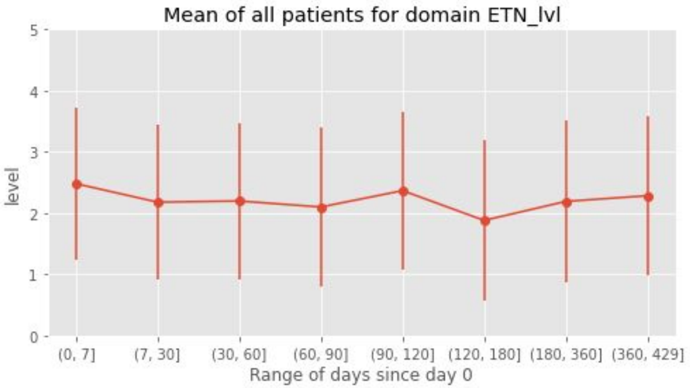
Mean evolution aggregated data for whole period



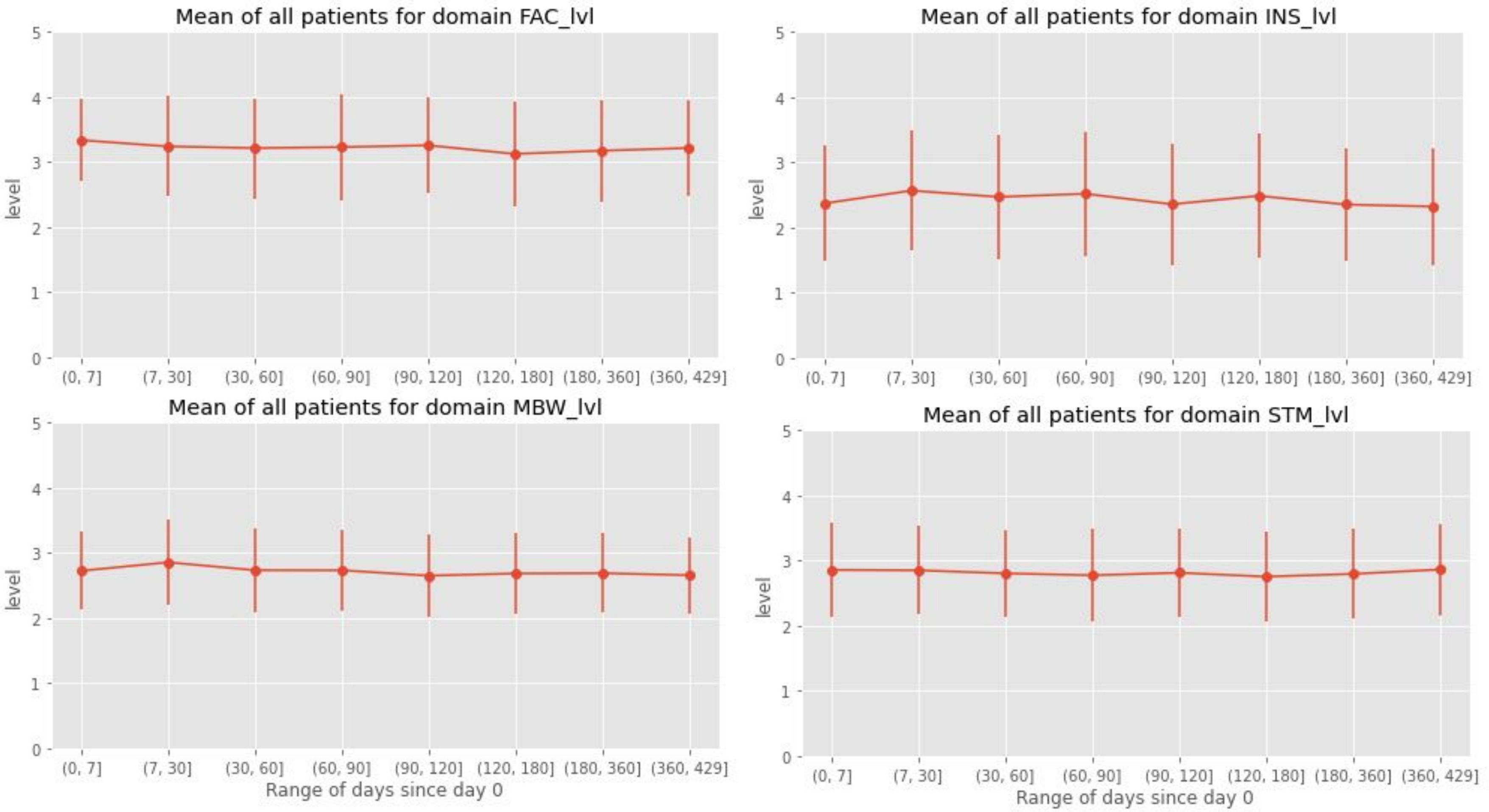






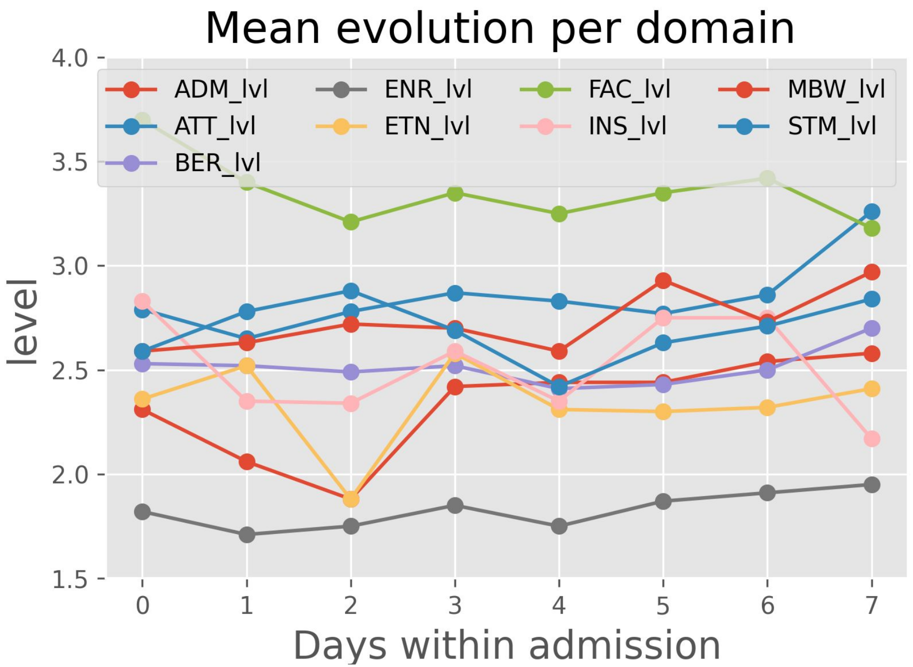


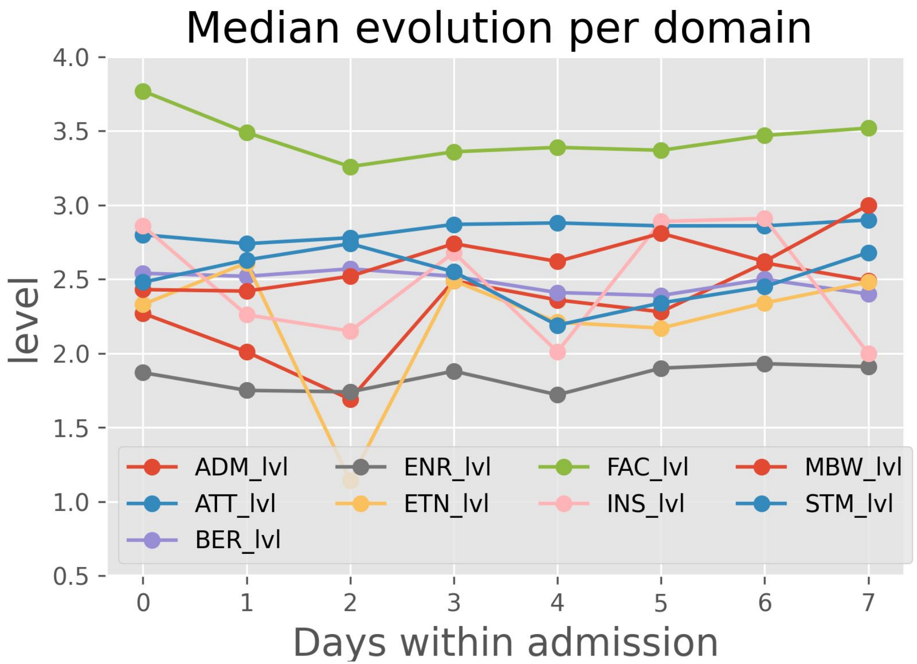
Mean evolution aggregated data for whole period



Within admission

Mean and median daily evolution within admission





After discharge

- Number of unique patients with discharge information: 1222
- Number of unique patients from discharge to 6w: 1222
- Number of unique patients from 6 weeks discharge to 3 months: 941

Discharge evolution per domain

	Mean			Difference		Count		
	day 0	0-6w	6w-3m	date1-0	date2-0	day 0	0-6w	6w-3m
ADM_lvl	2.19	2.12	2.16	-0.07	-0.03	960	2636	2580
ATT_lvl	2.72	2.76	2.79	0.04	0.07	26	61	89
BER_lvl	2.47	2.49	2.48	0.02	0.01	64	141	237
ENR_lvl	1.81	1.83	1.85	0.02	0.04	185	457	552
ETN_lvl	2.43	2.15	1.83	-0.28	-0.60	372	1111	1210
FAC_lvl	3.55	3.45	3.42	-0.10	-0.13	217	612	708
INS_lvl	2.61	2.69	2.68	0.08	0.07	172	428	574
MBW_lvl	2.72	2.65	2.69	-0.07	-0.03	77	225	353
STM_lvl	2.75	2.73	2.71	-0.02	-0.04	210	668	880

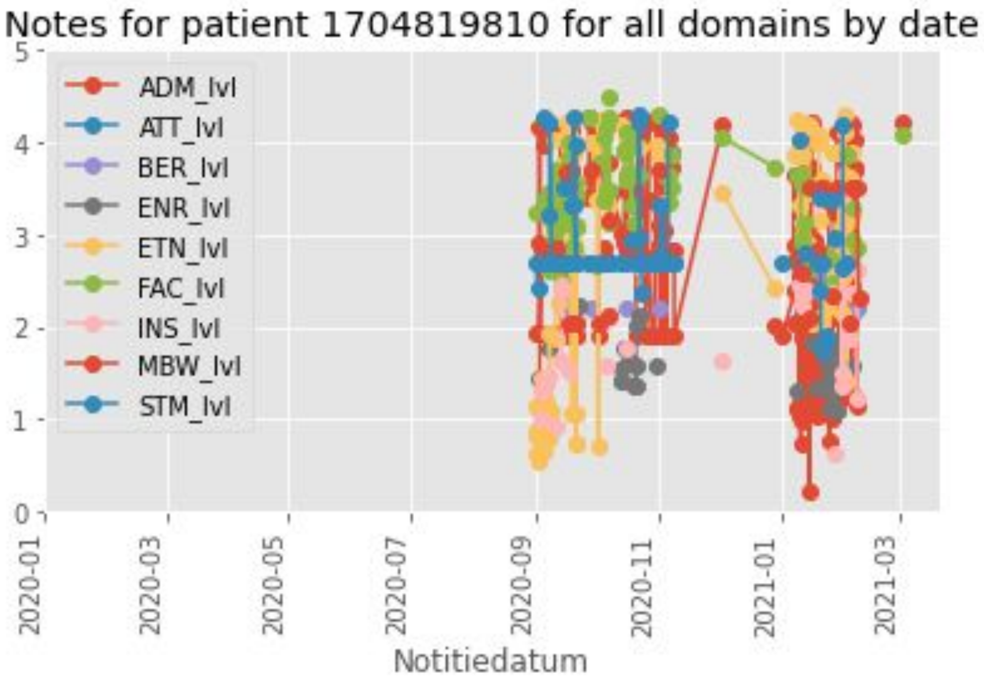
Tool for data analysis

Other analysis that can be done

Statistics from specific day

	Day 0				
	mean	min	max	median	count
ADM_lvl	2.31	-0.08	4.38	2.27	3454
ATT_lvl	2.79	0.00	4.00	2.80	37
BER_lvl	2.53	2.19	4.00	2.54	201
ENR_lvl	1.82	0.00	3.00	1.87	659
ETN_lvl	2.36	0.00	4.24	2.33	951
FAC_lvl	3.70	0.00	5.00	3.77	582
INS_lvl	2.83	0.00	4.29	2.86	445
MBW_lvl	2.59	1.00	4.00	2.43	277
STM_lvl	2.59	1.00	4.33	2.48	495

Case by case analysis of evolution per patient



Machine Learning

ADM prediction

Machine Learning Goal

Goal of project: Build a time series analysis model to predict rehabilitation behavior (levels) for ADM domain for a new patient overtime.

Filters applied for modeling:

- patients with minimum of 100 notes in total;
- and 100 distinct dates;
- with at least 50 notes in the same domain for at least one domain (to analyse the domains in a level layer)

from 1290 unique IDs to 149 when adding restrictions.

frequency over time (up to 12 annotations per ID per date)

Feature Engineering

Creation of time related features:

- Average domains
- Lag of 1, 2 and 3, for ADM feature
- Rolling mean - window sizes 3 and 7
- Rolling min - window sizes 3 and 7
- Rolling max - window sizes 3 and 7
- Expanding window mean

Other changes in dataset:

- Interpolation of ADM to solve missing values
- Discretization of ADM_lvl variable

Modelling and results

Learning setup:

Split of data: divided into 80% of patients in training and 20% of patients in test

- Shapes of DF Train: (13904, 24) | (13904,)
- Shapes of DF Test: (3695, 24) | (3695,)

Hypothesis: Model needs to reflect temporal info!

Models applied to training and test:

Accuracy on training data for each model:

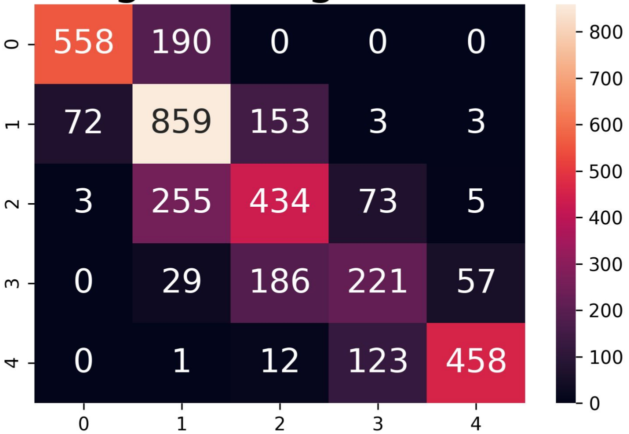
- Log Reg: 0.704
- KNN: 0.754
- Gauss Naive Bayes: 0.626
- Decision Tree: 1.0
- Random Forest: 0.999

F1-macro score for test set for each model:

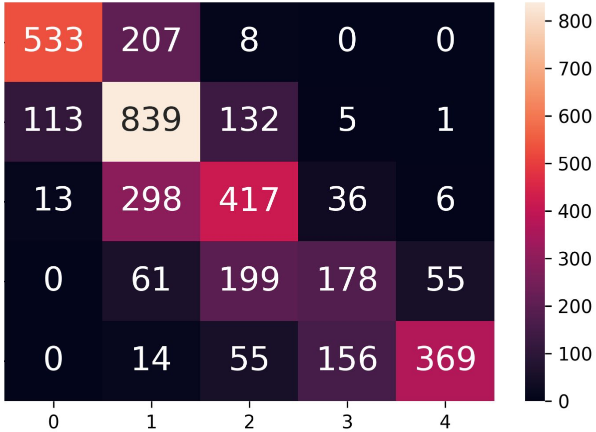
- Log Reg: 0.676
- KNN: 0.617
- Gauss Naive Bayes: 0.625
- Decision Tree: 0.850
- Random Forest: 0.852

Modelling results - Confusion matrix per model

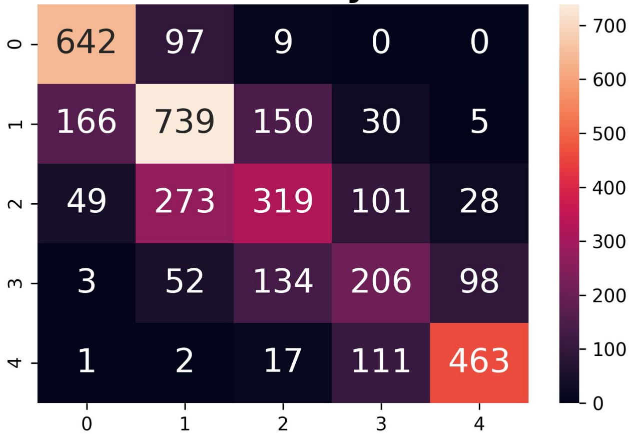
Logistic Regression



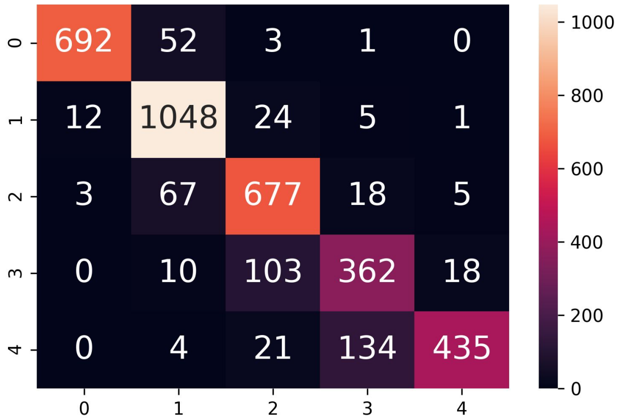
KNN



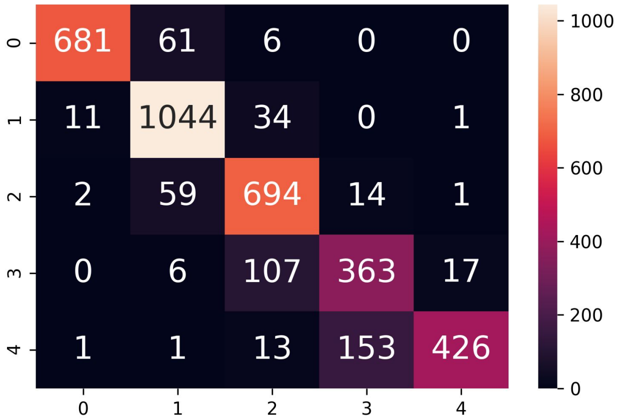
Naive Bayes



Random Forest



Decision Tree



Discussion and Outlook

For ADM feature Random Forest had the best performance in the test set closely followed by Decision Tree

The notebooks created allows a great cross of information to be explored, overall or by patient, domain-wise, frequency-wise and to track evolution

Many possibilities of more sophisticated models and using Predictive modeling with notion of time!