Technical Interview

DAVOS CK CARE - KÜHNE FOUNDATION

Mr JOBARD Aurélien | Data Management position | 31.12.2024

Introduction

- Atopic dermatitis (AD), also known as atopic eczema, is a long-term type of inflammation of the skin.
    - Regarding pathophysiology, the downstream JAK pathway is in particular activated and lead to the release of **antigen specific IgE.**
    - The cause of AD is not known, although some evidence indicates environmental, immunologic, and potential genetic factors.
    - For the diagnosis, the UK Diagnostic Criteria can be used and has been the most widely validated (Criteria used such as **asthma**, **allergic rhinitis** ...).
    - The most commonly used topical treatments for AD are **topical corticosteroids** and moisturisers to help keep control.
    - Clinical studies often measure the efficacy of treatments with a severity scale such as the **SCORAD score** or the **EASI score.**

Source: *Wikipedia.*

- Regarding the Swiss nLPD and due to the sensitive medical context, real world data from CK Care were transformed into synthetic data to be used in the context of this exercise.
- Available data in this exercise are :
    - A 1st dataset, called df1, with clinical data and a shape of (800, 90) including index column. Concepts available are patients' demographic, medical history, treatment details and outcomes.
    - A 2nd dataset, called df2, with information regarding availability for 4 types of biological sample and a shape of (800, 15) including index column.
    - A 3rd dataset corresponding to a data dictionary (German features to English meaning) and timepoint of data collection.
- Python version 3.10.12. was used for this work and the following libraries :
    - Pandas – version 2.2.2.
    - Numpy – version 1.26.4.
    - Matplotlib – version 3.8.0.
    - Seaborn – version 0.13.2.
- Global objectif of this exercise was to transform initial data into longitudinal data (Three timepoint t0, t1, t2), create a subset from the initial dataset to characterize Atopic Dermatitis (AD) / Non AD data and lastly achieve specific queries to identify patient of interest.
- Particular emphasis was placed on data quality control.

Quality Control diagnostic – General comment:

- **Duplicata:**

No duplicata regarding anonymized patient ID was detected both in df1 and df2.

- **Missing values or inconsistent values:**

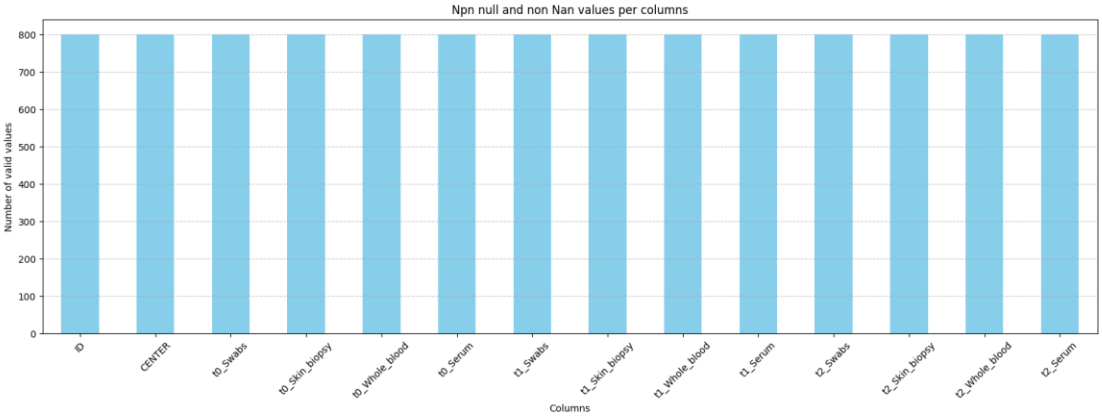
For the feature ‘Birth Month’, inconsistent values equal to 13 and -8 were detected as well as missing values.

For the feature ‘Birth Year’: inconsistent values equal to -8, 1387, 7001 and 7002 were detected as well as missing values.

NB: To calculate age of patients from features ‘Birth_Month’ and ‘Birth_Year’ inconsistent values were transformed into INTEGER format and replaced by 0 or 1 for a better identification. Then for basic statistic using age, it is recommended not to consider these patient to avoid misinterpretation of final results. Impact is insignificant due to the low number of inconsistent values (1, 75 % values not considerated).

For the features ‘(t0, t1, t2)_Visit_Date’: inconsistent values equal to 20271230, 20730721, 70230619, 70231127, and 70730818 were detected as well as missing values. No data treatment were applied to these features but should be achieved in case of data analysis based on these values.

A global diagnostic of missing values or NaN values for features in both df1 and df2 was established. For example, a graph shows below absence of missing/NaN values (YES/NO values only) into df2.



- **Wording and language:**

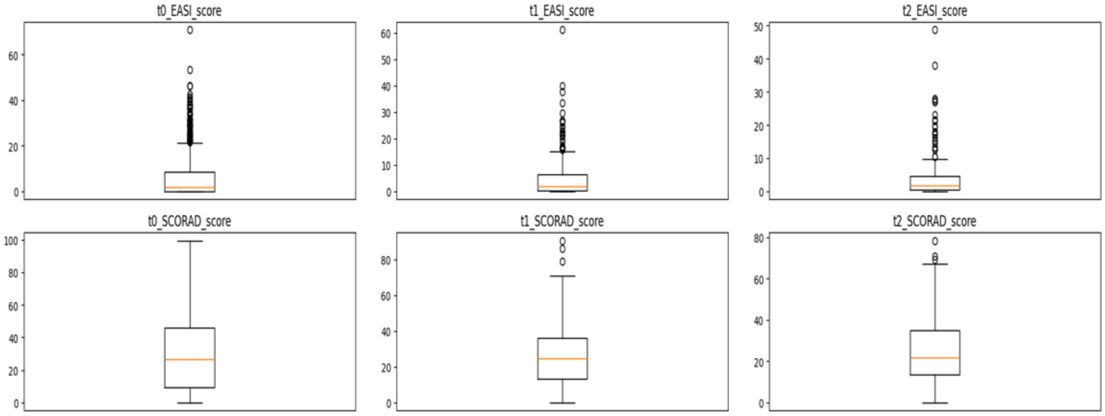
For the feature ‘Name_Treatment’, human reporting error were identified and a specific attention was considered for the writing of the name of treatment “DUPILUMAB” (See question3). A dictionary was constructed to consider all type of errors: {dupilumab, dupilumap, dupixent, dupilvmab}.

For optimized Python code, especially in loops, the feature ‘Name_Treatment_1_’ could be changed to ‘Name_Treatment_1’ such as the others one.

Binary answers (ja/nein) into df1 are written in German whereas binary answers (Yes/No) into df2 are written in English. Responses should be harmonized for a better data processing.

- **Outliers:**

Regarding EASI_Score and SCORAD_Score, we observed these values can reach respectively 61, 2 and 90, 3. From a statistical viewpoint (IQ method), these values are considered as outlier. Nevertheless, from a clinical viewpoint, these values could be possible. That's why, these data should be validated by a medical community. Here below are the boxplot related to both these scores at the three timepoint considered:



Regarding IGE values, we observed that standard concentration can reach 50 000 UI.mL-1. From a statistical viewpoint (IQ method), these values are considered as outlier. Nevertheless, from a pathophysiological viewpoint, these values could be possible. That's why, these data should be validated by a medical community.

- **Timepoint for data collection:**

We need to ensure that the right data is collected at the right time. For example, a quality control could be achieved on Visit_date to ensure that 't0_Visit_date' < 't1_Visit_date' < 't2_Visit_date'.

- **Bias:**

Regarding question 2, both numerical and categorical table "Patient characteristics" were calculated based on the longitudinal table without any distinction between the three timepoint which could be a bias. With more time, a more precise work could be done. Furthermore, due to the use of synthetic data, the data transformation process could be the source of statistical bias regarding distribution of values and could consequently bias interpretation of statistics.

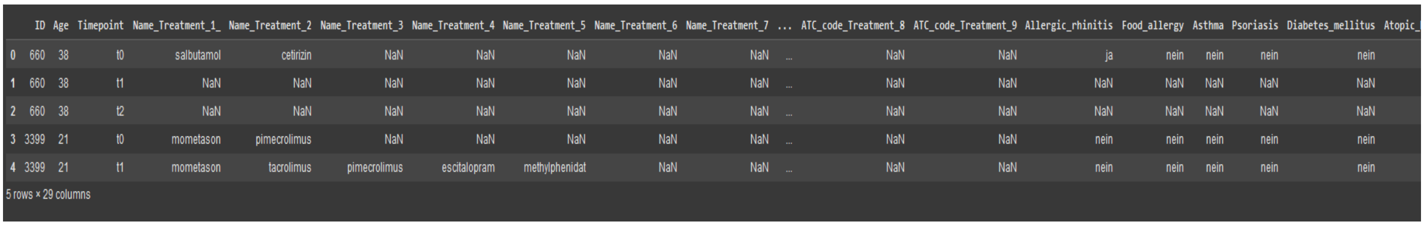
- **Number format:**

To avoid any mistakes, all numerical columns should be harmonized into the same number format. For example, the feature 't0_IGE_level' contains either FLOAT values or INTEGER values.

Question 1 – Longitudinal format.

- A feature engineering was achieved to build the new feature 'Age' :
  - For both features 'to_Birth_year' and 'to_Birth_month' missing values were filled and converted to INTEGER format.
  - Concatenation step was used to create the intermediary feature 'Birth_date'. Invalid dates were considerated as NaT.
  - A fixed reference date was used to calculate the difference between this reference and the 'Birth_date' previously calculated, corresponding then to the 'Age' of the patient. Missing or invalid ages were replaced with '-1' value for a better follow-up.
- Features such as 'Size' or 'Weight' were not available on df1. Consequently, the feature 'BMI' was impossible to calculate.
- To facilitate analysis of time-dependent data, a longitudinal format was generated :
  - Time dependant columns of interests (Treatments, Demographics, Score, Commodities) were selected and each variable's values were grouped under one column while a new a new column 'Timepoint' was created.
  - The 'Timepoint' column is filled by extraction of the timepoint indicator t0, t1 or t2.
  - The generated table is then merged iteratively with the common key ('ID') and with the new columns calculated ('Timepoint', 'Age' or 'Birth_date').
  - Lastly, the merged table is sorted by 'ID' and 'Timepoint' with a reset of index.

Here below are the first lines of the longitudinal table (Example with ID 660 and 3399):

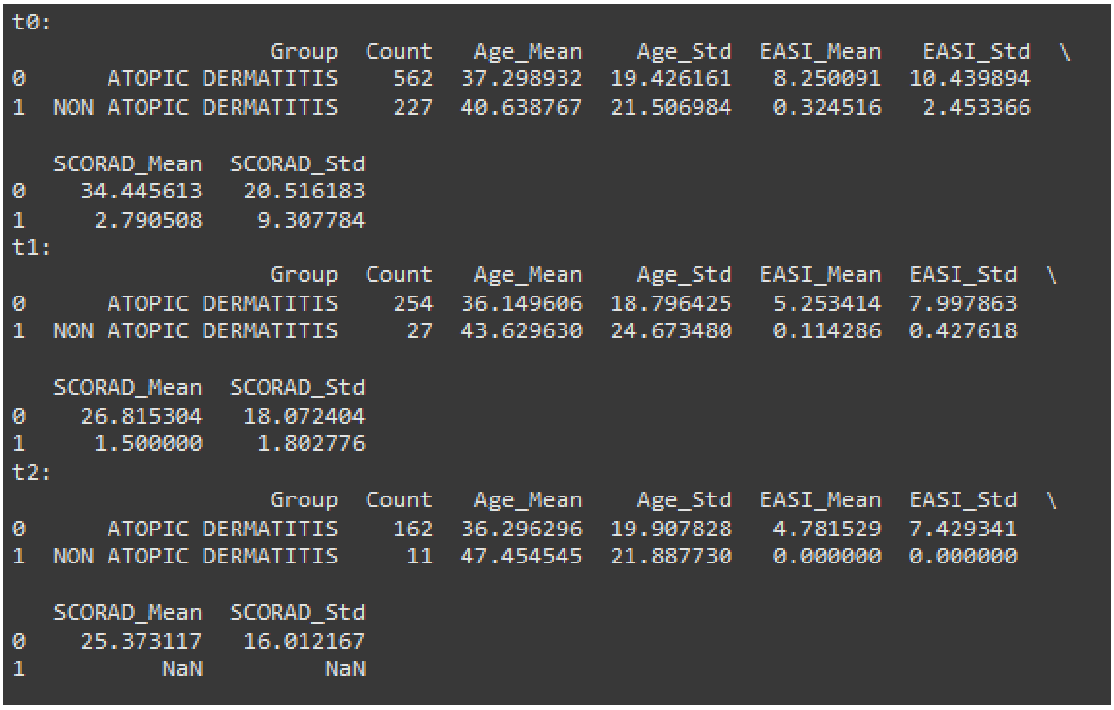


```
     ID  Age  Timepoint  Name_Treatment_1_  Name_Treatment_2  Name_Treatment_3  Name_Treatment_4  Name_Treatment_5  Name_Treatment_6  Name_Treatment_7  ...  ATC_code_Treatment_8  ATC_code_Treatment_9  Allergic_rhinitis  Food_allergy  Asthma  Psoriasis  Diabetes_mellitus  Atopic_
0   660   38         t0          salbutamol          cetirizin               NaN               NaN               NaN               NaN               NaN  ...                   NaN                   NaN                 ja          nein    nein       nein               nein
1   660   38         t1                 NaN                NaN               NaN               NaN               NaN               NaN               NaN  ...                   NaN                   NaN                NaN           NaN     NaN        NaN                NaN
2   660   38         t2                 NaN                NaN               NaN               NaN               NaN               NaN               NaN  ...                   NaN                   NaN                NaN           NaN     NaN        NaN                NaN
3  3399   21         t0           mometason       pimecrolimus               NaN               NaN               NaN               NaN               NaN  ...                   NaN                   NaN               nein          nein    nein       nein               nein
4  3399   21         t1           mometason         tacrolimus      pimecrolimus       escitalopram    methylphenidat               NaN               NaN  ...                   NaN                   NaN               nein          nein    nein       nein               nein
5 rows × 29 columns
```

Question 2 – General characteristics of subset datas.

Work for this question was divided into three parts:

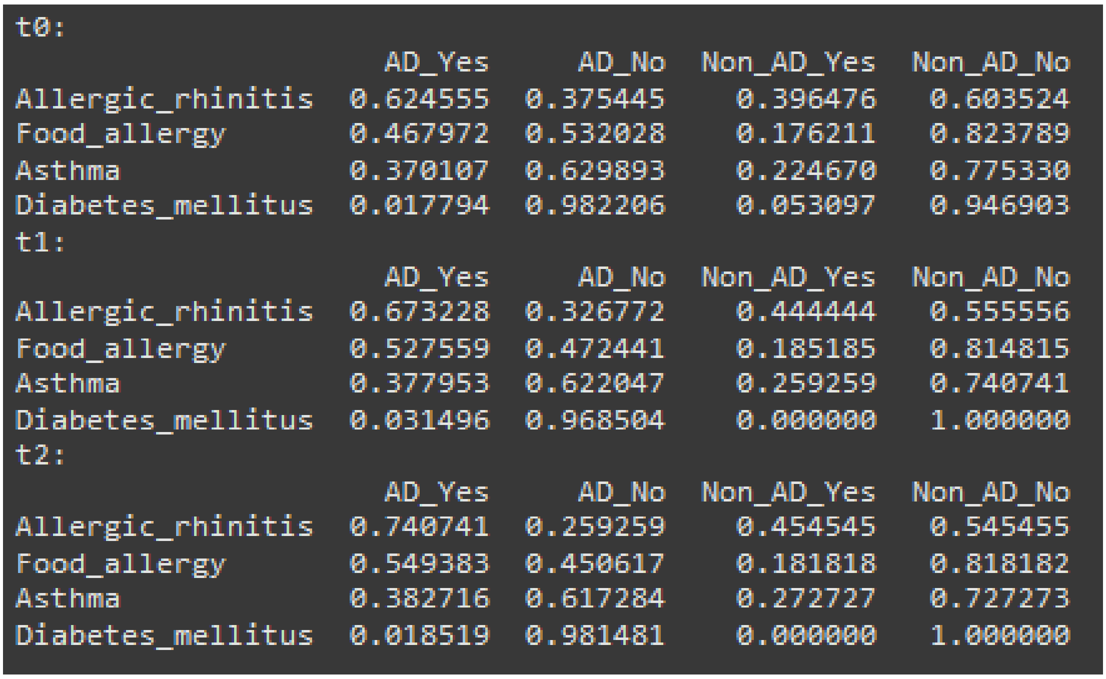
- For numerical features ('Age', 'EASI_Score' and 'SCORAD_Score') and at each three timepoint, mean and standard deviation were calculated for both cluster AD and non-AD. Results called "Patient characteristics" are showed below ("Summary" table on Python script) :



```
t0:
                   Group  Count   Age_Mean    Age_Std  EASI_Mean   EASI_Std  \
0      ATOPIC DERMATITIS    562  37.298932  19.426161   8.250091  10.439894
1  NON ATOPIC DERMATITIS    227  40.638767  21.506984   0.324516   2.453366

   SCORAD_Mean  SCORAD_Std
0    34.445613   20.516183
1     2.790508    9.307784
t1:
                   Group  Count   Age_Mean    Age_Std  EASI_Mean  EASI_Std  \
0      ATOPIC DERMATITIS    254  36.149606  18.796425   5.253414  7.997863
1  NON ATOPIC DERMATITIS     27  43.629630  24.673480   0.114286  0.427618

   SCORAD_Mean  SCORAD_Std
0    26.815304   18.072404
1     1.500000    1.802776
t2:
                   Group  Count   Age_Mean    Age_Std  EASI_Mean  EASI_Std  \
0      ATOPIC DERMATITIS    162  36.296296  19.907828   4.781529  7.429341
1  NON ATOPIC DERMATITIS     11  47.454545  21.887730   0.000000  0.000000

   SCORAD_Mean  SCORAD_Std
0    25.373117   16.012167
1          NaN         NaN
```

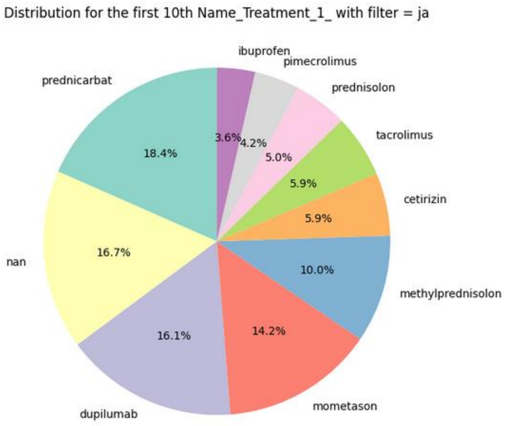
NB: Others basic statistics could be added to this table such as quartile, mode…

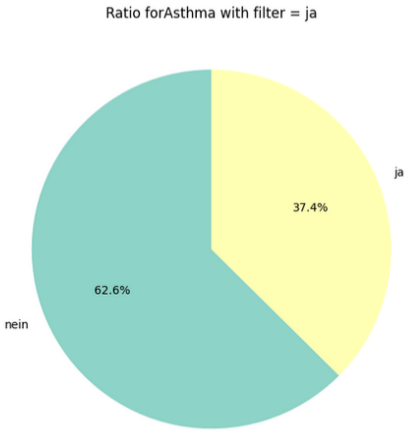
- For the 4 commodities features (‘Allergic_rhinitis’, ‘Food_allergy’, ‘Asthma’ and ‘Diabetes_mellitus’) and at each three timepoint, proportion of Presence/Absence of the disease were calculated for both cluster AD and non-AD. Results are showed below:



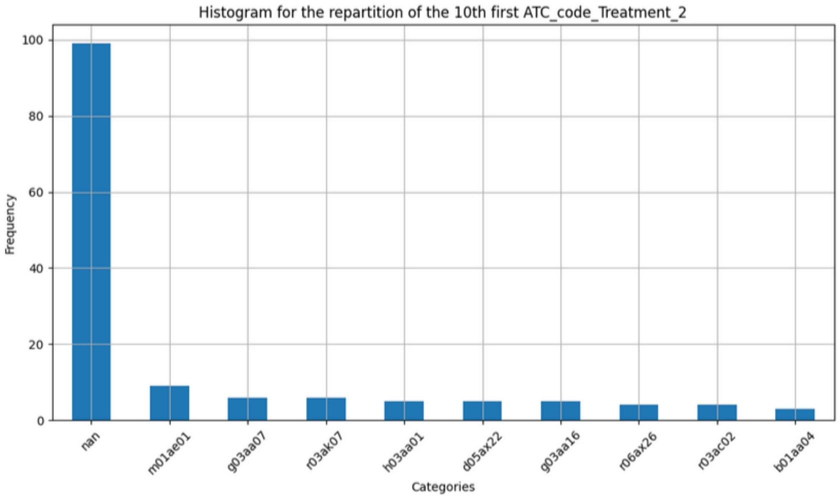
```
t0:
                     AD_Yes     AD_No  Non_AD_Yes  Non_AD_No
Allergic_rhinitis  0.624555  0.375445    0.396476   0.603524
Food_allergy       0.467972  0.532028    0.176211   0.823789
Asthma             0.370107  0.629893    0.224670   0.775330
Diabetes_mellitus  0.017794  0.982206    0.053097   0.946903
t1:
                     AD_Yes     AD_No  Non_AD_Yes  Non_AD_No
Allergic_rhinitis  0.673228  0.326772    0.444444   0.555556
Food_allergy       0.527559  0.472441    0.185185   0.814815
Asthma             0.377953  0.622047    0.259259   0.740741
Diabetes_mellitus  0.031496  0.968504    0.000000   1.000000
t2:
                     AD_Yes     AD_No  Non_AD_Yes  Non_AD_No
Allergic_rhinitis  0.740741  0.259259    0.454545   0.545455
Food_allergy       0.549383  0.450617    0.181818   0.818182
Asthma             0.382716  0.617284    0.272727   0.727273
Diabetes_mellitus  0.018519  0.981481    0.000000   1.000000
```

- Additionally, a data visualization work on Treatment and ATC values was achieved both for AD and non-AD. NB: In these graphs, we keep ‘NaN’ values which are very important in terms of meaning since they have to be interpreted as “No treatment” or “No information available”. Here below are some relevant Pie Chart or Histogram :





We observed, for example, that for the first line of treatment, prednicarbat or mometason or methylprednisolon (corticosteroids) are administrated to the patient in 50% of cases. Likewise, for patients with AD, 37 % of these patients suffer at the same time from Asthma. These results are consistent regarding the medical context (See introduction).



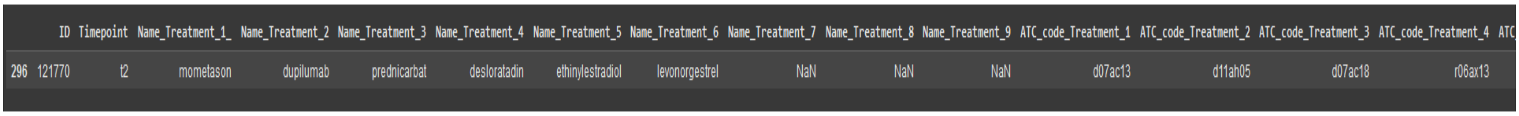
We observed, for example, that for the second line of treatment, most patients without AD didn’t receive any treatment and for some patient the first ATC code treatment given was ‘m01ae01’ (Ibuprofen) which is also consistent regarding the medical context.

Question 3 – Merging and specific filter/count.

- Sometimes for logistic or practical reason, some patient could be sampled in a given center and then clinically followed up in another center from the participating hospital network. The consequence is to have, for a same ‘ID’, two distinct lines due to the distinct center. Consequently, a pairs comparison for each (ID, CENTER) was achieved between df1 and df2 to check any eventual difference. No difference was observed which means that sample and clinical follow-up are achieved at the same place.

- Merging of both df1 and df2 was achieved on the join key ‘ID’. NB: The join type is not specified is this case (by default inner join). The final merged dataset contains 2400 lines corresponding to 3 Timepoint (t0, t1, t2) X 800 distinct ID patients.
- For the first subquestion 3.1 (EASI_Score > 25, Skin_biopsy sample available and patient without any Dupilumab treatment) :
  - At t0, 5 patients were identified with the IDs 139265, 353849, 411314, 922551 and 963943.
  - At t1, 1 patient was identified with the ID 649645.
  - At t2, 0 patient was identified.

- For the second subquestion 3.2 (SCORAD_Score < 50, Whole_blood sample available and patient under Dupilumab treatment) :
  - At t0, 3 patients were identified with the IDs 533988, 809919 and 814575.
  - At t1, 5 patients were identified with the IDs 77278, 173058, 174234, 337214 and 425021.
  - At t2, 1 patient was identified with the ID 121770. Here below is an example for this last query.



	ID	Timepoint	Name_Treatment_1_	Name_Treatment_2	Name_Treatment_3	Name_Treatment_4	Name_Treatment_5	Name_Treatment_6	Name_Treatment_7	Name_Treatment_8	Name_Treatment_9	ATC_code_Treatment_1	ATC_code_Treatment_2	ATC_code_Treatment_3	ATC_code_Treatment_4	ATC
296	121770	t2	mometason	dupilumab	prednicarbat	desloratadin	ethinylestradiol	levonorgestrel	NaN	NaN	NaN	d07ac13	d11ah05	d07ac18	r06ax13	

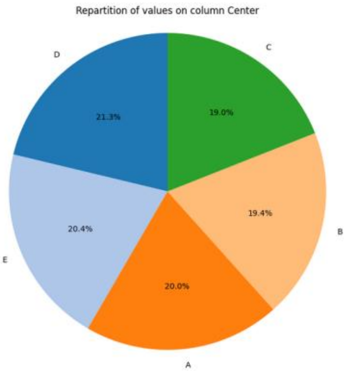
Conclusion:

A particular emphasis was placed on data quality control for this exercise particularly with df1 dataset which contains inconsistent, missing and outliers values. Furthermore, some medical precision should be given by a medical community to validate or not biological interpretation of some data.

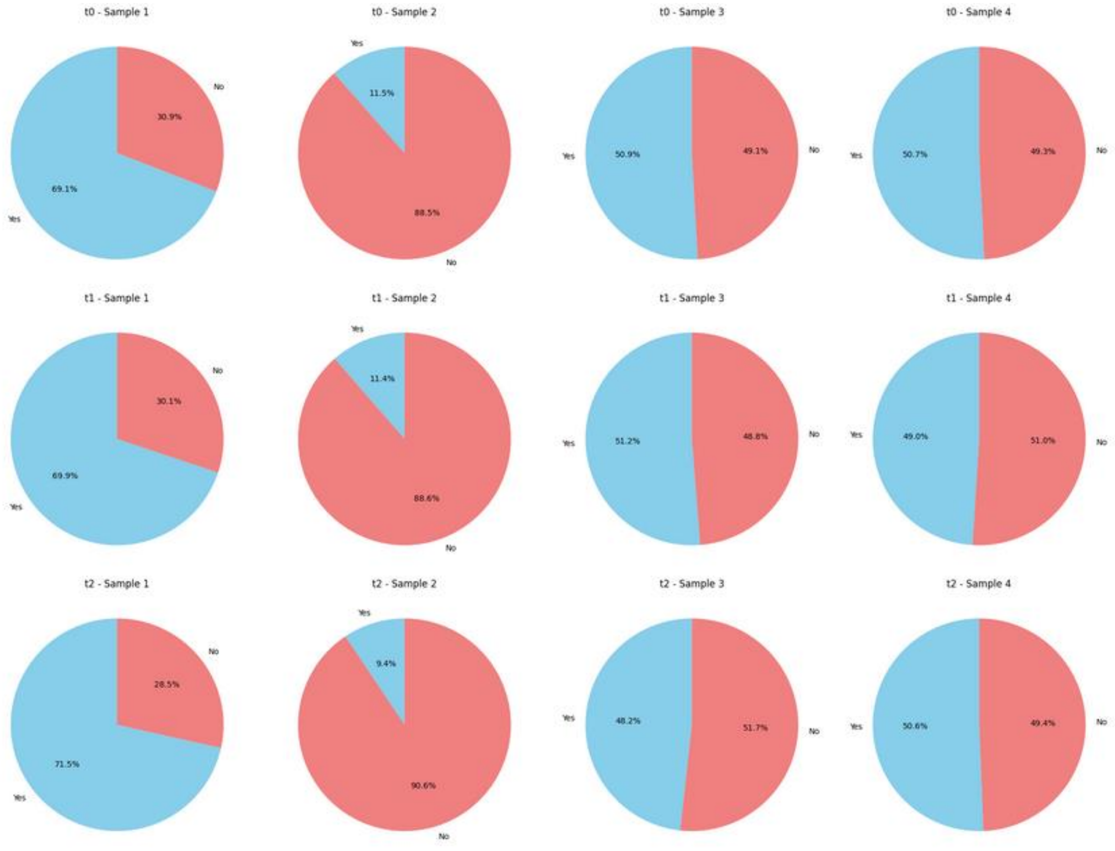
Data were transformed into a “Longitudinal format” regarding three timepoint of the study (t0, t1, t2) to characterize in particular two groups of patients with ATOPIC DERMATITIS or NON ATOPIC DERMATITIS. Lastly, some patients of interest were identified corresponding to 3 specific inclusion criteria for each queries.

Regarding subquestion 3.2, machine or deep learning method could be applied on microscopy photos taken from available whole blood sample in order to classify automatically immune cell, leucocyte for example, into predefined category for standardized diagnosis associated to clinical data.

Additional work – Data visualization:



Regarding this previous graph, we observed that all patients are equally distributed between the 5 centers participating in this longitudinal study.



This second graph shows that Whole_blood and Serum samples are collected in half of the cases whatever the timepoint. What's more, Skin_biopsy are achieved only for 10% for all patients whatever the timepoint. Lastly, Swabs sample are collected in almost three quarters of cases. These results are useful for sample management and future research work regarding cell culture in a context of personalized medicine for example.