# General Information

The list of chosen columns; It can be changed inside the program.

|  |  |
| --- | --- |
| Sheet | Selected Columns |
| OutputData | E, J, Q, T, V, W, X, DX, [OS-SV] |
| AllVisit | E, H, I, J, L, N, O, P, Q, [R-V], AA, AB, [AC-AF], AH, [CQ-EQ], [LZ-MG], [MT-MX], [MY-NH] |
| DiagnosisData | D, E |
| MedicationData | D, G, J, Q |

# Combining OutputData and AllVisit

* **First\_Biologic (StartDate)** < **AV\_Visit\_Date**
  + All samples of this patient in **AllVisist** sheet will be ignored. Patient with PatientID=9 is an example of this case, as shown below:

|  |  |  |
| --- | --- | --- |
| First\_Biologic(StartDate) (or DX) | AV\_Visit\_Date (or H) | Note |
| 01/09/2009 | 19/01/2021 | Since the Biological treatment date is before all other date, all rows from **AllVisist** will be ignored. Meaning for PatientID=9 columns of **AllVisist sheet** was NaN. |
| 09/04/2021 |
| 20/01/2022 |
| 14/04/2022 |
| 28/07/2022 |
| 27/10/2022 |
| 19/04/2023 |

* **First\_Biologic (StartDate)** Does not exist.
  + It returns the row with the latest date for this patient in **AllVisist** sheet. Patient with PatientID=11 is an example of this case, as shown below:

|  |  |  |
| --- | --- | --- |
| First\_Biologic(StartDate) (or DX) | AV\_Visit\_Date (or H) | Note |
| Not exist. | 13/01/2021 | Since there is no biological treatment date, last row (18/08/2023) was chosen. |
| 18/08/2023 |

* **Remaining cases**
  + Handled as mentioned in our meeting. Patient with PatientID=4 is an example of this case, as shown below:

|  |  |  |
| --- | --- | --- |
| First\_Biologic(StartDate) (or DX) | AV\_Visit\_Date (or H) | Note |
| 24/09/2020 | 14/08/2019 | The columns of row with H=**17/08/2020** waschosen. |
| **17/08/2020** |
| 24/09/2020 |
| 17/12/2020 |
| 10/03/2021 |
| 02/06/2021 |
| 23/03/2022 |
| 15/06/2022 |
| 07/09/2022 |

# Combining OutputData and DiagnosisData

The new column named **hasRheumatoidArthritis** was added to **OutputData sheet** with value [0-3]

* 571 patients do not have Rheumatoid Arthritis. Their values in **hasRheumatoidArthritis** are **zero**.
* 505 patients have one row in DiagnosisData. Their values in **hasRheumatoidArthritis** are **one**.
* 6 patients have two rows in DiagnosisData. Their values in **hasRheumatoidArthritis** are **two**.



* 1 patient has three rows in DiagnosisData. Its value in **hasRheumatoidArthritis** is **three**.



# Combining OutputData and MedicationData

Newly added columns

|  |  |
| --- | --- |
| Column Name | Descriptions |
| Multi | There are same cases when patient took the medication in multiple periods. For example, between 2019 to 2020; then stop and started again between 2021 to 2022. This column how for how many times the medication was taken. The default value is 1.   * **Multi == 0** 🡪 Have never taken this medication * **Multi == 1** 🡪 Take this medication in one period * **Multi > 1** 🡪 Take this medication in multiple period |
| Period | For how long the patient took the medicant. It is in day format; for example, 30 days. There are some special cases:   * If **Multi > 1** 🡪 Sums all periods. * If **isActive == True and** **Multi == 1** 🡪 the value is **-1** * If **isActive == True and** **Multi > 1** 🡪 Sums the period of previous sessions |
| isActive | Is a patient still use this medication? **True** for yes, otherwise return **False** |

The final output for column names of this sheet will be (**MedicamentGroupName** + **ColumnName**) As shown below:

A white rectangular object with black text

Description automatically generated

# First Update Points

|  |  |
| --- | --- |
| Point | Update |
| In the data extracted from OutputData you should have column "U" instead of "V" | DONE (updated the code) |
| Concerning 'Combining OutputData and DiagnosisData" of read.me file - if these 6 patients who have two records in "DiagnosisData" have "Rheumatoid Arthritis" in column E of "DiagnosisData", you may assign to them hasRheumatoidArthritis=1. The same holds for the patient witt ID=590 (three extries in "DiagnosisData" and three times "Rheumatoid Arthritis" in column E) | **Previously**, we had 0,1,2 and 3 in the hasRheumatoidArthritis column based on how many times the value appeared in DiagnosisData. **Now**, there are only two values: **one** if patients have Rheumatoid Arthritis and **zero** if they don't. Regardless of how many times it appears in DiagnosisData. (I understood your point in this way; please correct me if I'm wrong) |
| **First Part**: Concerning "Combining OutputData and MedicationData" of read.me file - like in the case of AllVisit sheet we should use only records earlier than than the date in column DX of OutputData.  **Second Part**: in the case DX is empty we may use all medication data. | **First Part**: I did that if there are no records and the column values will be NaN.  **Second Part**: I only take the latest sample if there are more than one row.  Please let me know if any actions are needed on this point. |
| I do not understand why you have applied the condition "If isActive == True and Multi == 1 => the value is -1" | I treated -1 as a symbol of uncertainty (The value that we don't know), not like a number. In the case of **isActive == True** and **Multi == 1**, it means that a patient is still using the treatment and it's their first time to use this treatment. For example, If the patient started using the treatment on 2023-10-01 and the data collection date on 2023-11-06 (I think this column is named **ExtractionDate** in the Excel file), the patient may stop using this treatment on 2023-11-07, 2024-01-01, etc. (infinite values) One option is using **ExtractionDate**column, for previous example the value will be 37; please let me know how to handle this case.  In the case of**isActive == True** and **Multi > 1**, I treated the current treatment session the same way as above. For example, if **Multi =2,**thefirst period is between 2022-01-01 and 2022-06-01, and the second period started on 2023-10-01 and is still active, the value of the**period column** will be six months. (The second period has been ignored) |
| **First Part**: I also think that it can be important if the patient stopped using some drug e.g. 10 days or 10 months before starting biological treatment. So there should be a column like "How long ago" for each of {DMARD, NSAID/COXIB, OTHER/SUPPLEMENTS}.  **Second Part**: I do not how this variable should be computed for patients which do not started biological treatment, maybe (-1) is an option. | **First Part**: DONE. **treatment\_StopDate**= (**biological\_start\_date** – **other\_treatment\_end\_date**) (value in days format)  **Note 1**: If (**Multi**>1) I only calculate **treatment\_StopDate** for the latest session  **Note 2**: There are cases that the treatment does not stop after starting biological treatment, I ignore those cases. PatientID=3  **Second Part**: for now, they are **NaN** |
| **First Part**: Could you, please, write to me how many patients with rheumathoid arthritic do we have in total. (**512 patients**)  **Second Part**: (and how many with biological treatment among them)? (**215** **patients**)  **Third Part**: Could you, please, summarize how big is the problem of missing data? (**Yes, quite big. the detail in Notebook**) | The code answers to those questions are in the last section of NoteBook. The section called **Questions**. After we finalize the structure, we will show all statistics with the graph and more details.  **Third Part**: Please let me know if you want to see any statistics in more detail |