Clinical Trial Prediction Report

Done by: Dinmukhamed Umbetzhan

Introduction

All Clinical Trials information was stored in the xml format and consists of more than 100.000 files. Most of the files are structured in the similar way, however some of them differ significantly. The sizes of each file might differ significantly as well, which complicates the extraction of the data stored in them.

Each xml file uses specific tags to store information related to the given tag name. For instance, Figure 1 illustrates an example of these tags. The information regarding the phase of the trial can be found between “<phase>” (opening tag) and “</phase>” (closing tag). The structure of those tag system is the same as in the html. And same as in the html there might be a nested tags, for example, the information regarding the intervention model is stored inside ”<study_design_info>“ tag.

```
<start_date>February 2000</start_date>
<completion_date type="Actual">March 2015</completion_date>
<primary_completion_date type="Actual">June 2005</primary_completion_date>
<phase>Phase 3</phase>
<study_type>Interventional</study_type>
<has_expanded_access>No</has_expanded_access>
<study_design_info>
  <allocation>Randomized</allocation>
  <intervention_model>Parallel Assignment</intervention_model>
  <primary_purpose>Treatment</primary_purpose>
  <masking>None (Open Label)</masking>
</study_design_info>
```

Figure 1. *An example of Clinical Trial Report*

Process of data collection

The first step was to choose a programming language and a convenient development environment (IDE). My choice was Python and the Pycharm development environment, the reason for that was the simplicity of Python when working with a large amount of data and it is also quite easy to extract files in Python.

My code can be found here https://github.com/JustForFun29/MQP_Clinical

Discovering patterns inside XML files

Before extracting the data, it is needed to understand – what kind of data is needed? To answer this question, I used an example of the data that was used by a graduate student, unfortunately, it was not enough. So, after many iterations I came to conclusion that the data that is needed consists of the following:

1. nct_id – unique id also called the ClinicalTrials.gov identifier.
2. org_study_id – unique identification code given to each study.
3. brief_title – short version of title for each study.
4. official_title - official version of title for each study.
5. overall_status – status of the study when it was submitted.
6. study_type – interventional/observational.
7. minimum_age – minimum age of participants of the study.
8. maximum_age – maximum age of participants of the study.
9. gender – gender of participants of the study.
10. criteria – criteria for participants (including the exclusion criteria).
11. intervention_model – model on which a particular intervention was tested.
12. primary_purpose – treatment/prevention.
13. masking – single/double/triple/quadruple.
14. source – the source of the study.
15. phase – the phase of the study.
16. start_date – start date of the study.
17. brief_summary – straightforward.
18. detailed_description – straightforward.

All the information mentioned above will be stored in the main table. But having these fields is not enough since I'm still missing the drugs and diseases fields. I decided to use separate tables for them. The reason is that if there are some studies that use the same drugs or cure the same diseases it will be easier to "observe" if they are the same if they somehow had specific ids. The fields in each table will only consist of id and drug/disease names.

Creating a SQL database tables

As mentioned in previous section there will be following tables: main table, drug table and disease table. And to show the relationship between specific drug and specific study I created relationship table to show it, and same for the diseases. The diagram of each table can be seen in Figure 2.

For instance, let's take a real example in first xml file we've got "NCT00000102" that uses drug called "Nifedipine" and treats "congenital adrenal hyperplasia". So, to show the relationship between all of them first I would add the drug to the drug table, then I would add disease to the disease table and only after that I would create relationship between main table field by using id (which is actually an nct-id) entry and drug table field by using id (which will be 1 since it is the first drug that would be added to the table) and some for the disease entry.

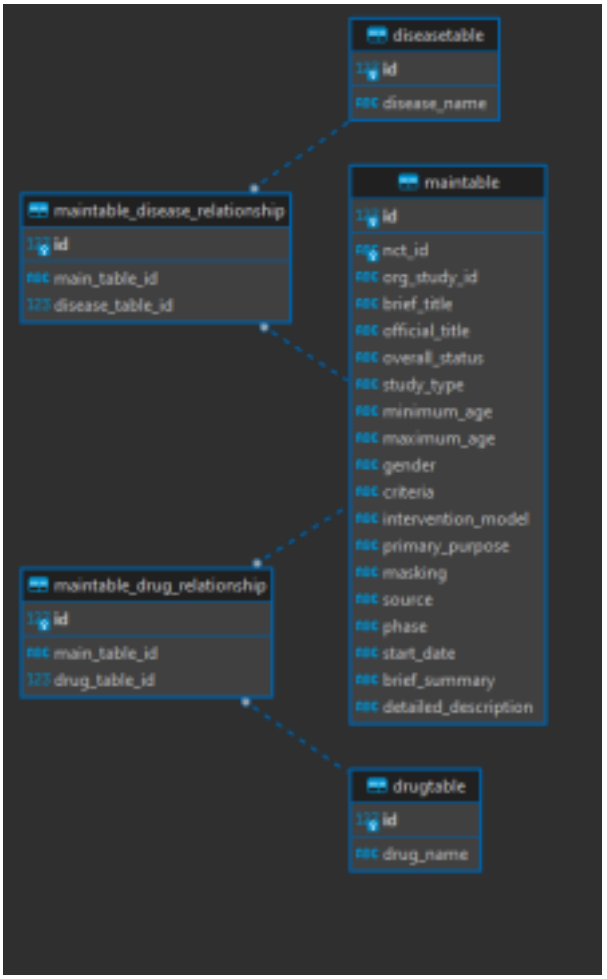

Figure 2. *The diagram of SQL tables*

The reason why I chose this path is simple – it is just much more convenient. By using only drug id in the maintable_drug_relationship I can retrieve all studies where this specific drug was used and same for the disease entries.

Extracting the data from XML files

The hardest part of this project was the extraction of data since as mentioned in the introduction section all xml files differ significantly. Some of the files might not have any information – even including the source of the study. Sometimes study can have multiple disease entries and sometimes they can have none of them and all of the cases must be handled. To retrieve the data from the XML file I used the strategy of nested dictionaries by using the library xmltodict. And by accessing this dictionaries it is much easier to retrieve nested tags as was shown in the Introduction section.

Firstly I decided to extract the drugs and diseases separately from the rest of the study information to be ensured that each value is unique. And the first problem that I encountered is that in some studies the disease and drug names might by capitalized and sometimes they are lower case. To resolve this issue, I decided to make all of them lower case.

Then I retrieved all information that was populated in the main table. And only after that I tried

to apply the relationships between each entry of the tables.

Results

As a result, I've got:

- 102,865 unique studies in the main table.
- 34,580 unique drugs in the drug table.
- 25,569 unique diseases in the disease table
- 56,650 unique relationships in the maintable_drug_relationship table •
  170,931 unique relationships in the maintable_disease_relationship table

The reasons for such great difference between the relationship tables (drugs relationships with main table is much higher) is because many reports doesn't use drugs (for instance, some of them uses procedures) and the second reason is that drug names not always included in the <drug_name> tag, but rather in the text fields and it makes retrieving such information more challenging.