Constructing a Knowledge Graph from Clinical Trials data

Timo Anthonijsz - i6333093

March 2024

Abstract

Clinical trials play a pivotal role in advancements within medical knowledge and treatment methods yet they have high costs and a low success rate. In this project, we propose a modest start for creating knowledge graphs from clinical trials data. Leveraging data obtained from clinicaltrials.gov and processing this data using various techniques, an attempt has been made to answer a few research questions. From the results, we can conclude that the knowledge provides a good starting point. However, some limitations are to be found in the metadata, some solvable with a lot of NLP, others not within our grasp.

1 Significance

Clinical trials can offer critical insight into advancing new treatment methods and the development of new medicine. As well as playing a key role in discovering the safety and risks of these new drugs and treatments for human subjects. Typically, before any drug can be introduced for public use, three phases of these clinical trials need to be completed. However, clinical trials are very expensive, increasing in cost through every phase. In addition to the cost, the success rate of clinical trials is very low. Sertkaya et al. (2016) reported that the cost of a single trial phase ranges from 1.4 million up to 52.9 million US dollars. Additionally, as reported by Wong et al. (2019), clinical trials in oncology only have a success rate of 3.4%. The combination of a high cost and a low success rate calls for obtaining increased knowledge on existing clinical trials. By analyzing previous trials and inferring information from them, useful insight might be obtained, such that future trials can be designed in a more sophisticated manner. In the last decade, there has been an incredible surge in the size and variety of data we have, increasing the significance of storing this data in a more homogeneous manner.

With that in mind, this project proposes a knowledge graph for clinical trials to try improve the analysis and the ability to infer information from existing clinical trials.

2 Related work

In 1986 DerSimonian & Laird (1986) proposed a meta analysis method in which a random effects model was used to "characterize the distribution of treatment effects in a series of studies." This method has been widely used ever since publication, and is commonly referred to as the "DerSimonian and Laird method". However, in the last decade, we have had an incredible surge in the amount of data. DerSimonian & Laird (2015) addresses this increase in data in a revision of their original paper from 1986. They discuss some more recent challenges. For instance, with this big data, analysis tends to be done on data summaries instead of combined big datasets, which questions the completeness and representativeness of those summaries. Additionally, data has become more heterogeneous, which has made it increasingly difficult to infer associations with a high degree of certainty.

In the context of lacking structure to properly infer knowledge from clinical trials, there have been recent developments. Lehman et al. (2019) proposed some heuristic and machine learning methods to infer whether a medical treatment works based on information obtained in clinical trials. Although neural networks have proven to have potential they still lack proficient attentional methods to infer relevant evidence. The attention problem, however, might be decreasing with the developments in the transformer models. A more similar approach to inferring data from clinical trials was proposed byChen et al. (2022). They have constructed a knowledge graph for clinical trials obtained from *https://clinicaltrials.gov/* (n.d.), called CTKG. In this knowledge graph, nodes are represented as medical entities (e.g., studies, drugs and conditions), and edges represent the relations between these entities (e.g., drugs used in studies). Resulting in a graph that can be used for applications like drug repurposing and similarity. The goal of this project is similar in the sense that it will aim to find similarities in the clinical trials as well. On the other hand, a secondary goal of this project is to facilitate searching for patterns within clinical trials.

3 Goals & Objectives

The main goal of this project is to develop an knowledge graph to represent clinical trials. The purpose of this graph is to enable clinical trial analysis with the help of queries. The graph is meant to facilitate researchers in finding interesting patterns in clinical trials, for instance, patient demographics or treatment effectiveness. The project's overarching goal is for researchers to infer knowledge from this graph, so that future clinical trials can be designed more efficiently, leading to reduced costs and hopefully a higher success ratio.

A. Validating the graph

- **What are the number of clinical trials connected to the condition "heart failure"** For this question, a query was used only filtering on the condition "heart failure" and counting this amount. This result can be checked using the site *https://clinicaltrials.gov/* (n.d.)
- **What are the amount of clinical trials held in the United States** For this question, I had to use a query extracting in which we filter on whether the location string contains United States. This result can be checked using the site *https://clinicaltrials.gov/* (n.d.)
- **What are the amount of clinical trials in which the intervention method uses ibuprofen** For this question I used a query that filtered the intervention method on DRUG: Ibuprofen. This result can be checked using the site *https://clinicaltrials.gov/* (n.d.)

B. Gaining insights

- **What are the most common conditions studied across all clinical trials?**
- **Which sponsors are involved in the highest number of clinical trials?**
- **What is the average study duration up until the primary completion date and what is the average time between the primary completion date and completion date?**

4 Methodology

4.1 Data

As mentioned before, the clinical trial data could be obtained from *https://clinicaltrials.gov/* (n.d.). I chose to obtain the data as *CSV* file, the other option was as *JSON* file. This was done leveraging the *Clinicaltrials.gov API.* The data includes the main information about a clinical trial. This data was stored as *CSV* file with 485574 rows representing the entries

4.2 Methodology

To complete this project, various techniques were employed. The process was divided in different parts, which are mentioned in further detail below.

A. **Data extraction**. Firstly, the clinical trials data had to be obtained from their API. This was done per 1000 files, as that was the allowed maximum. All the files have been merged into one big *CSV* file of approximately 1.5 GB.

B. **Data exploration, vocabulary development**. Secondly, the data was explored, and afterwards a vocabulary was created

C. **Knowledge Graph creation** Create the Knowledge Graph using RDFLib.

D. **Querying the Knowledge Graph** When the knowledge graph was created, it could be explored to try and find some interesting patterns. The data was loaded into GraphDB for querying purposes.

4.2.1 A: Data extraction

The first objective was to obtain the data, as mentioned, this was done using *Clinicaltrials.gov API.* This data was then split into a smaller chunk of 10000 rows first for testing purposes, as 485574 rows seemed like a lot to test everything on.

4.2.2 B: Exploration & Vocabulary

In this step, the vocabulary had to be constructed to structure the data for the Knowledge Graph. There were 30 columns in the data. Some of which had multiple entries. These were dealt with by splitting for '|' or ',' or ':'. The main structure is of the following form: Every trial is recognizable through its unique NCT code and is typed as ClinicalTrial. All the basic information is directly linked to this. There are 4 subgroups, all linked with the main trial: Participant Information, Medical Details, Collaboration Details, Location Information. These subgroups are typed. All of these groups have their own predicates linked to them. The full overview of the vocabulary can be found in appendix A.

4.2.3 C: Constructing the Knowledge Graph

After creating the vocabulary, we could start creating a turtle file containing my Knowledge Graph. This was done using RDFLib, to create a ttl file from a dataframe created with the *CSV* data file. This however ended up giving a memory error, so the data was split into 4 parts and afterwards combined and serialized. It finally took approximately 5.5 hours to serialize and parse all the data into a turtle file.

4.2.4 D: Querying the Knowledge Graph

The last step was to query the Knowledge Graph to answer some research questions. This was done by constructing queries in Python. A first thought was to load the data in a database like graphDB, but the size (+-2.5 GB) of my graph forced me to pay on most user friendly graph databases. Therefore, I just employed Python.

5 Results

In this section the results of the Python queries will be shown. Afterwards they can be discussed.

5.1 Validating the graph

What are the number of clinical trials connected to the condition "heart failure? To answer this question, the following query was constructed:

```
PREFIX ct: <http://example.org/clinicalTrial/>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>

SELECT (COUNT(DISTINCT ?trial) AS ?numTrials)
WHERE {
  ?trial rdf:type ct:ClinicalTrial ;
         ct:hasMedicalDetails ?medDetails .
  ?medDetails ct:conditions ?condition .
  FILTER(CONTAINS (?condition, "Heart Failure"))
}
```

Listing 1: SPARQL Query to extract studies which have Heart failure as condition

This query resulted in returning 5227 trials with Heart Failure. When checking with *https://clinicaltrials.gov/* (n.d.) there are 6461 trials to be found. The difference can be accounted for after some manual search. For instance, studies linked to "Heart Failure" on their website could have the following conditions: "Congestive Cardiac Failure", "Cardiac Insufficiency", among others. These differences are hard to account for using just a query.

What are the amount of clinical trials held in the United States? To answer this question, the following query was constructed:

```
PREFIX ct: <http://example.org/clinicalTrial/>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>

SELECT (COUNT(DISTINCT ?trial) AS ?numTrials)
WHERE {
  ?trial rdf:type ct:ClinicalTrial ;
         ct:hasLocationInfo ?locationInfo .
  ?locationInfo ct:locations ?location .
  FILTER(CONTAINS(?location, "United States"))
}
```

Listing 2: SPARQL Query to extract studies performed at an location in the United States

This query returned 170559 trials, which were held in the United States. When checking with *https://clinicaltrials.gov/* (n.d.) there are 171.286 trials to be found. This difference is less than a percent, which still seems very good. However there are probably some entries that only mention the state in the United States or mention only a facility there.

What are the amount of clinical trials in which the intervention method uses ibuprofen? To answer this question, the following query was constructed:

```
PREFIX ct: <http://example.org/clinicalTrial/>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>

SELECT (COUNT(DISTINCT ?trial) AS ?numTrials)
WHERE {
  ?trial rdf:type ct:ClinicalTrial ;
         ct:hasMedicalDetails ?medDetails .
  ?medDetails ct:interventions ?intervention .
  FILTER(CONTAINS(lcase(?intervention), "
    ibuprofen"))
}
```

Listing 3: SPARQL Query to extract studies using Ibuprofen as intervention

This returns 630 instances. When checking with *https://clinicaltrials.gov/* (n.d.) there are however 934 studies to be found. Similarly to the first query, we have the issue that there are a lot of similar words representing some form of Ibuprofen, leading to this lower number of instances returned.

5.2 Gaining Insight

What are the most common conditions studied across all clinical trials? To answer this question, the following query was employed:

```
PREFIX ct: <http://example.org/clinicalTrial/>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>

SELECT ?condition (COUNT(?trial) AS ?numTrials)
WHERE {
  ?trial rdf:type ct:ClinicalTrial ;
         ct:hasMedicalDetails ?medDetails .
  ?medDetails ct:conditions ?condition .
}
GROUP BY ?condition
ORDER BY DESC(?numTrials)
LIMIT 10
}
```

Listing 4: SPARQL Query to extract most common conditions in a clinical trial

The results can be found in Table 1. Healthy is actually the first of all conditions, which does make sense, as preventing a disease is infinitely better than getting it. This result did raise another question in me. Obesity has been an increasing problem, and I was wondering if that could be seen through the amount of trials on Obesity per year over the years. Therefore, this question is extended, and we employ the following query:

```
PREFIX ct: <http://example.org/clinicalTrial/>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>

SELECT ?trial ?startDate
WHERE {
  ?trial rdf:type ct:ClinicalTrial ;
         ct:hasMedicalDetails ?medDetails ;
         ct:startDate ?startDate .
  ?medDetails ct:conditions ?condition .
  FILTER(CONTAINS(?condition, ”Obesity”))
}
ORDER BY ?startDate
}
```

Listing 5: SPARQL Query to extract years in which obesity trials were held

Condition	Number of Trials
Healthy	9759
Breast Cancer	7523
Obesity	6585
Stroke	4099
Hypertension	4035
Depression	3932
Prostate Cancer	3847
Pain	3798
HIV Infections	3760
Asthma	3381

Table 1: Number of Trials by Condition

The results can be seen in Figure 2. There is a definite increase in clinical trials regarding obesity, this however is not enough to prove it is not just because of the increase in the amount of trials. Therefore, as some supplementary evidence Figure 3. This shows the increase in trials over the years. Comparing the shapes over both graphs makes it seem like there are a couple of years in which obesity becomes more apparent as a condition. This was checked by going over the percentage of obesity trials in the total amount of trials. What is found is a 1 percent increase in the total amount of studies from the early 2000's till the covid period.

Which sponsors are involved in the highest number of clinical trials? To answer this question, the following query was employed:

```
PREFIX ct: <http://example.org/clinicalTrial/>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>

SELECT ?sponsor (COUNT(?trial) AS ?numTrials)
WHERE {
    ?trial rdf:type ct:ClinicalTrial ;
              ct:hasCollaborationDetails ?
    collabdetails .
    ?collabdetails ct:sponsor ?sponsor .
}
GROUP BY ?sponsor
ORDER BY DESC(?numTrials)
LIMIT 10
}
```

Listing 6: SPARQL Query to extract most apparent sponsors of clinical trials

The results can be found in Table 2. For me, it was mainly interesting to see that specifically 2 Egyptian universities are top contributors as that is something I would not have expected

Sponsor	Number of Trials
GlaxoSmithKline	3482
National Cancer Institute (NCI)	3411
Assiut University	3373
Cairo University	3057
Pfizer	3053
AstraZeneca	3031
Assistance Publique - Hôpitaux de Paris	2961
Mayo Clinic	2777
M.D. Anderson Cancer Center	2710
Novartis Pharmaceuticals	2398

Table 2: Number of Trials by Sponsor

What is the average study duration up until the primary completion date and what is the average time between the primary completion date and completion date? To answer this question, the following query was employed

```
PREFIX ct: <http://example.org/clinicalTrial/>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>

SELECT ?trial ?startDate ?primaryCompletionDate ?
    completionDate
WHERE {
  ?trial rdf:type ct:ClinicalTrial ;
         ct:startDate ?startDate ;
         ct:primaryCompletionDate ?
    primaryCompletionDate ;
         ct:completionDate ?completionDate .
}
```

Listing 7: SPARQL Query to extract starting and completion dates

From this query, 3 different dates were obtained to find the average duration. The data was converted to an amount of days using the datetime package. To find the first average, we do $primarycompletiondata - startdate$. For the second one we do $completiondate - primarycompletiondata$. This resulted in the following durations.

- Average duration of clinical trials until primary completion: 921.1285448853705 days

- Average duration of clinical trials from primary completion till completion: 166.72414066386935 days

So approximately 2,5 years before the primary outcome measure is satisfied, and another half year before a trial is completely done on average.

6 Discussion

From the result section, we may conclude that the knowledge graph is quite an okay starting point. There are difficulties with the metadata. We found these during validation of the graph. This can mainly be accounted for by a combination of inconsistent entries and synonyms for things like conditions or different columns in which the entries are even more open, like Outcome measures. This could partially be solved employing a lot of NLP to convert the data to a more similar standard, but definitely will not fix all the issues.

Due to my lack of knowledge in the medical domain, the understanding I have of all the results is still lacking. Combining the knowledge with someone of the medical domain might lead to more interesting insights.

If one were to continue this work, it would definitely be worth investigating the JSON version of the data, as this one should have more different columns for a more comprehensive graph. More complex queries could be employed to find different and more complex insights. Finally, linking the graph to existing ontologies might be a worthwhile step, such that it can be reused.

Moreover, I really wanted to load my graph into a graph database, but as mentioned, due to the size of the graph I created a paywall for myself.

Lastly, *https://clinicaltrials.gov/* (n.d.) has definitely already done quite a good job with their own pretty advanced search engine.

7 Conclusion

This project provided a starting point for creating a KG of clinical trial data. Approximately half a million clinical trials have been processed for this graph. Due to lack of expertise, differences in data, resources, the analysis of the clinical trial data is still limited. For the scope of this project, it might have been better to use a slightly smaller dataset, such that free database programs are available. Overall. a fine starting point and a good learning process

References

Chen, Z., Peng, B., Ioannidis, V. N., Li, M., Karypis, G., & Ning, X. (2022, 3). A knowledge graph of clinical trials ($$\mathop {\mathtt {CTKG}}\limits$$). *Scientific Reports*, *12*(1), 4724. doi: 10.1038/s41598-022-08454-z

DerSimonian, R., & Laird, N. (1986, 9). Meta-analysis in clinical trials. *Controlled Clinical Trials*, *7*(3), 177–188. doi: 10.1016/0197-2456(86)90046-2

DerSimonian, R., & Laird, N. (2015, 11). Meta-analysis in clinical trials revisited. *Contemporary Clinical Trials*, *45*, 139–145. doi: 10.1016/j.cct.2015.09.002

https://clinicaltrials.gov/. (n.d.).

Lehman, E., DeYoung, J., Barzilay, R., & Wallace, B. C. (2019, 4). Inferring Which Medical Treatments Work from Reports of Clinical Trials.

Sertkaya, A., Wong, H.-H., Jessup, A., & Beleche, T. (2016, 4). Key cost drivers of pharmaceutical clinical trials in the United States. *Clinical Trials*, *13*(2), 117–126. doi: 10.1177/1740774515625964

Wong, C. H., Siah, K. W., & Lo, A. W. (2019, 4). Estimation of clinical trial success rates and related parameters. *Biostatistics*, *20*(2), 273–286. doi: 10.1093/biostatistics/kxx069

A Code

The notebooks used to create everything in this project can be found on `https://github.com/Timszy/KG-clinical-trials`

B GPT usage

ChatGPT was used during this project, mainly during the coding part of this project. During EDA, it was of help with the code to identify the structure of the data and create ways to split the data. During graph creation, GPT was used to speed up the process after coming up with the hierarchy myself. It was used to extend my own idea for the queries. GPT was employed for the report after the first draft was done to critique my creation. It did however help with creating the vocabulary in the appendix from the code I had. Finally, it was used for support with the abstract.

C Vocabulary

Trial Information: a ct:ClinicalTrial

This is the main item the other 4 are linked to this main item.

- **NCT Number:** Unique identifier for a clinical study (`string`)
- **Study Title:** Official or lay title of the study (`string`)
- **Study URL:** Link to the official study page (`URL`)
- **Acronym:** Shortened form of the study title (`string`)
- **Study Status:** Recruitment status (`string`)
- **Brief Summary:** Concise summary of the study (`string`)
- **Study Results:** Indicates availability of study results (`boolean`)
- **Start Date:** Date when the study began (`date`)
- **Primary Completion Date:** Date when primary data collection ended (`date`)
- **Completion Date:** Actual completion date of the study (`date`)
- **First Posted:** Date when the study was first posted on ClinicalTrials.gov (`date`)
- **Results First Posted:** Date when study results were first posted (`date`)
- **Last Update Posted:** Date of the most recent update to the study record (`date`)
- **Study Documents:** Documents associated with the study, including filenames and links (`string + URL`)
- **Collaboration Details:** Information about collaboration details for the clinical trial (`URI`)
- **Location Info:** Information about the locations associated with the clinical trial (`URI`)
- **Medical Details:** Information about the medical details of the clinical trial (`URI`)
- **Participant Info:** Information about the participants involved in the clinical trial (`URI`)

Participant Information: a ct:ParticipantInfo

- **Sex:** Gender of study participants (`string`)
- **Age:** Age eligibility criteria for participants (`string`)
- **Enrollment:** Number of participants in the study (`integer`)

Medical Details: a ct:MedicalDetails

- **Conditions:** Health conditions being studied (`string`)
- **Interventions:** Treatments or procedures being studied (`string`)
- **Primary Outcome Measures:** Main outcomes assessed in the study (`string`)
- **Secondary Outcome Measures:** Additional outcomes assessed (`string`)
- **Other Outcome Measures:** Other outcomes of interest (`string`)
- **Phases:** Stage of the clinical trial (`string`)
- **Study Type:** Interventional or observational (`string`)
- **Study Design:** Investigative methods and strategies used in the study based on interventional (4 adds) or observational (2 adds) the graph additions differ (`string`)
- **Other IDs:** Additional identifiers assigned to the study (`string`)

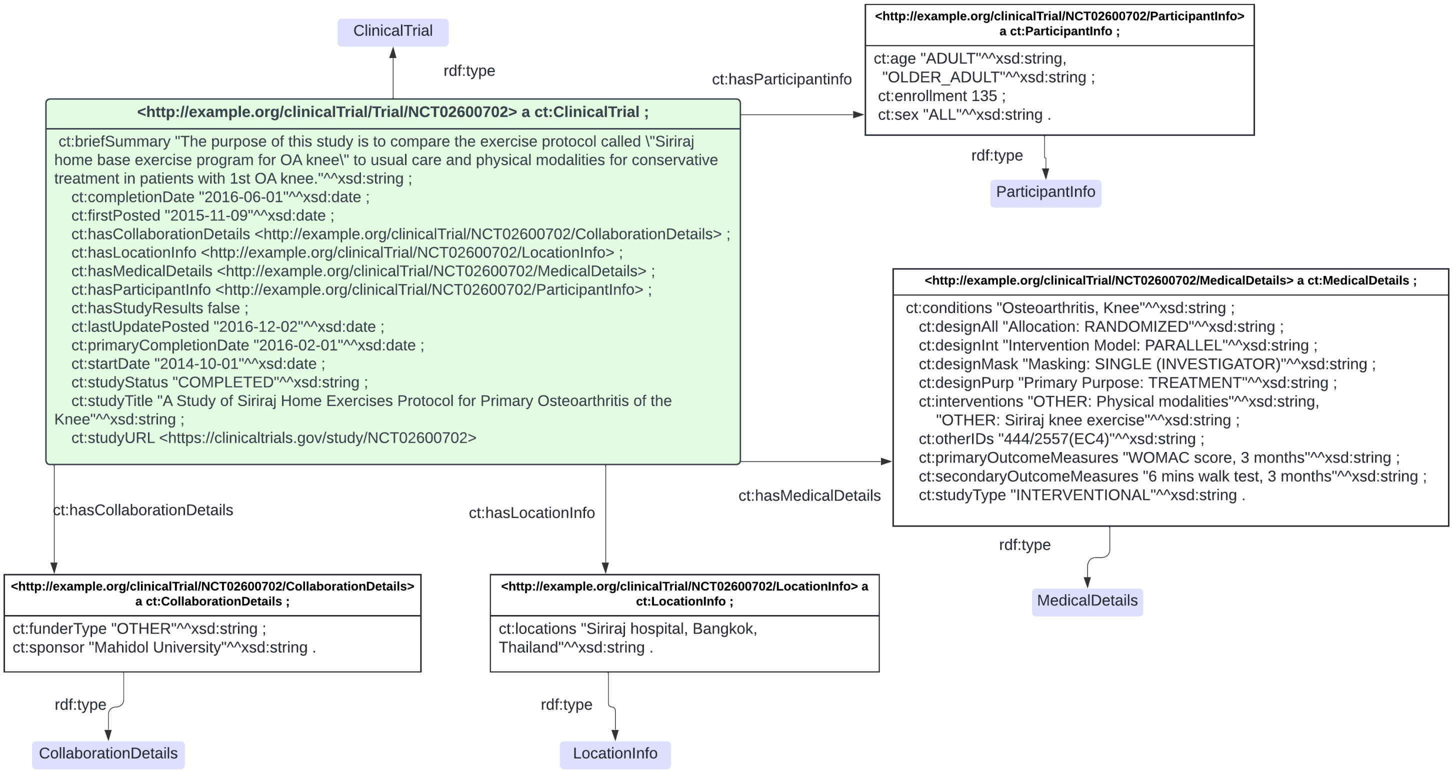
Collaboration Details: a ct:CollaborationDetails

- **Funder Type:** Type of organization providing funding/support (`string`)
- **Sponsor:** Entity initiating and overseeing the study (`string`)
- **Collaborators:** Organizations providing support for the study (`string`)

Location Information: a ct:LocationInfo

- **Locations:** Locations associated with the study (`string`)

D Figures



```
<http://example.org/clinicalTrial/Trial/NCT02600702> a ct:ClinicalTrial ;
 ct:briefSummary "The purpose of this study is to compare the exercise protocol called \"Siriraj
home base exercise program for OA knee\" to usual care and physical modalities for conservative
treatment in patients with 1st OA knee."^^xsd:string ;
    ct:completionDate "2016-06-01"^^xsd:date ;
    ct:firstPosted "2015-11-09"^^xsd:date ;
    ct:hasCollaborationDetails <http://example.org/clinicalTrial/NCT02600702/CollaborationDetails> ;
    ct:hasLocationInfo <http://example.org/clinicalTrial/NCT02600702/LocationInfo> ;
    ct:hasMedicalDetails <http://example.org/clinicalTrial/NCT02600702/MedicalDetails> ;
    ct:hasParticipantInfo <http://example.org/clinicalTrial/NCT02600702/ParticipantInfo> ;
    ct:hasStudyResults false ;
    ct:lastUpdatePosted "2016-12-02"^^xsd:date ;
    ct:primaryCompletionDate "2016-02-01"^^xsd:date ;
    ct:startDate "2014-10-01"^^xsd:date ;
    ct:studyStatus "COMPLETED"^^xsd:string ;
    ct:studyTitle "A Study of Siriraj Home Exercises Protocol for Primary Osteoarthritis of the
Knee"^^xsd:string ;
    ct:studyURL <https://clinicaltrials.gov/study/NCT02600702>
```

```
<http://example.org/clinicalTrial/NCT02600702/ParticipantInfo>
a ct:ParticipantInfo ;
ct:age "ADULT"^^xsd:string,
  "OLDER_ADULT"^^xsd:string ;
 ct:enrollment 135 ;
 ct:sex "ALL"^^xsd:string .
```

```
<http://example.org/clinicalTrial/NCT02600702/MedicalDetails> a ct:MedicalDetails ;
ct:conditions "Osteoarthritis, Knee"^^xsd:string ;
   ct:designAll "Allocation: RANDOMIZED"^^xsd:string ;
   ct:designInt "Intervention Model: PARALLEL"^^xsd:string ;
   ct:designMask "Masking: SINGLE (INVESTIGATOR)"^^xsd:string ;
   ct:designPurp "Primary Purpose: TREATMENT"^^xsd:string ;
   ct:interventions "OTHER: Physical modalities"^^xsd:string,
       "OTHER: Siriraj knee exercise"^^xsd:string ;
   ct:otherIDs "444/2557(EC4)"^^xsd:string ;
   ct:primaryOutcomeMeasures "WOMAC score, 3 months"^^xsd:string ;
   ct:secondaryOutcomeMeasures "6 mins walk test, 3 months"^^xsd:string ;
   ct:studyType "INTERVENTIONAL"^^xsd:string .
```

```
<http://example.org/clinicalTrial/NCT02600702/CollaborationDetails>
a ct:CollaborationDetails ;
ct:funderType "OTHER"^^xsd:string ;
ct:sponsor "Mahidol University"^^xsd:string .
```

```
<http://example.org/clinicalTrial/NCT02600702/LocationInfo> a
ct:LocationInfo ;
ct:locations "Siriraj hospital, Bangkok,
Thailand"^^xsd:string .
```

Figure 1: Example of an instance

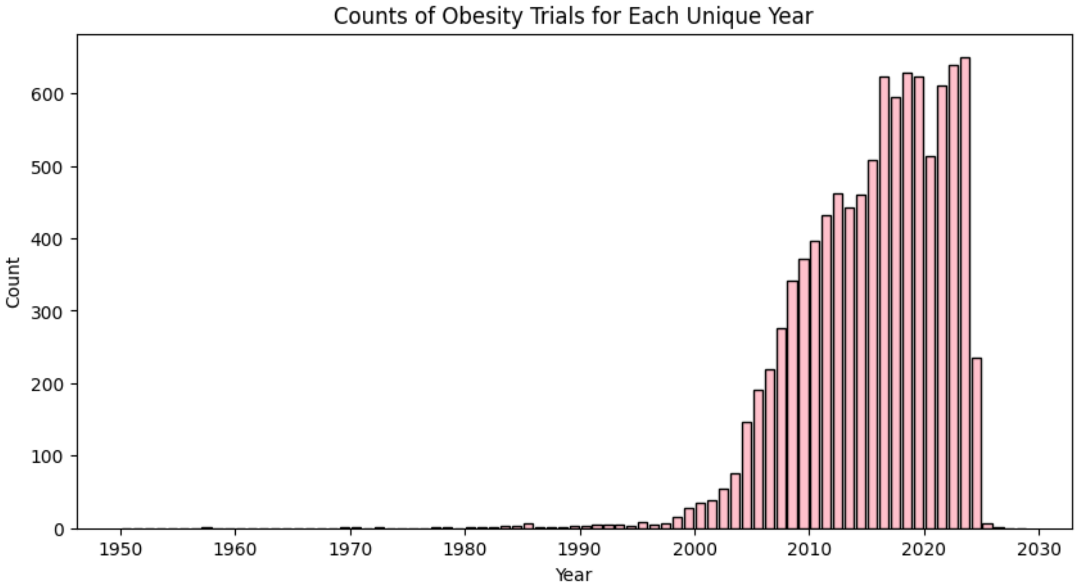


Figure 2: Clinical trials on obesity over the years

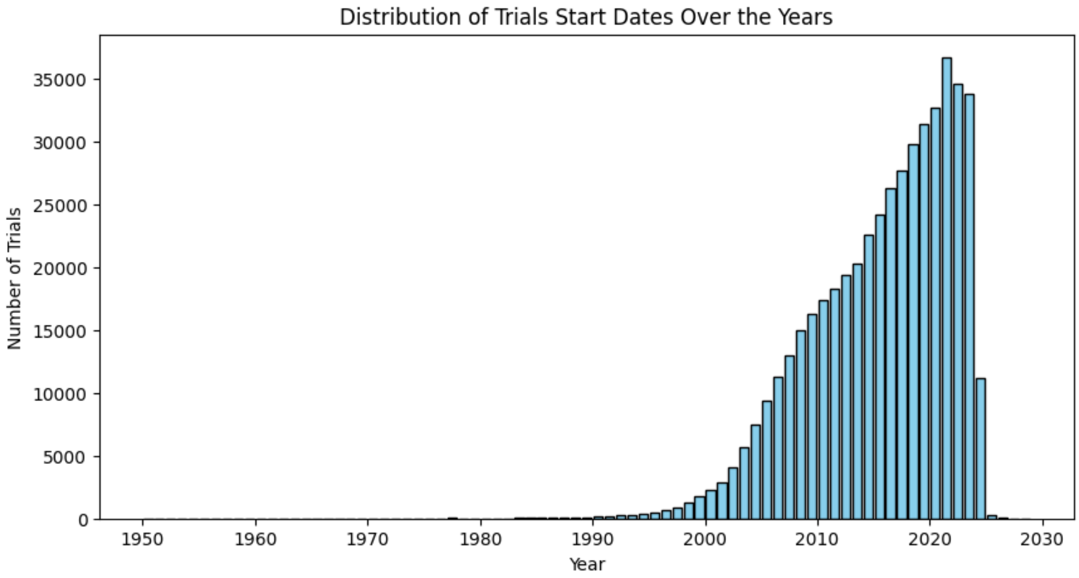


Figure 3: amount of clinical trials over the years