Heta Shah
hetas96@gmail.com

Aggregated Analysis of ClinicalTrials.gov Database

Problem Definition

Using the publicly available Aggregated Analysis of ClinicalTrials.gov dataset (AACT) please download this data and host it in a local postgres database to complete the following. Please complete each task using SQL.

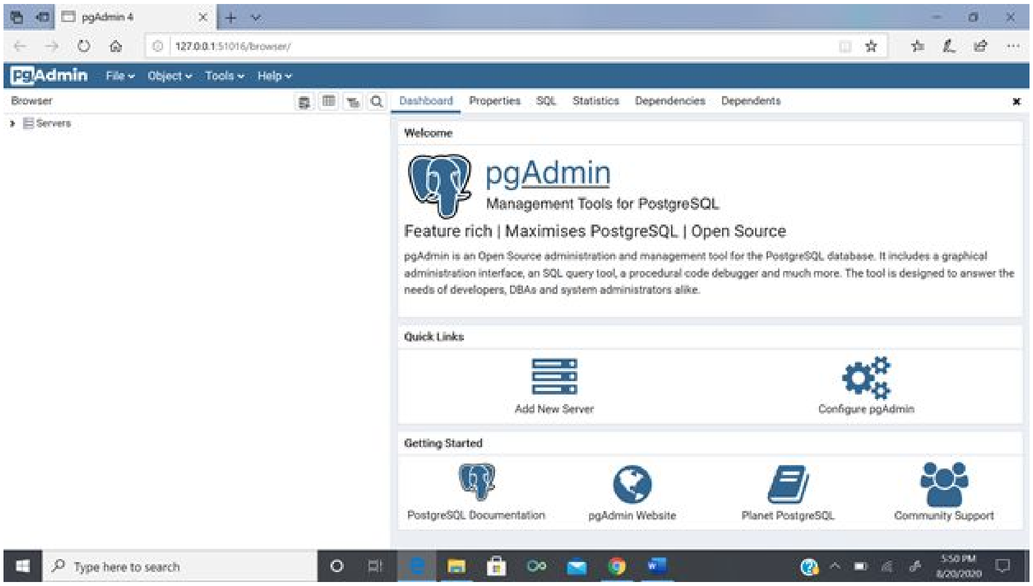
1. Create a view of all prospective cancer related clinical trials that are completed (no longer recruiting and not prematurely terminated)
   a. This view should include an nct_id, the cancer condition, inclusion/exclusion criteria for the trial, location of the trial, and the intervention of study, total participants in the study
   b. Use this view to subset/answer all below requests
2. Create view for all observed adverse events and outcomes recorded for each trial.
3. Find the trial that had the most patients with a complete response to the intervention of study (using outcome_measurements table)
4. Find the number of trials that started after 2005 and ended before 2010
5. Distribution of trials by state

Heta Shah
hetas96@gmail.com

Loading Data to PostgreSQL Database using pgAdmin 4

Connecting to postgres Server

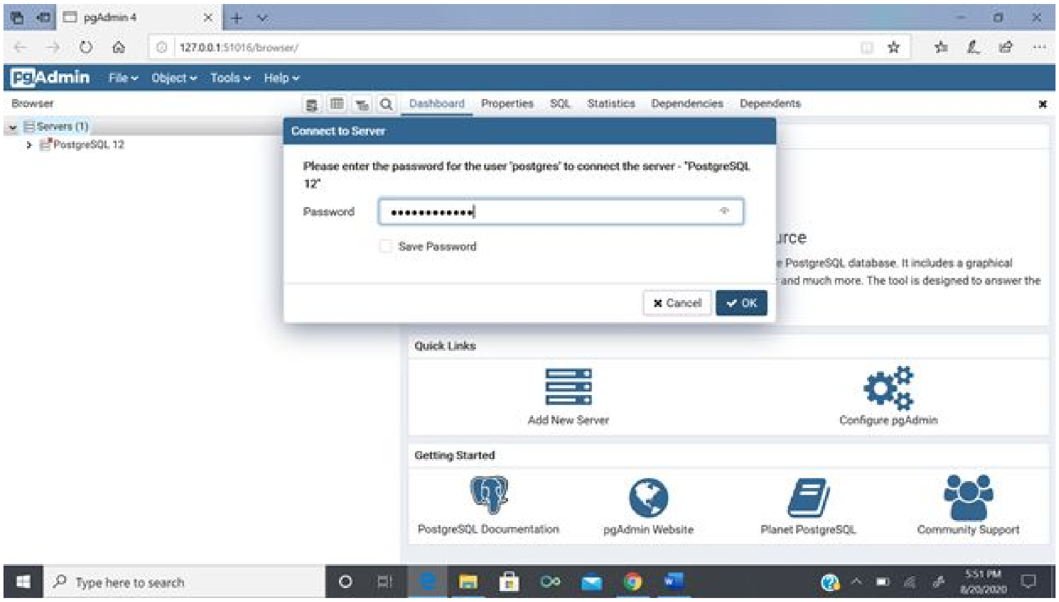
Launch pgAdmin 4:



Feature rich | Maximises PostgreSQL | Open Source

pgAdmin is an Open Source administration and management tool for the PostgreSQL database. It includes a graphical administration interface, an SQL query tool, a procedural code debugger and much more. The tool is designed to answer the needs of developers, DBAs and system administrators alike.

Double-click on “Servers” section under the “Browser” section on the left-side of the window. Feed the password when prompted, and click on OK:



Please enter the password for the user 'postgres' to connect the server - "PostgreSQL 12"

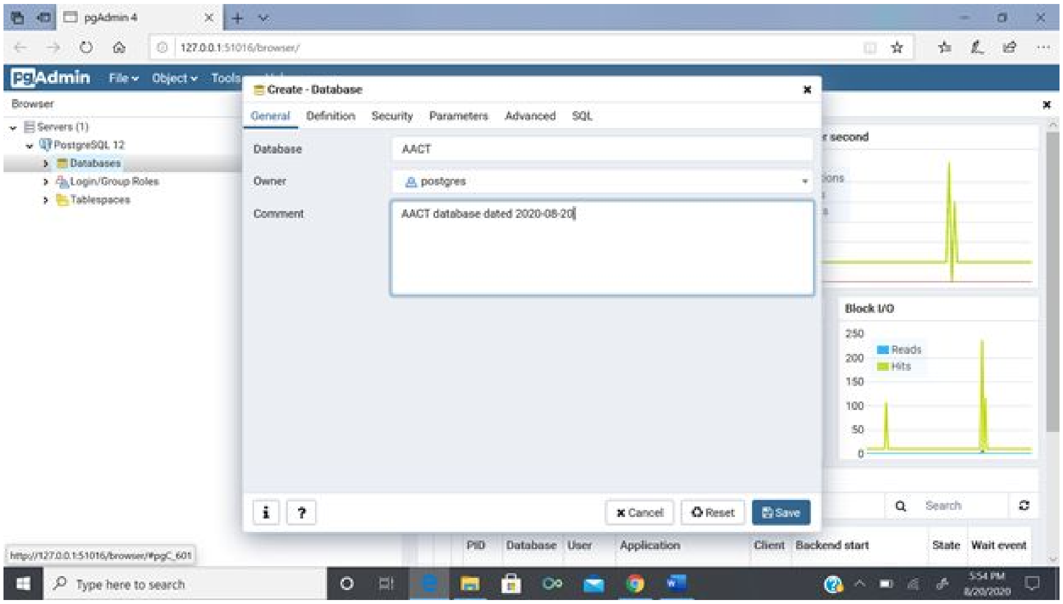
Password ••••••••••••

□ Save Password

Heta Shah
hetas96@gmail.com

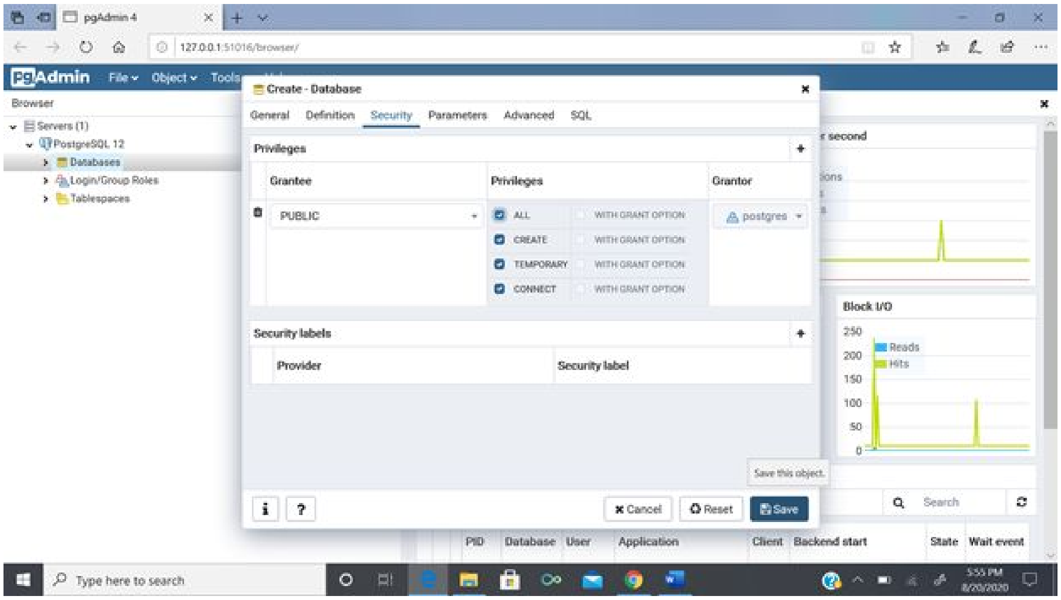
Creating AACT database

a). Select “Databases” under Servers. Then from the menu ribbon on the top of the page, select “Object -> Create -> Database”, and feed the following information:



Database: AACT
Owner: postgres
Comment: AACT database dated 2020-08-20

b). Under the Security tab, click on “+” under Privileges – select PUBLIC as Grantee, check ALL under Privileges, and keep Grantor as postgres. Then click Save:

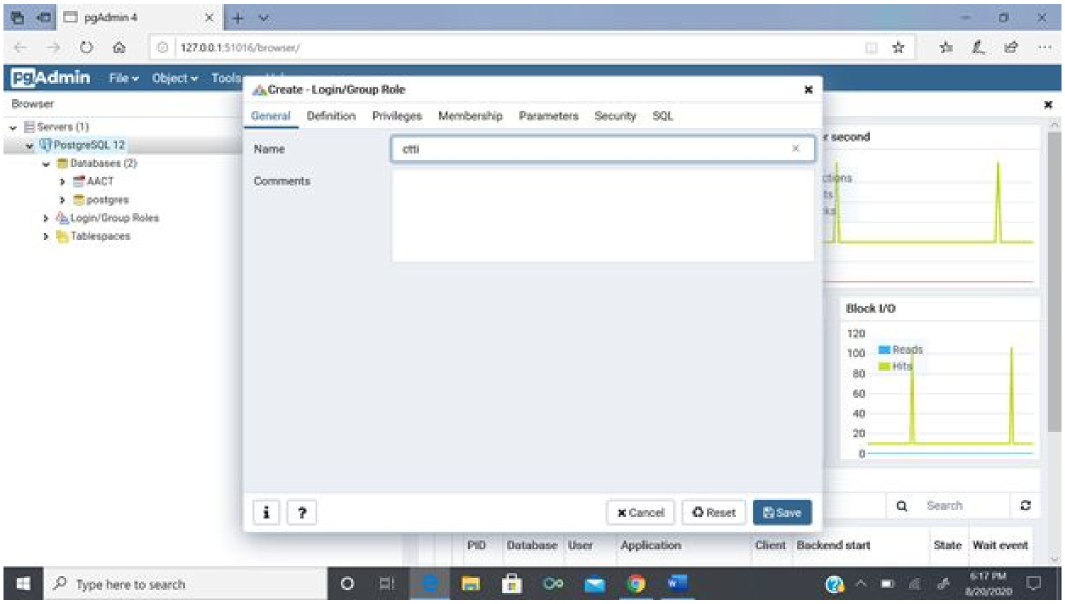


Privileges

Grantee	Privileges		Grantor
PUBLIC	☑ ALL	□ WITH GRANT OPTION	postgres
	☑ CREATE	□ WITH GRANT OPTION	
	☑ TEMPORARY	□ WITH GRANT OPTION	
	☑ CONNECT	□ WITH GRANT OPTION	

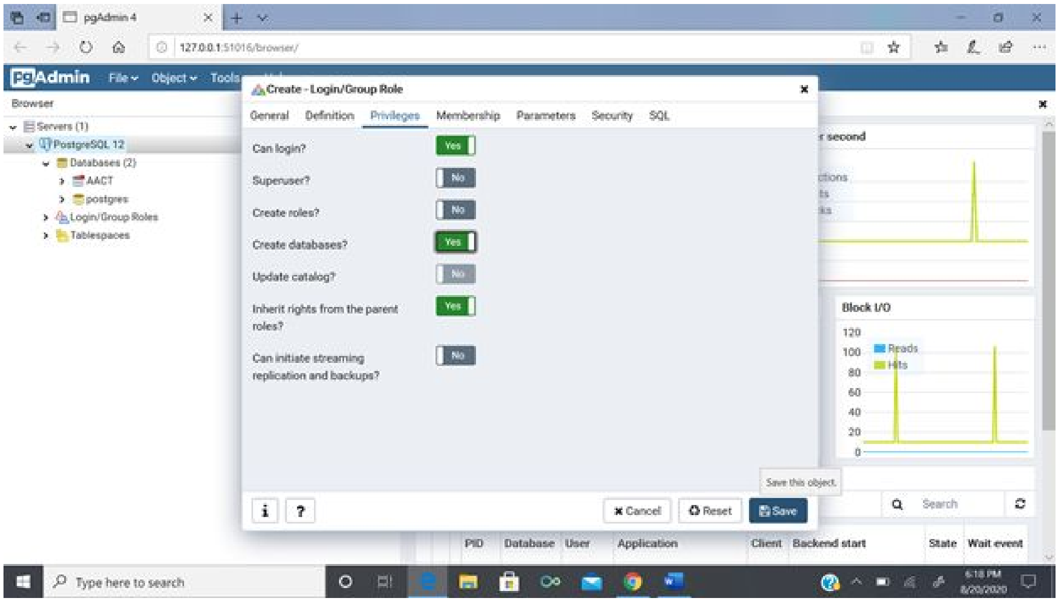
Creating new database user/role to import data from the AACT dump file. This is because all database objects in the dump file are owned by a user named ctti. If the new role doesn't exist, we will get an error -

a). To do so, select PostgreSQL 12 from the left pane. Then click on Object -> Create -> Login/Group Role. Give the name as ctti and click on Privileges tab:



Name: ctti

b). Under the Privileges tab, toggle on "Can Login?" and "Create databases?" options, and click on Save:

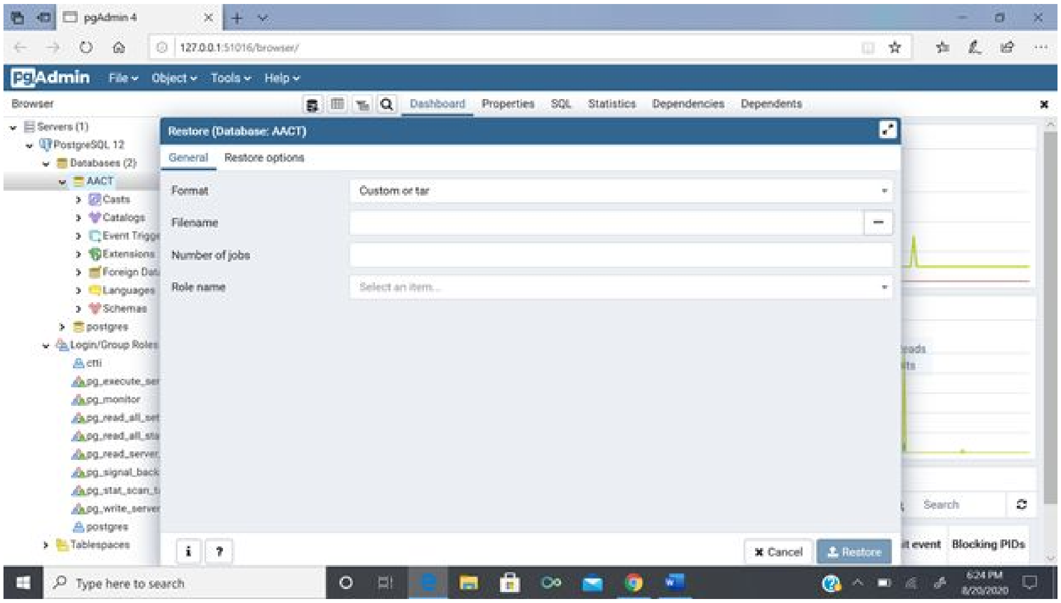


Can login? Yes
Superuser? No
Create roles? No
Create databases? Yes
Update catalog? No
Inherit rights from the parent roles? Yes
Can initiate streaming replication and backups? No

Heta Shah
hetas96@gmail.com

Importing data from AACT dump file into AACT database

Right click on AACT database and select Restore option:

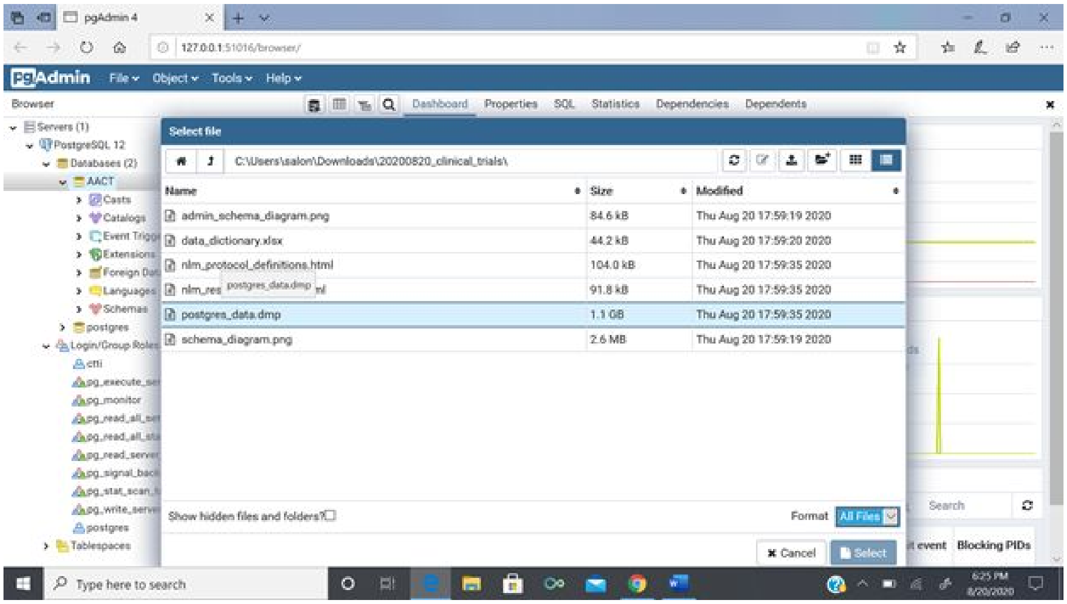


Restore (Database: AACT)

General | Restore options

Format	Custom or tar
Filename	
Number of jobs	
Role name	Select an item...

Under Filename, browse the location where your unzipped AACT downloaded package is, and select the postgres_data.dmp file:



Select file

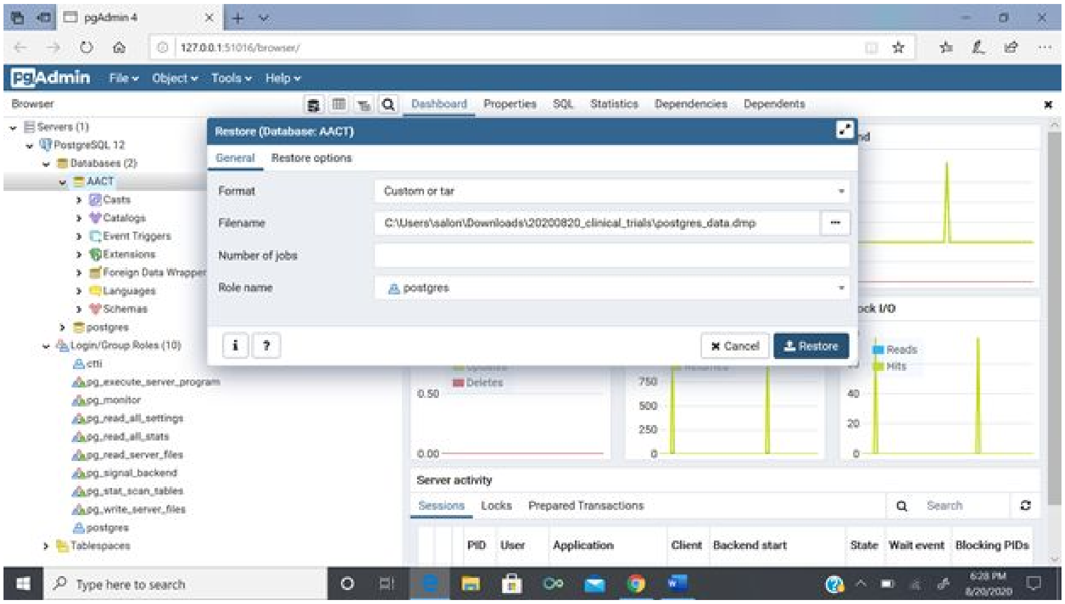
C:\Users\salon\Downloads\20200820_clinical_trials\

Name	Size	Modified
admin_schema_diagram.png	84.6 kB	Thu Aug 20 17:59:19 2020
data_dictionary.xlsx	44.2 kB	Thu Aug 20 17:59:20 2020
nlm_protocol_definitions.html	104.0 kB	Thu Aug 20 17:59:35 2020
nlm_res postgres_data.dmp ml	91.8 kB	Thu Aug 20 17:59:35 2020
postgres_data.dmp	1.1 GB	Thu Aug 20 17:59:35 2020
schema_diagram.png	2.6 MB	Thu Aug 20 17:59:19 2020

Show hidden files and folders? □

Format: All Files

Under Role name, select postgres, and click on Restore:



Restore (Database: AACT)

General | Restore options

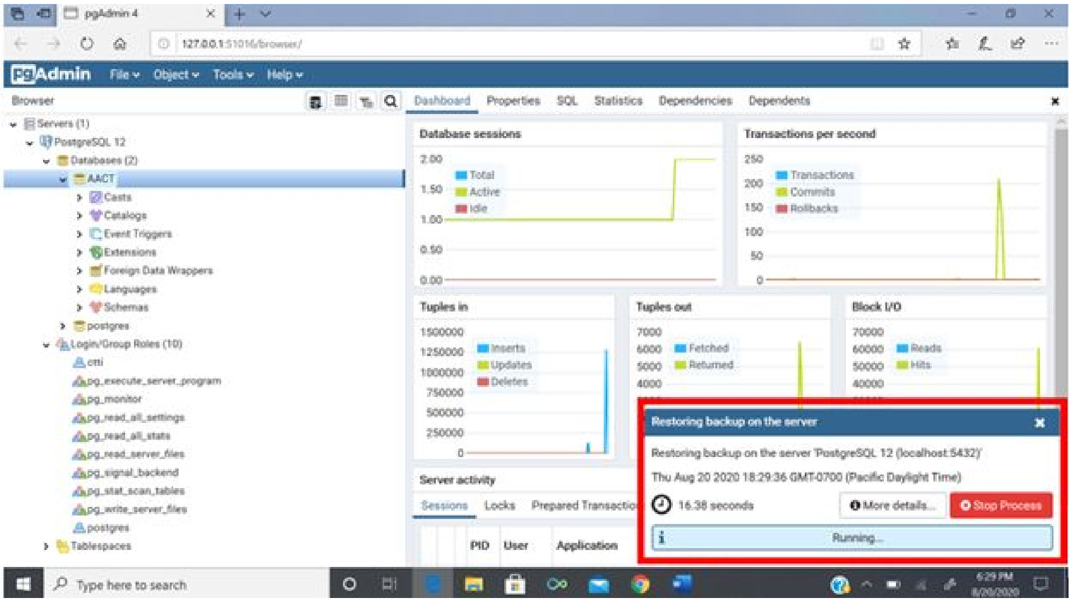
Format: Custom or tar

Filename: C:\Users\salon\Downloads\20200820_clinical_trials\postgres_data.dmp

Number of jobs:

Role name: postgres

Data restoring takes some time. Progress can be tracked at the bottom-right of the screen (highlighted in red):



Restoring backup on the server

Restoring backup on the server 'PostgreSQL 12 (localhost:5432)'

Thu Aug 20 2020 18:29:36 GMT-0700 (Pacific Daylight Time)

16.38 seconds

Running...

Heta Shah
hetas96@gmail.com

SQL code to Query the database

To be able to perform any tasks, I first took time to familiarize myself with the database using the Data Dictionary provided. I also did some Exploratory Data Analyses (EDA) for the required tables to get an idea of the structure of the data. The aim was to know

- how many records are in each table,
- how many distinct records are in the table,
- which column from one table relates to which other column from the other table,
- how many values are missing (if any) and how it will affect my analyses,
- whether the data in a column is numeric, categorical or descriptive text, etc.

Task 1

Create a view of all prospective cancer related clinical trials that are completed (no longer actively recruiting and not prematurely terminated)
a. This view should include an nct_id, the cancer condition, inclusion/exclusion criteria for the trial, location of the trial, and the intervention of study, total participants in the study
b. Use this view to subset/answer all below requests

Solution:
I identified the following relations as necessary:
I. ctgov.conditions
II. ctgov.studies
III. ctgov.eligibilities
IV. ctgov.facilities
V. ctgov.interventions

Approach:
Instead of joining complete tables, first obtain required subsets and then perform necessary joins. This helps in reducing computation, and thus, quicker query processing:

I. Select only a subset of the ctgov.conditions table to limit the data to include only cancer-related trials (name ilike '%cancer%'). This yields a subset of all rows where the condition name contains the word "cancer", ignoring the case, and it yields ~53,000 records. To get data about completed clinical trials, I needed ctgov.studies relation.

II. Select only a subset of the ctgov.studies table to limit the data to include only completed cancer-related trials (overall_status = 'Completed'). This yields ~190,000 records. I joined the nct_id column of both tables to identify the trial conditions, status and the number of participants in each trial. This join returns only ~25,000 records that satisfy all conditions.

III. Obtain the inclusion/exclusion criteria from ctgov.eligibilities table which returns ~350,000 records. Joining on nct_id column to the previous subset returns ~25,000 rows about trial ID, condition, status, eligibility criteria, etc.

IV. Getting city, state and country columns from ctgov.facilities table gives us details about the location of the trial. This subset of ~6,000 rows is joined on nct_id of conditions table, since we don't want other trials.
NOTE: facilities table is not filtered on "status = 'Completed'" condition since that condition is already taken care of while selecting data from conditions table.

V. Select intervention type (eg: Drug) and name of intervention from ctgov.interventions table. This returns ~600,000 records. However, we want only records that match our above-mentioned criteria so I join on nct_id of conditions table.

VI. Create a view on this query, for further use.

SQL Code Snippet:

```sql
CREATE VIEW cancer_view AS
SELECT sub1.nct_id, sub1.condition, sub1.condition_name,
            sub2.overall_status, sub2.participants,
            e.criteria,
            f.city, f.state, f.country,
            i.intervention_type, i.name AS intervention
FROM (SELECT c.nct_id, c.name AS condition, c.downcase_name AS
condition_name
            FROM ctgov.conditions AS c
-- selecting only cancer trials before joining
            WHERE c.name ILIKE '%cancer%') sub1
JOIN (SELECT s.nct_id, s.overall_status, s.enrollment AS participants
            FROM ctgov.studies AS s
-- selecting only completed cancer trials before joining
            WHERE s.overall_status = 'Completed') sub2
      ON sub1.nct_id = sub2.nct_id
JOIN ctgov.eligibilities AS e
      ON sub1.nct_id = e.nct_id
JOIN ctgov.facilities AS f
      ON sub1.nct_id = f.nct_id
JOIN ctgov.interventions AS i
      ON sub1.nct_id = i.nct_id;
```

Analysis:
The resulting View contains 725,646 records giving details about the NCT ID of completed cancer-related clinical trials, cancer condition, trial locations, eligibility criteria, type and name of intervention, number of participants, etc.

Task 2:

Create a view for all observed adverse events and outcomes recorded for each trial
Solution: Using the data dictionary, I identified the following new tables to be used:
I. ctgov.reported_events
Along with the existing view:
II. cancer_view

Approach:
Select adverse events and outcomes recorded for only completed cancer trials.

SQL Code Snippet:

```sql
CREATE VIEW cancer_trials_adverse_outcomes AS
SELECT *
FROM ctgov.reported_events
-- observed adverse events and outcomes recorded for each completed cancer trial
WHERE nct_id IN (SELECT DISTINCT nct_id
                              FROM cancer_view)
```

Analysis:
This view returns 556,976 records of adverse outcomes from cancer-related clinical trials. Some of these adverse outcomes include Diarrhoea, Anameia, Abdominal Pain, Constipation, Cough, Rash, etc.

Task 3

Find the trial that had the most patients with a complete response to the intervention of study (using outcome_measurements table)

Solution:
Tables and Views required for this task are:
I. cancer_view
II. ctgov.outcome_measurements
III. ctgov.studies

Approach:
I. Calculate count of cancer trials that have received complete response on intervention study from the patients
II. Select the trial with the greatest number of patients who showed complete response
III. Display basic details about this trial

SQL Code Snippet:

```
-- additional details about the trial
SELECT nct_id, start_date, completion_date, study_type, official_title,
source
FROM ctgov.studies
WHERE nct_id = (SELECT sub.nct_id
                  FROM (SELECT nct_id, category, COUNT(*) AS
category_ranking
                              FROM ctgov.outcome_measurements
                              -- completed cancer trials
                              WHERE nct_id IN (SELECT DISTINCT nct_id
                                                FROM cancer_view)
                              GROUP BY 1, 2
                              -- with complete response
                              HAVING category ILIKE '%complete response%'
                              ORDER BY 3 DESC
                              -- the only trial having the most patients
with complete response to intervention study
                              LIMIT 1) sub)
```

Analysis:
The trial sourced by AstraZeneca with nct_id NCT02127710 had the most patients with a complete response to the intervention of study. This trial was started on April 30, 2014 and completed on April 20, 2020.

Task 4

Find the number of trials that started after 2005 and ended before 2010

Solution:
Table and View used for this task:
I. cancer_view
II. ctgov.studies

Approach:
I. Identify cancer-related completed trials that started after 2005 and completed before 2010
II. Count the number of such trials

SQL Code Snippet:

```
-- number of trials
SELECT COUNT(*) AS trials_count
FROM (SELECT DISTINCT nct_id, start_date, completion_date
            FROM ctgov.studies
            WHERE nct_id IN (SELECT DISTINCT nct_id
                              FROM cancer_view)
        -- that started after 2005 and completed before 2010
                  AND EXTRACT(YEAR FROM start_date) > 2005 AND EXTRACT(YEAR
FROM completion_date) < 2010) sub
```

Analysis: 555 distinct cancer-related trials, which are now completed, were started after 2005 and completed before 2010.

Task 5

Distribution of trials by state

Solution:
I have used cancer_view for this task.

Approach:

Heta Shah
hetas96@gmail.com

Select the distribution of number of trials by state, country (ordered in descending order of trial count)

SQL Code Snippet:

```
SELECT DISTINCT state, country, COUNT(*) AS trials_count
FROM cancer_view
-- distribution of trials by state
GROUP BY state, country
ORDER BY trials_count DESC
```

Analysis:
Most cancer-related trials are conducted and completed in the United States. Top 3 leading states are Ohio, California and Illinois, with a total of 113,125 trials in these states alone.

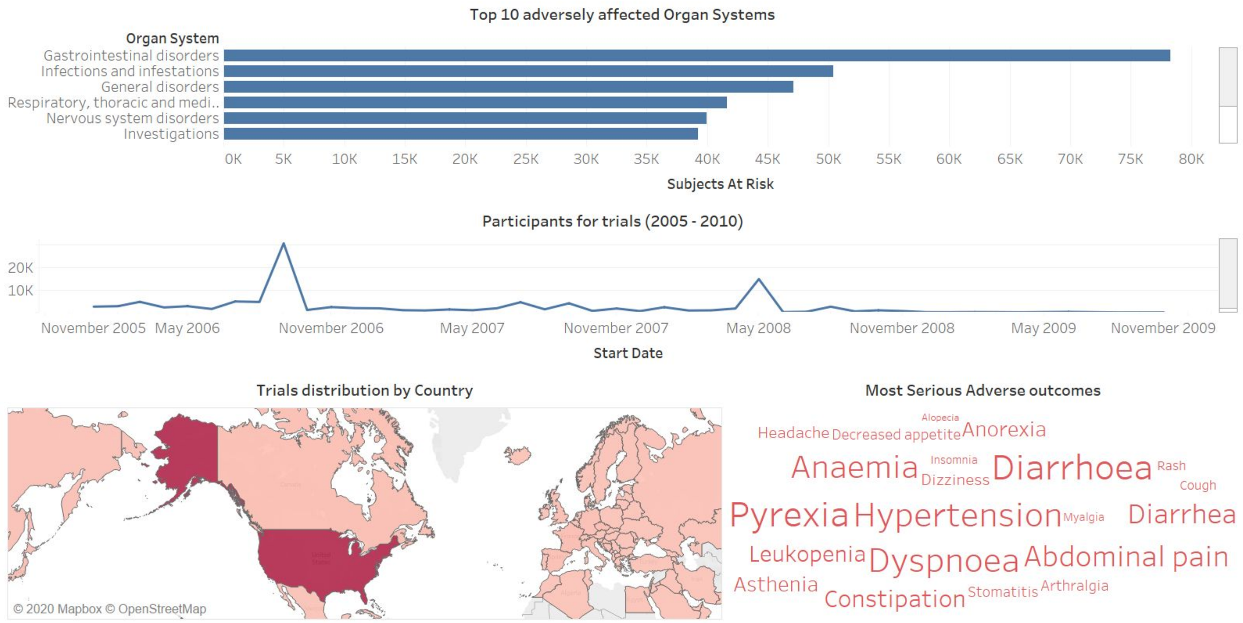
Heta Shah
hetas96@gmail.com

Tableau Analysis

Some further analysis leveraging Tableau helped me identify:

- Around 30,000 trials were started in September 2006
- Diarrhoea, Pyrexia Hypertension and Dyspnoea are the most serious adverse outcomes of clinical trials
- USA conducted the maximum number of clinical trials (~500,000 trials)
- Gastrointestinal disorders due to clinical trials have put 75,000+ human subjects at risk
- Other adverse effects on organ systems include infections and infestations, respiratory disorders and nervous system disorders

Here is a Tableau dashboard providing insights into some analytical insights:



Future Scope

Given more time, some noticeable area of improvements are:

- Handling missing/null values. Some data can be imputed with the average, while some other records can be dropped altogether
- Dirty data, or data with logical spelling errors can be replaced with the correct spelling
- Get deeper understanding of all relations in the database, and consequently prepare more interactive dashboards and stories