COMP 533 Spring 2018 Assignment #1

Due January 24, 2018 at 11:55PM.

1 Description

The goal of this assignment is to write Relational Calculus and Relational Algebra expressions, as well as several straightforward SQL queries.

The RC/RA questions are based on the relations listed in this document. The SQL queries will answer questions about clinical trials registered with the US government. The data is a **subset** of the data made available through the Clinical Trials Transformation Initiative. Since the data provided is a subset of the full data set, so be sure to use the version of the data available on the course Canvas site. If you are interested, you can learn more about the dataset at http://aact.ctti-clinicaltrials.org.

1.0.1 What's In and Out of Scope

With regards to SQL, this is intended to be a declarative SQL assignment. Therefore, you must write queries in SQL (not stored procedures or functions). You may use VIEWs as needed and you may use standard built-in MySQL functions (e.g. ROUND, IF or CASE statements). If you're not sure if something is allowed, ask!
Some helpful functions:

- LIMIT: Limit the number of records returned from a query by using the ending term "LIMIT N", where N is the number of records to return.
  For example, "SELECT * FROM studies LIMIT 100" returns 100 records from the studies table.
- DATE_ADD: Adds the number of specified intervals to a date.
  DATE_ADD([date],INTERVAL [expr unit])
  Where date is the starting date and [expr unit] is the quantity and unit to add.
  For example DATE_ADD(start_date, INTERVAL 5 DAY)
- EXTRACT: Extracts the specified information from the date.
  EXTRACT([unit] FROM [date])
  Where [unit] is one of {YEAR, MONTH, DAY, HOUR_MINUTE, ...} and [date] is the date to extract the unit from.

  A complete list of allowable units may be found here:
  `https://dev.mysql.com/doc/refman/5.7/en/date-and-time-functions.html#function_date-add`

2 Turnin

Create a document that contains your RC/RA expressions and SQL code, as well as the results from running your code. By 11:55P on the due date, submit this document electronically to Canvas. You must submit a text or pdf file with a .txt or .pdf extension. Other formats (such as Microsoft Word) are not acceptable. Your assignment must be typed - that is, a scanned, hand-written document is not acceptable (we have to be able to easily read what you wrote).

3 Grading

Each question is worth 5 points. If you don't get the right answer or your expression or query is not correct, you won't get all of the points; partial credit may be given at the discretion of the grader.

In some cases, you are asked to provide a certain number of answers. If you provide more than the requested number, the grader will only consider the first n, where n is the number of answers asked for.

4 Academic Honesty

The following level of collaboration is allowed on this assignment: You may discuss the assignment with your classmates at a high level. Any issues and help getting MySQL running is totally fine. What is not allowed is direct examination of anyone else's RC/RA expression or SQL code (on a computer, email, whiteboard, etc.) or allowing anyone else to see your RC/RA expressions or SQL code. You MAY post and discuss query results with your classmates.

You may use the search engine of your choice to lookup the syntax for SQL commands, but may not use it to find answers to queries.

5 Relational Calculus & Relational Algebra

The Relations

Consider the following relations, which describe patients and visits to see clinicians:

PATIENT(MRN, FIRSTNAME, LASTNAME, AGE)
VISIT(VISIT_ID, MRN, DATETIME, LOCN)
PROCEDURE(VISIT_ID, NAME, CLIN_ID)
CLINICIAN(CLIN_ID, FIRSTNAME, LASTNAME, CERT)

- Relation PATIENT includes information about patients, including their Medical Record Number (MRN), name and age.
- VISIT describes a medical visit where a patient is seen at a certain time and place.
- Relation PROCEDURE describes each named procedure that was performed during the visit and specifies who performed the procedure. Since multiple clinicians can participate, all fields are part of the primary key.
- CLINICIAN lists all the known clinicians and their highest certification.

1. Give one reason why MRN is a poor choice for the primary & foreign keys?
2. Write Relational Calculus expressions for the following:
   (a) Who is 25 years old?

   (b) Who had a medical visit in December, 2017?

(c) Who got a 'flu shot'?

(d) Who did NOT get a 'flu shot'?

(e) What is the first and last name of all patients who have seen MD Paula Jones?

3. Write Relational Algebra expressions for the following questions. Return the relation primary key to identify the tuples, unless otherwise specified.

   (a) Who is 25 years old?

   (b) What is the first and last name of all patients who have seen a Physician's Assistant (cert is 'PA')

   (c) Which patients have the same names and age?

   (d) Which patients who got a flu shot also got a measles immunization during the same visit?

   (e) Which patients who have seen an MD have not seen a PA?

6 Queries

6.1 Getting Started

First, go to your database, and create the following tables. Paste these definitions into a query window and click on the lightening bolt to execute them. If you want to execute just one at a time, highlight the SQL code you want to execute and click on the lightening bolt.

```
CREATE TABLE studies (
    nct_id CHARACTER(11),
    start_date DATE,
    start_date_type CHARACTER(11),
    completion_date DATE,
    completion_date_type CHARACTER(11),
    study_type  VARCHAR(35),
    brief_title VARCHAR(350),
    overall_status   VARCHAR(35),
    phase   VARCHAR(20),
    enrollment INTEGER,
    enrollment_type CHARACTER(11),
    source VARCHAR(150),
    why_stopped VARCHAR(200),
    is_fda_regulated_drug BOOLEAN,
    PRIMARY KEY studies_pk(nct_id)
);
```

```
CREATE TABLE reported_events(
    id INTEGER,
    nct_id CHARACTER(11),
    event_type VARCHAR(15),
    subjects_affected INTEGER,
    subjects_at_risk INTEGER,
    description VARCHAR(500),
    event_count INTEGER,
    organ_system VARCHAR(100),
    adverse_event_term VARCHAR(100),
    PRIMARY KEY reported_events_pk(id)
);
CREATE TABLE designs (
    id INTEGER,
    nct_id CHARACTER(11),
    allocation VARCHAR(15),
    intervention_model VARCHAR(25),
    observational_model VARCHAR(25),
    primary_purpose VARCHAR(25),
    time_perspective VARCHAR(25),
    masking VARCHAR(25),
    masking_description VARCHAR(1000),
    intervention_model_description VARCHAR(1000),
    subject_masked BOOLEAN,
    caregiver_masked BOOLEAN,
    investigator_masked BOOLEAN,
    outcomes_assessor_masked BOOLEAN,
    PRIMARY KEY designs_pk(id)
);
CREATE TABLE conditions (
    id INTEGER,
    nct_id CHARACTER(11),
    name VARCHAR(200),
    downcase_name VARCHAR(200),
    PRIMARY KEY conditions_pk(id)
);
```

6.2 Load the data

Load the data needed for the assignment. You should do this in MySQL Workbench. The files are located on the class Canvas site in https://canvas.rice.edu/courses/10475/files/folder/A1/A1data.zip. They are .sql files. Copy the contents of each file into a query window in MySQL Workbench and execute all of the queries.

You must use the table and attribute names provided. Do not rename anything.

6.3 Questions

Answer all of the questions below by writing and executing SQL queries. The queries must contain ONLY the answer to the question (no extra rows or columns). You many only use SQL to answer the questions. You may need to explore the database a bit prior to generating your final solutions.

1. List the nct_id and study_type from the study whose brief_title is "Autologous Cell Therapy After Stroke".
2. List the different values for study_type, in alphabetical order.
3. Which studies that started in 2016 have reported events?
4. Which of the studies that started in February 2016, but on or after the 15th, are expected to complete (or have completed) within 6 months of their start date?

7 Reading / Short answer

Read the article "The Database for Aggregate Analysis of ClinicalTrials.gov (AACT) and Subsequent Regrouping by Clinical Specialty" by Tasneem et al.(pone.0033677.pdf), focusing on the database design content (vs. the medical classification content) and answer the following questions. You might need to run some queries to examine the data in order to answer the questions.

1. List three reasons the database was "normalized."
2. On page 4, the authors list a number of data elements that are contained in the XML download files from ClinicalTrials.gov. How are these different elements implemented in the design table in the AACT database?
3. Which non-description field in the design table is least populated? That is, which field is most often left blank?
4. In the conditions table there is an attribute called "name" and an attribute called "downcase_name". What is the difference? Why might a database provide both of these fields? What trade-offs are involved?
5. Look at some of the name, downcase_name pairings. Do you see any anomalies? Give 2 examples of what challenges might these anomalies pose to a user of the database.