Database Theory and Applications for Biomedical Research and Practice (BMIN 502/EPID 635)
Spring 2019
Assignment 7: Working with external data, Data importing and exporting, and Views

1. Create a database for the PennOmics sample data.

2. Using the MySQL Workbench, import the PennOmics data into your database.

3. Write a query to create a view the reflects a join with demographics *and* diagnosis
    a. Use the following as sample code:
    ```
    CREATE OR REPLACE VIEW `patient_dx` AS
    SELECT demographics.patient_id AS demo_pat_id,demographics. age,
    demographics.race, diagnosis.patient_id AS dx_pat_id, dx,diagnosis_code,
    dx.diagnosis description
    FROM demographics
    JOIN diagnosis ON demographics.patient_id = dx.patient_id
    ```
    b. Create and run a SELECT query that returns the results of the view you created above
    c. Run the queries and submit them and a snapshot of the result set to Assignment 7 as: **yourlastname_BMIN502_19_7a.pdf**

4. Write a query to create a view the reflects a join with demographics *and* diagnosis *and* procedures. Run the queries and submit them and a snapshot of the result set to Assignment 7 as: **yourlastname_BMIN502_19_7b.pdf**

Pointers for Assignment 7

- Download the PennOmics sample data from the Files folder on Canvas and place in a folder on your laptop
- In MySQL Workbench, create a new connection, called PennOmics
- Create a new schema, called PennOmics
- Right-click on “Tables” in the schema, and use the Table Data Import Wizard to import the three tables in the PennOmics sample data (Demographics, Diagnoses, and Procedures)
    - After each file is imported, click on the Log button (lower left in the pane) to get a sense of the errors, but don’t try to fix anything (yet).
- Click on the refresh icon in the Schemas bar (it’s really tiny, and looks like a recycle icon). All of your tables should now appear.
- In the lower left-hand pane (Object Info), you will see a listing of the fields in each table. Take a look at these, and you will see that you have potential problems. For example, the Demographics table has field names that contain illegal characters (for use in other programs). And, the PATIENT_IDENTIFIER field doesn’t show up for Diagnoses or Procedures. Thus, All three files in the PennOmics sample data folder have dirty data, and you will have errors in executing the SQL code on the previous page. For now, just edit the Demographics file
    - Replace the spaces in the field names with underbars (“_”)
    - Remove the parentheses in (DOB)
    - Save the file as DemographicsClean.csv
    - Drop the old Demographics table from the schema (right-click on the table name, click on Drop Table)
    - Import DemographicsClean. Click on refresh to make sure it imported correctly.
- Enter the SQL code on the previous page, and try running it. Take a look at the error message at the bottom. Pretty mysterious, right? When you see messages like this, be sure to refer to the Object Info pane and that will give you a clue as to the source of the error.
- Now, fix the other two files. You will need to open each one in a text editor and do the following:
    - Diagnoses and Procedures:
        - There is a hidden non-ASCII character in the first fieldname (PATIENT_IDENTIFIER). You need to completely replace this field name by typing it in from scratch as: PATIENT_IDENTIFIER with no spaces.
        - Save the files as DiagnosesClean and ProceduresClean.
- Drop the old Diagnoses and Procedures tables and import the clean ones, respectively. Click on refresh to make sure everything imported. Then check on the Object Info pane to ensure that PATIENT_IDENTIFIER and the others fields are in place.
- Run the SQL code to create the view, and then finish up the assignment!