**Project 2**

Your team is part of the newly established Data Analytics group at a nation-wide insurance company. A public health expert from a top university in the northeastern area was invited to give a seminar in your company. The expert shared a recent research finding that among all patients receiving pacemaker implantation, males and females are diagnosed differently in their emergency room visits that occur previous to the implantation. It is possible that male patients are overlooked (under-diagnosed) for some medical conditions in the emergency room that might lead to heart problems. If those male patients can be diagnosed earlier, they might receive preventative treatment in more timely fashion and avoid the pacemaker surgery.

Your company has collected hospital inpatient care records in 2010 and emergency room visiting records in 2009 and 2010 for a state in the southeastern region.  Your group is being tasked to  investigate if the same pattern occurs in your patient population by analyzing the data in your database.

**Data**

Download three data files as contained in the [Project 2 data zip file](https://drive.google.com/a/uncc.edu/file/d/0ByF-_HZqXT00Wjd4VVdiYVZQZ0U/view?usp=sharing) (this is a UNCC Google Drive Share, so you will need to be logged in to your UNCC Google Account). The file sid\_2010.csv contains all hospital inpatient care records in 2010 for the covered state. The file sedd\_2009.csv and sedd\_2010.csv capture all emergency room visit records in 2009 and 2010. The files are in csv (comma separated value) format. The first line is the header line and tells what attributes are included in each dataset.

<https://www.hcup-us.ahrq.gov/db/state/siddist/sid_multivar.jsp> (for inpatient care records)  
<https://www.hcup-us.ahrq.gov/db/state/sedddist/sedd_multivar.jsp> (for emergency room visit records)

**Analysis Steps**

You will submit 2 files for this assignment.  The first file (named p2.pdf) will be a PDF document that has (a) answers to the text response questions and (b) screen shots illustrating the execution of the SQL queries.  The second file (named p2.sql) will be a text file containing only the SQL queries.  Each part should be identified by the step and sub-step (e.g., 1a, 1b, etc).

**Step 1.** Identify the patient population. The patients to be included for this analysis should meet two conditions:

*Condition 1* the person was diagnosed with heart block or bradycardia, and received a pacemaker implantation in his or her inpatient care records. Diagnosis codes for heart block include '4260', '4261', '4262', '4263', '4264', '4265', '4266','4267', '4268', and '4269'; and the diagnosis code for bradycardia is '42789'. Principal procedure codes for pacemaker implantation include '3781', '3782', and ’3783’.

* **a.** Do your own research and explain why so many codes are representing one medical condition.
* **b.** Do your own research and explain why there are 31 diagnosis columns in the inpatient care dataset.
* **c.** Create an SQL query to identify the patient population that meets ￼this condition. Explain how the query works in the report document.
* **d.**Create an SQL View to represent this patient population, based on your query.

*Condition 2* the same person visited emergency room within 365 days before the pacemaker implantation. For the patients you identified by first criterion, you need to narrow down to a subset where the patient has at least visited an emergency room once within 365 days before the implantation. If a patient visited the emergency rooms multiple times during that period, you only keep the record of his or her last emergency room visit.

* **e.** What is DaysToEvent? Why does the database use DaysToEvent, instead of actual dates? How do you calculate the time difference between pacemaker implantation and an emergency room visit?
* **f.** Create an SQL query to identify the patient population that meets this second condition. Explain how the query works in the report document.
* **g.** Create an SQL View to represent this patient population, based on your query.
* **h.** Create an SQL query using the defined views to determine the patients that meet both criteria. How many are there? How many are females and how many are males? Also show the distribution of ethnicity / race.

**Step 2.** Use the baseline queries from step 1 to make the following analysis.  Record any supplemental queries that you use in the SQL file. Identify the top 15 diagnosis codes determined by physicians (columns Dx1 to Dx10) for those patients' last emergency room visit before pacemaker implantation. Summarize your results in the following tabular format. List the diagnosis code, diagnosis name (you can look up names using an ICD9 look up table), count how many times this diagnosis appears in the defined patient population, calculate the percentage of male and female patients with the diagnosis.

| *Exemple List of Top Emergency Room Diagnoses* | | | | |
| --- | --- | --- | --- | --- |
| **Diagnosis Code for ER Visit** | **Diagnosis Name** | **# Patients** | **% Male** | **% Female** |
| 9999 | Name of disease/injury | 90 | 40 | 60 |

**Step 3.** Use the baseline queries from step 1 to make the following analysis.  Record any supplemental queries that you use in the SQL file. Compare the top diagnoses between the two genders. Are there any particular diagnosis codes that occur much lower in male than in female patients? Answer the research question: Is it possible that male patients are overlooked (under-diagnosed) for some medical conditions in emergency room?  And why - support your answer.

**Step 4.** Use techniques discussed in class, such as indexing to try and improve performance in this application context. Give an example of how you improved the database query performance. Document (a) What does the query try to accomplish? (b) What did you do in order to improve the query performance? (c) Describe the performance improvement in terms of running time or etc.