CS-3200 Homework #3

Using SQL in a well-known yet rather messy database

Schema Description

This homework uses a subset of the WHO (World Health Organization) 2015 mortality database (MDB). It is an invaluable resource for population health researchers. The limited schema consists of 7 tables: country, population, deaths, icd9, admin1, subdiv1, and rlist. WHO has been tracking populations, deaths and cause of death since 1950. The data is typically collected at the country and reporting year level; however there are a few exceptions to this rule.

Cause of death is coded according to the International Classification of Disease (ICD). The coding schema has evolved over the last 50 years; the complete dataset has data representing cause of death in 10 different formats; we limit the homework database to a subset of ICD9 Basic. ICD9 was typically used between the years 1980 through 2000, however different countries adopted and abandoned the ICD9 coding schema in different years. Within the deaths table the cause of death counts within the database has a hierarchical rollup feature (pointing out yet another flaw in the data schema), please ignore this fact and assume each icd-9 code is independent from each other.

As you know, geography has also evolved over the last 50 years. Another evolved method is the stratification of statistics by different age bands as well as collection of data for sub-populations.

In porting this dataset to My SQL, my goal was to keep the variable names and the table names consistent with the WHO names, so that you could use the documentation provided by WHO. There are a few exceptions to that rule: 1) I made all table names and field names lower case 2) I renamed those fields that had names that conflicted with key words in My SQL.

NB – The queries are written such that the given cause of death is stated as a description as opposed to an ICD-9 code. Make sure your resultant set contains all causes of deaths associated with the specific disease (take advantage of the LIKE operator and regular expressions). Also, make sure you are returning sets as opposed to multi-sets for your answer (take advantage of the GROUP BY clause).

Purpose of Assignment

This assignment exposes you to the problem of the evolution of a data schema. Please be extremely critical of the given schema since you will be asked to propose changes to it.

This assignment also gives you a sense of the messiness existing in datasets that you may have to work with. You are being handed a somewhat ill-defined data schema and being asked to understand the schema. A few of the tables are not even part of the database schema and are only represented in the documentation. I created tables within the My SQL database for these tables (Admin1, Rlist, icd9, and SubDiv1 concepts), so you have a representation of this data within your schema. I recommend data exploratory of the tables to get a real sense of the schema and the problems associated with it. It is important you develop a critical eye when you are exposed to an ambiguously defined data schema.

Initial Step:

Download the WHO database from the class website. It can be downloaded in a compressed or an uncompressed format. The data expands to a megabyte so make sure you have enough disk space to accommodate it. Use the My SQL Workbench to import the database into your system. This is done from the server administration panel. Look for the task 'Data Import' in the left most blue tab labeled task and object browser.

Homework: Write SQL code to answer questions (1 -14):

1) How many people have died from cause 'B54'? (5 points)
2) What is the *description* of 'B54' that is listed in the icd9 table? (5 points)
3) How many tuples in the admin1 table do not have a description listed as = 'Country'? (5 points)
4) What are the names of the countries that have reported death counts using ICD9? (5 points)
5) What countries have reported deaths due to leprosy? (5 points)
6) How many people are coded as dying from AIDS ? (Remember not all countries represent their data in ICD9 or report data to WHO so the count will be lower than what you would expect; also the country table has statistics on subpopulations make sure you limit the results to countries) (5 points)
7) What countries have reported deaths attributed to AIDS? (5 points)
8) What was the first year that reported a death due to AIDS? (5 points)
9) How many people are coded (using ICD9) as dying from AIDS in the US (Unites States of America)? (5 points)
10) In what years did the US report deaths from AIDS using ICD9? (5 points)
11) In 1980, what is the most commonly coded cause of death for women in the United Kingdom? (5 points)
12) What percentage of the United Kingdom's female population died from the leading cause of death? (10 points)
13) What are the ten topmost reported cause of death in the United Kingdom for the year 1979? (10 points)
14) Determine the number of deaths associated with malignant neoplasms for each year that has reported a death associated with malignant neoplasms. Malignant neoplasm icd9 codes can be retrieved from the icd9 table where the description can be matched to '%MN OF %' (10 points)
15) The admin1 table and the subdiv1 table were introduced to allow reporting on subdivisions of a country's population? Was the concept of subpopulations properly introduced into the schema with these tables? Can you suggest an alternative method? (5 points)
16) Please name at least 5 ways you could improve the layout of this database. (5 points)
17) Currently there is no systematic representation of the merging and dividing of countries throughout the years. The schema does not even represent countries that are currently in existence versus defunct. Can you suggest a method to represent country evolution within the schema? Can you describe a method that would enforce statistical collection only on currently existing countries? (5 points)