Data Analyst Second Round Evaluation Project

Overview

Imagine you are building a claims analytics module from scratch to highlight certain key findings for the healthcare community and build a case for innovation in the current healthcare delivery system. To do so you will need to create a scalable model to track certain data. Correspondingly this project has two parts: "Part 1 – Data Skills" which is mandatory while "Part 2 – Statistical Skills" is optional. Below are some of the tools and data that you will need to utilize to complete this project.

Data

You will be using the CMS 2008-2010 Data Entrepreneurs Synthetic PUF files for this project. The files can be found at this link: https://www.cms.gov/Research-Statistics-Data-and-Systems/Downloadable-Public-Use-Files/SynPUFs/DESample01.

Translations for a lot of the common terms provided in the Beneficiary Summary file can be found at this link: https://www.cms.gov/files/document/de-10-codebook.pdf-0.

Tools

Part 1 of the project must be done MySQL, but Part 2 which is optional is flexible in terms of the tools you may use. For Part 1, use your preference in Query Editors. MySQL Workbench is a common free option which can be found at this link: https://www.mysql.com/products/workbench/.

Part 1 – Data Skills

This part will showcase your ability to think critically about data in a scalable fashion and perform basic operations within a database.

1. Architect the Database

You will need to create a scalable database infrastructure to house this data. As much possible, try to lay a foundation and structure that may scale well beyond the current project needs. EMR databases can be a good starting point for such work (although significant opportunity for improvement in EMR data structure exists). If you do not have healthcare experience and have not seen an EMR data structure before, you can use OpenEMR (an open-source EMR) as a reference. Their database structure can be found at this link: https://www.open-emr.org/wiki/index.php/Database_Structure

2. *Import the data*

Bring the data from the CMS 2008-2010 Data Entrepreneurs Synthetic PUF files into your database, according to your "proprietary" data architecture. Feel free to use as many of the claims types & files you'd like (more is more) but at a minimum you will need to use the below two files which span beneficiary demographic information and inpatient claim data:

- DE1.0 Sample 1 2010 Beneficiary Summary File (ZIP)
- DE1.0 Sample 1 2008-2010 Inpatient Claims (ZIP)

3. *Create a stored procedure*

Create a stored procedure that calculates the average expenditures per patient based on date ranges provided by the user (feel free to add other filter options if beneficial) and sliced by disease cohort. See the detailed requirements of this stored procedure below:

Calculated Metric: Average annual total expenditure per patient
Source: Inpatient claims (at a minimum)
Inputs: Date ranges (at a minimum)
Required Disease Cohorts:

- COPD
- CHF
- DIABETES

Within each Cohort there should be two subcategories:

1) Single Condition – patients where the only major condition listed is the one governing the cohort
2) Comorbid – patients that have the condition that is governing cohort but also have at least one other condition as defined by the Beneficiary Summary File

4. *Export the data*

Dump the entire database, including schema, data, stored procedure, and submit back the dump to us using any file share method of your choosing.

Part 2 –Statistical Skills (OPTIONAL)

In any tool of your using, conduct a multi-variate regression analysis to identify if any of the below factors have a significant impact in average expenditure per patient over a reporting period of your choice.

- Age
- Race
- Sex
- County

Include all program files, direction to use aforementioned program files, and summary of your findings, including commentary, as part of your final submission.