**Preparing Raw Data for Effective Healthcare Analytics – A Peer Reviewed Assignment**

Imagine that you are a data analyst working with a team of data scientists and statisticians for a large healthcare system called Acme Healthcare. This healthcare system includes numerous clinics and hospitals. Your mission is to provide analytical solutions to the executive leaders at Acme Healthcare to help them solve one of the following analytical problems:

* Providers at Acme Healthcare may vary with respect to quality and utilization. How can administrators adjust for patient-level risk factors?
* Administrators at Acme Healthcare want to identify providers or groups of providers that might have more adverse events. How can they adjust for patient-level risk factors?
* Some providers at Acme Healthcare may be engaging in fraud with respect to documentation and billing. How can they be identified after controlling for patient-level risk factors?

You will create and upload a one to two-page PDF report for the executive leaders of Acme Healthcare who are mandated to solve the problem you have selected. The report has multiple steps and will include a description of the problem area you want to focus on, the data, and how you might address the necessary challenges and possible solutions to the problem. Your report will include an Appendix to illustrate how you would modify the data dictionary and where you can put additional descriptive text or examples about how you plan to solve some of the complex ETL issues.

Here is a summary of the steps for this report, which build on each other and for which you will be graded:

1. Choose one of the analytical problems and suggest a possible analytical solutions.
2. Evaluate how groupers can help you solve aspects of the analytical problem. You can consider how to group diagnoses, procedures, and medication codes into analytical categories.
3. Create a one paragraph analytical plan about how you will solve the problem.
4. Answer questions about what ETL processes are required to create analytical file.
5. Create an appendix where you include suggestions for improvements to the data dictionary and summarize likely analytical output.

**Tools and Data**

The tools and data you will use for this assignment are:

* Excel
* Access to the already transformed CMS 2008-2010 Data Entrepreneurs’ Synthetic Public Use File (from lessons in Module 4)
* Optional (statistical software or various programming languages to transform and analyze the data)

**Step 1 – Summary of Analytical Problem Requiring Risk Adjustment**

Select one of the three topic analytical areas that could benefit from risk adjustment:

* Provider profiling for quality
* Risk adjusting patient safety indicators
* Provider profiling for fraud analysis

Within your report to Acme Healthcare administrators, address the following questions in a paragraph:

* Why did you choose the topic?
* How can the problem benefit from an analytical solution?
* Why risk adjustment is helpful or necessary?
* What general conceptual steps will be required to perform risk adjustment?

**Step 2 – Using Groupers to Prepare Analytic Datasets**

To prepare for your risk adjustment analysis, consider how you will group diagnoses, procedures and drugs into more manageable categories.

For this first part of the project you will review data files that contain grouper logic for the following systems:

* Healthcare Cost and Utilization Project (HCUP). (2016). [Clinical Classifications Software (CCS) for ICD-9-CM](https://www.hcup-us.ahrq.gov/toolssoftware/ccs/ccs.jsp).
* U.S. National Library of Medicine, National Institutes of Health. (2014). [Unified Medical Language System](http://www.nlm.nih.gov/research/umls/rxnorm/).
* University of California - San Diego. (Undated). [Chronic Illness and Disability Payment System](http://cdps.ucsd.edu/).
* [Berenson-Eggers Type of Service (BETOS) Codes](https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/MedicareFeeforSvcPartsAB/downloads/betosdesccodes.pdf)

For each file, address the following question:

* How can you aggregate many codes into a smaller number of analytical categories?

**Step 3 - Describe Analytical Plan**

Using the SEMMA methodology, describe how you will use the data sets provided to solve your analytical problem, here are some questions to consider in your description:

* Sample: Will you include all rows of the data?
* Explore: What descriptive analyses might you perform to learn about the data? How might this help you select fields to include in final analysis?
* Modify: Although you will go into details about this step in Part 4, briefly describe what data transformations might be required and why these are necessary.
* Model: Briefly consider some of your knowledge about data science and statistics to describe some possible methods used for risk adjustment.
  + Be clear in your discussion of datasets (e.g., rows, tables), and use concrete definitions of terms related to predictive modeling (e.g., structured vs. unstructured)
* Assess: Describe how you will assess your model and output.

**Step 4 – Creating an Analytical File**

Based on the lessons about how to perform risk-adjustment, the objective for this part of the project is to describe what types of data transformations and processing are queried to prepare the data for the risk adjustment analysis.

Address all 12 of the following questions in your response *(please note: some answers can be answered in one to two sentences, but for others you may need you to expand your answer to three to five sentences)*:

**Concepts, Fields, Groupers**

* What concepts are required to in the analysis?
* Which fields from the datasets will you select for each concept?
* Continuing from the earlier section about groupers, which grouper categories will you use?

**ETL**

* Which tables have multiple rows per patient?
* When you join data from the various tables, will your output include duplicates?
* In looking at the data dictionary and the data tables, do you see any need for mapping to more standard codes?
* Is there evidence that the data might vary through time, or by different regions/states?
* Would you consider conditional programming logic to recode data values?
* What type of aggregation of data might be helpful?
* Would it be helpful to select specific rows (filter)?
* Is there are need to transpose any fields?
* Are there temporal aspects of data related to dates that could cause problems?

**Step 5 – Appendix: Data Dictionary and Output Interpretation**

One of the most important parts of analytical projects is to have documentation about the source data so that the data science teams can produce reliable information. In addition, once the analytics are complete, the data scientist teams should explain how they transformed data and created their models.

You need to include the following into your appendix:

* Improve the rudimentary data dictionary that you worked on in the lessons for Module 4. For this assignment, create a sample data dictionary that has at least 5 fields (additional points available for more than 5).
  + Include fields that were not included in the original example provided by the instructor
  + Consider including derived fields that you might create for the analysis. For example, if you create a new variable that combines two variables, it would be important to describe how this was done. It is also important to describe fields created by groupers.
* Based on your analytical and modeling plan, summarize what types of output might be created for the risk adjustment analytics.