Objective

- This assignment uses a data source from the National Center for Biotechnology Information to perform an analysis that is similar to ones the Data Science Practice collaborates on with the Brown Center for Biomedical Informatics.
- The goal is to perform a *comorbidity analysis* for obesity, meaning an analysis of other/additional diseases that frequently co-occur with obesity.

Data Sources

- The primary data source you will use is the PubMed/MEDLINE records from NCBI, which contain metadata about publications in the biomedical literature (http://www.ncbi.nlm.nih.gov/books/NBK25501/).
- The records contain keywords called “Medical Subject Headings” (or MeSH descriptors) that summarize the content for each publication (https://www.nlm.nih.gov/mesh/).
- The MeSH descriptors are grouped into “Semantic Types” in this XML file: ftp://nlmpubs.nlm.nih.gov/online/mesh/2015/desc2015.xml (298MB).

Methods

Perform the following using programming tools of your choice:

1. Formulate PubMed/MEDLINE search for articles between 2000 and 2012 with obesity indicated as the major MeSH descriptor.
2. Obtain PubMed/MEDLINE records (in MEDLINE or XML format) for the formulated search using NCBI E-Utilities (https://www.ncbi.nlm.nih.gov/books/NBK25501/). Bindings are available in most programming languages, including Python and R.
3. Create mapping of MeSH descriptors to semantic types using the MeSH Vocabulary file (desc2015.xml) and identify descriptors with type “Disease or Syndrome.”
4. Using a statistical or modeling approach of your choice, identify and rank comorbidities for obesity based on publications that share MeSH descriptors for multiple diseases/syndromes.

Results

- Provide any source code used for retrieval, extraction, and analysis.
- Summarize your overall result for obesity commodities, as well as your methodology for each step above.
- You do not need to submit the raw data you downloaded from NCBI.