Readme file for healthcare referral graph database

All work was done on a Windows 7 machine. Any command line tools were used in Windows Command Prompt.

Data distribution processing:

Used Anaconda distribution of Python 2.7 (https://store.continuum.io/cshop/anaconda/).

Used Spyder IDE in Anaconda to write various Python scripts to perform small-scale changes including modifying headers, truncating whitespace, removing bad data, etc.
Sample python file: (https://www.dropbox.com/s/v9um31rju88vpqp/changeheader.py)

Used pip in Anaconda to install 2 useful packages to assist with processing.

massedit v.66 (https://github.com/elmotec/massedit): replace commas with tabs en masse for format requirements (see “Build”), replace strings.

csvkit v.1.0.0 (http://csvkit.readthedocs.org/en/latest/): toolbox of command line utilities to work with large CSV files. Cut relevant columns from base data with csvcut, join columns after modifications with csvjoin, find relevant data with csvgrep, check for erroneous data with csvsort.

All tools are free, massedit and csvkit are open-source. massedit and csvkit tools were used in Windows Command Prompt.

Build:

Database was built on a Windows 7 machine using the Neo4j platform (http://www.neo4j.com). Used Max de Marzi's batch-import tool 2.0 (binary download, https://github.com/jexp/batch-import/tree/20) to build database.

Using the Database:

After downloading, use the command line based Neo4j shell (http://www.neo4j.org/download/linux, documentation here: http://www.neo4j.org/develop/shell) to use it. One would need to be using a Linux OS to download it normally. Alternatively, request the Windows version from the developers.