

README FILE for Mesh Pair Metrics dataset
Neil R. Smalheiser and Gary Bonifield, February 2 2016

This dataset comes in the form of a compressed (.gz) file named meshPair_metrics.txt. The uncompressed size of the tab-delimited file should be ~16GB. These data are being released under the terms of the Creative Commons AttributionNonCommercial-ShareAlike CC BY-NC-SA International Public License 4.0. (<https://creativecommons.org/licenses/by-nc-sa/4.0/legalcode>).

How was the dataset created?

This dataset contains similarity metrics for relating any two Medical Subject Headings (MeSH terms) to each other. Article MeSH pairs were formed by creating all possible MeSH pairs for each article in the MEDLINE 2014 baseline dataset. Author MeSH pairs were formed by creating all possible MeSH pairs in the body of articles written by the same individual using the 2009 Author-ity author name disambiguation dataset. Only MeSH terms appearing in at least 25 articles were included in the dataset.

What metrics are in the dataset and how are they defined?

Each line in the data file corresponds to a MeSH pair identified by two MeSH terms and includes 5 article based metrics and 3 author based metrics. There are 201,866,854 MeSH pairs in the dataset and blank fields indicate that the MeSH pair was not present in either the article or author based datasets. The tab-delimited data file contains the following data columns:

MeSH Term 1. The first MeSH term making up the MeSH pair.

MeSH Term 2. The second MeSH term making up the MeSH pair.

Article Geometric Mean. The geometric mean of the number of individual article occurrences for MeSH term 1 and for MeSH term 2.

Article Count. The number of articles in which each MeSH term occurred.

Article Co-occurrences. The number of times the MeSH pair co-occurred across all articles.

Article Bin Avg. Average article co-occurrence score across all MeSH pairs in the same bin. Bins were formed by ranking all MeSH pairs in the dataset by article geometric mean and dividing into bins of 5000 pairs.

Article Odds Ratio. The observed article co-occurrence score divided by the average article co-occurrence score in the same bin.

Author Co-occurrences. The number of Author-ity 2009 author clusters in which the MeSH pair co-occurred.

Author Bin Avg. Average author co-occurrence score across all MeSH pairs in the same bin. Bins were formed by ranking all MeSH pairs in the dataset by article geometric mean and dividing into bins of 5000 pairs.

Author Odds Ratio. The observed author co-occurrence score divided by the average author co-occurrence score in the same bin.