Understanding Cardiovascular Health and Revealing Pathogenic Insights via Text-Mining Approaches

Peipei Ping, HeartBD2K Center at UCLA; *David Liem, HeartBD2K Center at UCLA; Doris Xin, KnowEnG Center; Quan Cao, HeartBD2K Center at UCLA; Vincent Kyi, HeartBD2K Center at UCLA; Leah Briscoe, HeartBD2K Center at UCLA; Karol Watson, HeartBD2K Center at UCLA; Alex Bui, HeartBD2K Center at UCLA; Jiawei Han, KnowEnG Center*

Over the past decades, mounting information on cardiovascular disease (CVD) from natural language in text is rapidly accumulating. Our pilot study demonstrates the feasibility of applying machine-learning and text-mining techniques on textual data in CVD groups to identify novel classifications, to facilitate predictive analytics, and to aid the clinical decision process. In our study, we applied a combination of phrase-mining algorithms and network-embedding techniques to 551,358 publications (dating from 1995 to 2016) in the Pubmed database, as well as the top-250 proteins that are highly relevant to the cause and treatment of CVD. Six CVD groups are studied, including Cerebrovascular Accidents (CVA), Cardiomyopathies and heart failure (CM), Ischemic Heart Diseases (IHD), Arrhythmias, Valve Dysfunction (VD), and Congenital Heart Disease (CHD).

The top 25 most relevant proteins in CM have a similar score pattern to both IHD and CVA, with the majority of proteins in both heart diseases revealing inflammatory function. Whereas, when we considered CVA and IHD as clustered diseases; VD, CHD, and Arrhythmias share little overlap with those top 25 proteins in CM. Furthermore, contractile protein, Titin, has a high relevance in CM and VD compared to the other CVDs. As expected, Troponin-I has a very high score in IHD. Moreover, Platelet-activating factor acetyl hydrolase, which is a mediator of many inflammatory functions, was identified as relevant for all 6 CVD groups.

Taken together, we demonstrate that a combination of phrase-mining algorithms and network-embedding techniques is effective to recognize hidden patterns underlying textual data contained in Pubmed literature. Novel insights are gained from characterizing these relationships among 250 proteins and 6 major CVDs, offering better understanding of CVD. We believe this new data acquisition strategy will be suitable to extract clinical relevant information from the vast amount of unstructured data in the public domain (e.g., Pubmed).