Abstract Text Mining to Create an Exhaustive Disease-disease Correlation Database

Misra, Suchir (School: Jericho High School)

Disease-disease correlations lend to improved treatment modalities, however, current disease-disease correlation databases fail to include poorly-studied diseases. Craniosynostosis, despite being the second most common craniofacial abnormality, is one such disease typically overlooked for disease-disease correlations. Text mining, specifically abstract text mining, may serve as a reliable process to establish disease-disease correlations. In this study, a computational approach was designed to create a disease-disease correlation database using solely abstract text mining. Python programs were written to collect abstract IDs for all genetic (i.e. terms related to genetics) papers (N = 2,056,144) to identify disease-disease associations. Abstracts were used to extract gene names. Disease-disease correlations were determined via gene overlaps. The top ten disease-disease connections overall and shared drugs were previously elucidated in literature, validating the effectiveness of the proposed algorithm and the power to improve disease treatment modalities through use of it. Of the top ten disease-disease connections for craniosynostosis, four were newly elucidated, illustrating the power of the database to find novel correlations for lesser studied diseases. This study created the most comprehensive disease-disease database available and the first to be created using solely abstract text mining. Future iterations of the database will include real-time search functions and automatic updates to the data collection. Physicians and researchers will both be able to use the database to design disease treatments for both rare and common diseases that lack viable treatment options today.