BIPOLAR DISORDER SENTENCE DATABASE AND LEXRANK SIMILARITY ASSESSMENT OF SENTENCES ON LITHIUM TREATMENT

Matthew N. Anyanwu¹, Tolulola O. Oyeleye², Wellington K. Ayensu³, Mehdi Pirooznia³ and Raphael D. Isokpehi²

¹Department of Computer Science, University of Memphis, Memphis TN 38152
²Center for Bioinformatics & Computational Biology, Jackson State University, Jackson, MS 39217, USA ³Department of Biological Sciences, University of Southern Mississippi, Hattiesburg MS

Abstract: The extraction of facts and information from unstructured natural language text such as PubMed abstracts is increasingly recognized as a crucial step for translational biomedical and behavioral research. Bipolar disorder (BPD) is a highly heritable, severe and chronic mental illness characterized by episodes of elation and high activity; alternating with periods of low mood and low energy. This condition is less prevalent but more persistent and more impairing than major depressive disorder (MDD). A search with the text "Bipolar Disorder" of the PubMed using the MiSearch Adaptive PubMed Tool revealed over 23,000 citations. We assume that these citations contain descriptors relevant to uncovering novel insights into various aspects of Bipolar Disorder. We have developed a software tool that splits the Title and Abstract text in Pubmed XML files into sentences. We implemented this software on a set of PubMed citations annotated with the MeSH term "Bipolar Disorder". A total of 125,988 sentences were obtained and can be queried using keywords and PubMed identifiers at http://compbio.jsums.edu/bpdsd. The database was developed to support a study on the Genetic Predisposition of African-American Women to Bipolar Disorder and Substance Abuse. Thus, we designed use cases to identify sentences and then abstracts that could guide further studies. Searches with the following keywords: female, genetic, woman, and women retrieved 1404, 2730, 267 and 1808 sentences respectively. These subsets of sentences were further analyzed for co-occurrence of descriptors such as alcohol, comorbid and substance. We used LexRank (a National Center for Integrative Biomedical Informatics [NCIBI] Tool) to determine the similarity of sentences that were from published bipolar disorder case reports on women that included lithium treatment. Filtering the LexRank Graph with the Cosine and Salience measures resulted in five significantly similar sentences including those from abstracts on case reports of side-effects of long-term lithium treatment in women (PMID:15730030; PMID:11050737). Further work will include improving the sentence splitting tool that we have developed and implementing LexRank algorithm to summarize results from queries.

Keywords: Bipolar Disorder, software tool, sentences, Pubmed identifiers, NCIBI, Genetic Predisposition, African-American Women.

Acknowledgments: The Research Centers in Minority Institutions (RCMI) – Center for Environmental Health at Jackson State University (NIH-NCCR 2G12RR013459-11); National Center for Integrative Biomedical Informatics (NIH U54DA021519); Mississippi NSF-EPSCoR (NSF-EPS-0556308).