Identifying natural health product and dietary supplement information within adverse event reporting systems

Vivekanand Sharma and Indra Neil Sarkar

Center for Biomedical Informatics, Brown University
Providence, RI 02912, USA
Email: vivekanand_sharma@brown.edu and neil_sarkar@brown.edu

Data on safety and efficacy issues associated with natural health products and dietary supplements (NHP&S) remains largely cloistered within domain specific databases or embedded within general biomedical data sources. A major challenge in leveraging analytic approaches on such data is due to the inefficient ability to retrieve relevant data, which includes a general lack of interoperability among related sources. This study developed a thesaurus of NHP&S ingredient terms that can be used by existing biomedical natural language processing (NLP) tools for extracting information of interest. This process was evaluated relative to intervention name strings sampled from the United States Food and Drug Administration Adverse Event Reporting System (FAERS). A use case was used to demonstrate the potential to utilize FAERS for monitoring NHP&S adverse events. The results from this study provide insights on approaches for identifying additional knowledge from extant repositories of knowledge, and potentially as information that can be included into larger curation efforts.

Keywords: Natural Health Products; Dietary and Herbal Supplements; Adverse Event Detection; Terminology Mapping; Natural Language Processing.

1. Introduction

The biomedical community has benefitted from continuous development and improvement of automated methods for knowledge acquisition from heterogeneous data sources. A fundamental requirement for such tasks includes identification of entities of interest and their resolution to standard terminologies\(^1\). The process of converting unstructured free text fields from data into structured format creates opportunities to attain actionable knowledge by designing analytic enquiries. The heterogeneity of data from different sources poses challenges when seeking to perform comprehensive, multi-source analyses. Previous studies have demonstrated the utility of interlinked data from multiple sources to identify potential new knowledge\(^2\)\(^-\)\(^4\). The growing amounts of biomedical data from multiple sources suggest that an essential prerequisite for biomedical knowledge discovery will be the potential to leverage terminology resources for facilitating efficient indexing and subsequent retrieval. The biomedical domain is equipped with standard vocabularies from several sources that are used to facilitate access, retrieval and analysis of data from disparate data and knowledge sources. For example, the Unified Medical Language System (UMLS) Metathesaurus, maintained by the National Library of Medicine (NLM), is a repository of over one million biomedical concepts from more than 100 sources\(^5\).

To support standardization and integration of available information about drugs and health related outcomes, the Observational Health Data Sciences and Informatics (OHDSI) workgroup was established with the goal of developing an open-source standardized knowledge base\(^6\). The