Identifying Rare Disease Studies by Natural Language Processing

Huan He, PhD[1], Feichen Shen, PhD[1], Michael Lin[2], Hongfang Liu, PhD[1]
[1] Division of Digital Health Sciences, Mayo Clinic, Rochester, MN, USA;
[2] Center for Clinical and Translational Science, Mayo Clinic, Rochester, MN, USA

Introduction: Developing treatments for rare diseases is challenging due to the insufficient evidence and poor access to innovative drugs and technologies[1], additionally, there is limited mechanisms to screen related studies from massive information resources such as clinicaltrials.gov and PubMed. Most existing search engines provide complex parameters to filter rare disease query results, and requires users to have sufficient search skills and specific knowledge. To address this issue, this study aims to promote the exploration of rare disease studies in existing study repositories by using natural language processing and data visualization techniques.

Materials and Methods: We identified the relevant information needed for identifying rare disease studies and extracted the free text information of research proposals in our institutional review board (IRB) system through a research data warehouse. This information includes the study titles, abstract, the summary of methods and disease/condition coding. A total of 11,690 active studies conducted in Mayo Clinic during the past 10 years were targeted. Among them, 3,247 studies had their diseases and conditions being studies annotated manually by domain experts in SNOMED terms for both our research and internal use, while the rest 8,402 studies remain to be annotated.

To identify whether a study is related to rare diseases using unstructured text data, we designed an ensemble model that includes a text search model based on keyword matching and a concept search model based on rare disease ontology (as shown in Figure 1(a)). In the text search model, we use regular expression technique to query rare disease terms based on a keyword list from the Genetic and Rare Diseases (GARD) information center and the Orphanet. In the concept search model, we first extracted medical concepts from the texts of studies by MetaMap[2], and then searched these concepts in a "rare disease – medical concept" dictionary based on the rare disease ontology from Orphanet. Moreover, we designed a visual exploration tool for interpreting the identification results.

Results: Evaluation on the annotated dataset shows that the text search model achieve higher performance (Precision: 74.26%, Recall: 60.48%, F1 Score: 66.67%) than the concept search model (Precision: 76.55%, Recall: 32.42%, F1 Score: 45.55%), and the ensemble model achieves overall better performance with a minor decrease in precision (Precision: 73.11%, Recall: 64.41%, F1 Score: 68.49%). With the ensemble model, we identified 1,912 rare disease studies out of the 8,402 and visualized the results in an interactive interface (as shown in Figure 1(b)).

Conclusions: The ensemble model could identify rare disease studies from unstructured clinical text and the visual design could improve the interpretability of the identification results, demonstrating the feasibility to translate the developed approach to internal usage on identifying rare disease studies.

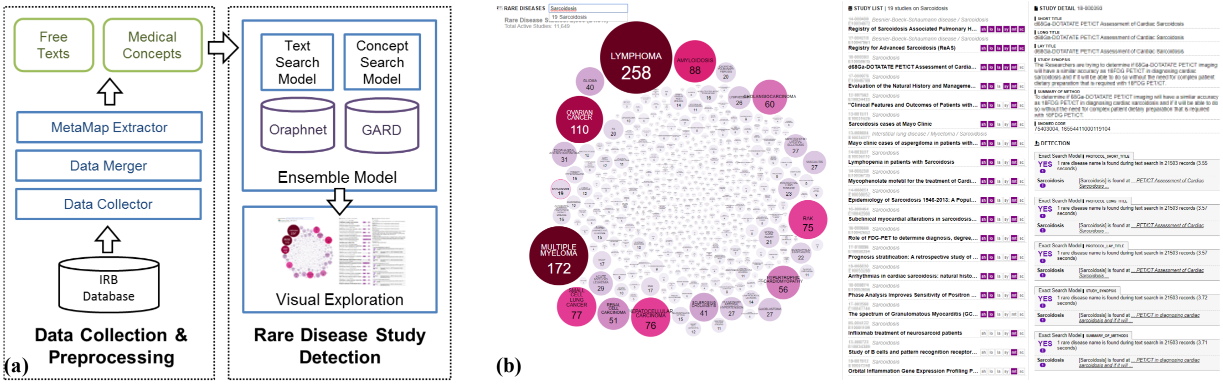

Figure 1. (a) The data processing procedure and (b) the screenshot of the visual exploration tool.

References:

1. Bogaerts J, Sydes MR, Keat N, et al. Clinical trial designs for rare diseases: Studies developed and discussed by the International Rare Cancers Initiative. *Eur J Cancer*. 2015;51(3):271-281.
2. Aronson AR, Lang F-M. An overview of MetaMap: historical perspective and recent advances. *J Am Med Inform Assoc*. 2010;17(3):229-236.