Temporal Pattern Discovery and Risk Stratification for Patients with Cancer

Saeed Mehrabi, Mathew Palakal (Thesis Advisor)
School of Informatics, Indiana University, Indianapolis

I. INTRODUCTION

Most of the Clinical decision support systems (CDSS) are rule based [1] which requires regular maintenance and frequent updates of the knowledge base [2]. Data driven CDSS where the information is derived from decisions made previously in similar patients, are an opportunity to remove this barrier. Most of the clinical observation within the electronic health records (EHR) are recorded in free text format. Rapid adoption of EHRs and satisfying its meaningful use criteria is strong incentives for development of Natural language processing (NLP) methods to extract the clinical knowledge embedded within these free texts to be used by CDSS [3]. The goal of this research is to contribute to physicians’ ability to offer high quality care by mining the existing data of previously treated similar patients and providing personalized decision support for current patients based on their unique characteristics. Experiments will be carried out on pancreatic cyst patients’ data to evaluate our methodologies.

II. PROPOSED METHODOLOGY

A. NLP Challenges

EHR contains valuable longitudinal clinical information that can be used for various applications such as CDSS, however these applications are not readily feasible because much of the information are in free text format. Extracting clinical concepts using NLP is not an easy task because modifiers such as negation and family history section significantly affect the meaning of a concept. Finding the scope of negation terms is one of the challenges in negation algorithms, for instance conjunctions such as “but” and “however” restrict the scope of negation terms. Dependency parser can be used to find the relation between negation terms and desired concepts. Clinical notes are organized into sections with section headers such as family history. In order to detect section headers, we will use a two-step approach where sequential tagging algorithm will be used to identify section boundaries and machine learning algorithm to label each section with its appropriate section category. Once the section is detected, family members and diagnosis and the relation between them will be identified.

B. Risk Stratification and predicting patients’ prognosis based on similar patients temporal patterns

Longitudinal patient health records consist of phenotypic information, disease characteristics, treatment and outcome. Analysis of this time varying physiological data may reveal patterns of disease progression (cancer stages) and treatment response. Not every two patients with the same disease have the same pattern of progression, treatment (based on patients’ condition and preferences) and outcome. These variations can be due to genetic profiles of patients, patient life style and environmental factors. EHR would be a valuable source containing patient phenotypic data, their treatment and results over the course of disease management. Predictive models including temporal data mining and dynamic Bayesian networks will be developed to analyze disease progression pattern, predict patient outcome and finally guide treatment.

C. Experimental evaluation on patients with pancreatic cysts

Pancreatic cancer is the fourth leading cause of cancer death in the US [5]. Screening the general population for pancreatic cancer is not feasible because of its low incidence. Pancreatic cyst and family history of pancreatic cancer represent two windows of opportunity for early detection of pancreatic cancer. NLP methods will be used to identify these patients. Natural history of pancreatic cysts and why some cysts progress to cancer is not very well known. Longitudinal information on previous patients can be helpful in pattern discovery and outcome prediction of new patients and eventually guiding their treatment and increasing their chance of survival.

III. PROGRESS/RESULTS

NLP methods were developed with improvement over NegEx algorithm using dependency parser in an Unstructured Information Management Architecture (UIMA) framework to identify patients with pancreatic cyst with 98.5% precision and 97.43% recall [4].

ACKNOWLEDGMENT

THIS WORK WAS SUPPORTED IN PART BY THE AGENCY FOR HEALTHCARE RESEARCH AND QUALITY R01 HS19818-01 AND BY A GRANT FROM THE OFFICE OF THE VICE PRESIDENT FOR RESEARCH AT IUPUI.

REFERENCES