Pancreatic Cancer Risk Stratification based on Patient Family History

Anand Krishnan [1], C. Max Schmidt[2], Alexandra M Roch[2], Chris Beesley[3], Saeed Mehrabi[1], Joe Kesterson[3], Paul Dexter[3], Mohammed A. Al-Haddad[4], Mathew Palakal[1]

[1]*School of Informatics, Indiana University, Indianapolis, IN, USA;* [2]*Department of Surgery, Indiana University, Indianapolis, IN, USA;* [3]*Regenstrief Institute, Indianapolis, IN, USA;* [4]*Department of Medicine, Division of Gastroenterology, Indiana University, Indianapolis, IN USA.*

Background: Pancreatic cancer is the fourth leading cause of cancer-related deaths in the US with an annual death rate approximating the incidence (38,460 and 45,220 respectively according to 2013 American Cancer Society). Due to delayed diagnosis, only 8% of patients are amenable to surgical resection, resulting in a 5-year survival rate of less than 6%. Screening the general population for pancreatic cancer is not feasible because of its low incidence (12.1 per 100,000 per year) and the lack of accurate screening tools. However, patients with an inherited predisposition to pancreatic cancer would benefit from selective screening. **Methods**: Clinical notes of patients from Indiana University (IU) Hospitals were used in this study. A Natural Language Processing (NLP) system based on the Unstructured Information Management Architecture framework was developed to process the family history data and extract pancreatic cancer information. This was performed through a series of NLP processes including report separation, section separation, sentence detection and keyword extraction. The family members and their corresponding diseases were extracted using regular expressions. The Stanford dependency parser was used to accurately link the family member and their diseases. Negation analysis was done using the NegEx algorithm. PancPro risk-prediction software was used to assess the lifetime risk scores of pancreatic cancer for each patient according to his/her family history. A decision tree was constructed based on these scores. **Results**: A corpus of 2000 reports of patients at IU Hospitals from 1990 to 2012 was collected. The family history section was present in 249 of these reports containing 463 sentences. The system was able to identify 222 reports (accuracy 87.5%) and 458 sentences (accuracy 91.36%). **Conclusion**: The family history risk score will be used for patients' pancreatic cancer risk stratification, thus contributing to selective screening.

Mentors: Mathew Palakal, School of Informatics, IUPUI; C. Max Schmidt, Department of Surgery, Indiana University.