DOI:10.1111/hpb.12375

ORIGINAL ARTICLE

Automated pancreatic cyst screening using natural language processing: a new tool in the early detection of pancreatic cancer

Alexandra M. Roch[1], Saeed Mehrabi[2,3], Anand Krishnan[2], Heidi E. Schmidt[1], Joseph Kesterson[4], Chris Beesley[4], Paul R. Dexter[4], Mathew Palakal[2] & C. Max Schmidt[1]

[1]Department of Surgery, [4]Regenstrief Institute, Indiana University School of Medicine, [2]School of Informatics and Computing, Indiana University, Indianapolis, IN, and [3]Divisions of Biomedical Statistics and Informatics, Mayo Clinic, Rochester, MN, USA

Abstract

Introduction: As many as 3% of computed tomography (CT) scans detect pancreatic cysts. Because pancreatic cysts are incidental, ubiquitous and poorly understood, follow-up is often not performed. Pancreatic cysts may have a significant malignant potential and their identification represents a 'window of opportunity' for the early detection of pancreatic cancer. The purpose of this study was to implement an automated Natural Language Processing (NLP)-based pancreatic cyst identification system.

Method: A multidisciplinary team was assembled. NLP-based identification algorithms were developed based on key words commonly used by physicians to describe pancreatic cysts and programmed for automated search of electronic medical records. A pilot study was conducted prospectively in a single institution.

Results: From March to September 2013, 566 233 reports belonging to 50 669 patients were analysed. The mean number of patients reported with a pancreatic cyst was 88/month (range 78–98). The mean sensitivity and specificity were 99.9% and 98.8%, respectively.

Conclusion: NLP is an effective tool to automatically identify patients with pancreatic cysts based on electronic medical records (EMR). This highly accurate system can help capture patients 'at-risk' of pancreatic cancer in a registry.

Received 28 April 2014; accepted 12 November 2014

Correspondence

C. Max Schmidt, Surgery, Biochemistry and Molecular Biology, Director, IU Health Pancreatic Cyst & Cancer Early Detection Center, 980 West Walnut Street C522, Indianapolis, IN 46202, USA. Tel.: +1 317 278 8349. Fax: +1 317 278 4897. E-mail: maxschmi@iupui.edu

Introduction

With an annual death rate approximating the incidence, pancreatic adenocarcinoma has been termed the 'deadliest cancer'. It is the fourth leading cause of cancer mortality in the United States with an annual incidence of 43 920 and death rate of 37 390.[1] In spite of a marked improvement in cancer care over the past several decades, the 5-year survival associated with pancreatic adenocarcinoma has changed little, rising from 3% in the 1970s to 6% in 2013.[1] Pancreatic adenocarcinoma is still diagnosed in more than 80% cases at an advanced stage where available systemic therapies remain largely ineffective.

While there is a significant pursuit of novel treatments targeted at established pancreatic cancer, the collective research effort on pancreatic cancer early detection and prevention (aside from general smoking cessation and physical fitness programs) is relatively small. Unlike colon, breast and prostate cancer, screening the general population for pancreatic cancer is not feasible owing to its low incidence (12.2/100 000/year) and the lack of effective screening tests to identify patients at earlier stages of the disease.[1,2] Pancreatic cancer screening may be applicable, however, only in select groups of patients with a higher risk of pancreatic cancer.

Funding sources: This work was supported in part by the Agency for Healthcare Research and Quality R01 HS19818-01 and a grant from the Office of the Vice President for Research at Indiana University-Purdue University Indianapolis.

This study has been presented on a Best Oral session at the 11th World IHPBA Congress, 22–27 March 2014, Seoul.

Patients at higher risk of pancreatic cancer include those with pancreatic cysts and/or those with a strong family history of pancreatic cancer. Both of these higher risk groups represent potential windows of opportunity for pancreatic cancer early detection and prevention. Pancreatic cysts, especially mucinous types including intraductal papillary mucinous neoplasms (IPMN) and mucinous cystic neoplasms (MCN), harbour a malignancy in 20% to 90% of patients undergoing a pancreatic resection.[3,4] Hereditary or familial pancreatic cancer is estimated to be the principal aetiology in 10% of pancreatic cancers.[5] The Johns Hopkins Hospital established in 1994 the National Familial Pancreas Tumor Registry as a research registry of families with more than one first-degree relative diagnosed with pancreatic cancer.[6] However, to our knowledge, no similar effort has focused on patients with pancreatic cysts.

The incidence of pancreatic cysts ranges from 2.6% in computed-tomography (CT) studies[7] to 19.6% in magnetic resonance imaging (MRI) studies[8] and up to 24.3% in a Japanese autopsy study.[9] Although most pancreatic cystic lesions do not require surgical resection, a recent review of 19 studies of mostly surgical series from 1997 to 2011, including 1060 patients with indeterminate pancreatic cystic lesions and final pathology found that 41.7% of them were malignant/ aggressive.[10] Considering the high incidence of pancreatic cysts, radiologists have established imaging recommendations to better guide their management.[11] However, these recommendations do not factor in main pancreatic duct dilation, which may be a manifestation of main-duct involved intraductal papillary mucinous neoplasm, a high-risk lesion. Furthermore, the imaging recommendations are based on cyst size (which is a less reliable criteria than previously thought) and the ability of cross-sectional imaging studies to correctly diagnose the cyst (in spite of the low reported accuracy of <50% of CT/ MRI for a specific diagnosis and the 15–20% cysts with crossover morphology[11,12]). In light of these limitations and because pancreatic cysts are often asymptomatic and incidentally detected, many pancreatic cystic lesions are ignored and never evaluated by a pancreatologist (surgeon or gastroenterologist).

With an increasing adoption of electronic medical records (EMR) systems by medical centres, more data from the patients' charts are becoming electronic and thus available for computational processing. However, in contrast to numerical data (such as laboratory values or blood pressure readings), data in medical documents is narrative free text, and thus, unstructured and not amenable to computerized applications. Natural language processing (NLP) is the formulation and investigation of computer-effective mechanisms for communication through natural language.[13,14] It allows computers to 'understand' natural language (i.e. the language humans use to communicate) by opposition to 'artificial' language used by computers. NLP allows automation and prospective tracking and is already used in hospitals for bio surveillance and quality measures by tracking adverse events.[15]

The aim of this study was to automatically identify patients with pancreatic cysts through EMR using NLP. Once feasibility is established, the plans are to track the patients, notify their primary care providers of their patient's condition and provide resources for medical decision making. The objective of this work is to optimize the management of pancreatic cysts, and ultimately, early detection and prevention of pancreatic cancer. In addition, through these efforts, we seek to create a patient registry to help improve the current knowledge of the malignant potential and natural history of pancreatic cysts.

Methods

Population

From March 2013 to September 2013, we conducted a prospective pilot study at a single medical centre (Wishard Memorial Hospital). Wishard Memorial Hospital is a 340-hospital bed institution located in a major city. Longitudinal EMR of all patients who visited this institution over the 7-month timeframe were retrieved from the Indianapolis Network for Patient Care (INPC, 94 hospitals including teaching hospitals, 110 clinics and surgery centres and other healthcare organizations within the state of Indiana).[16] Longitudinal EMR included all types of clinical, radiological, surgical and pathological narrative reports. Data were analysed on batches of monthly patients. The multidisciplinary team in charge of this pilot study included informaticians, hospital administrators, pancreatologists and pancreatic surgeons.

Data were collected and reported in strict compliance with patient confidentiality guidelines as defined by the Indiana University Institutional Review Board.

Natural language processing

'Pancreatic cyst' concept

A list of keywords and acronyms to define the concept of a 'pancreatic cyst' was created after a literature review and United States National Library of Medicine (National Institute of Health) Unified Medical Language System (UMLS) review.[17] Manual analysis of clinical reports was also performed to determine commonly used 'pancreatic cyst' descriptors. The final assembled list of 'pancreatic cyst' concepts was used in the NLP software for the identification of patents with a pancreatic cyst. The extraction process was first performed on a training set (obtained after randomization) to confirm relevant concepts, exclude irrelevant concepts and finally identify additional/missed concepts, thus improving the initial keywords list. The final list of keywords used by the query and their different patterns ('regular expression') and abbreviations are presented in Table 1.

Extraction process

An Unstructured Information Management Architecture (UIMA) framework was used for our NLP system development. UIMA is a platform that facilitates the implementation of multiple NLP tasks in a pipeline manner where each component's output will be used as the input to the next step/component.[18]

A rule-based algorithm was created to automatically identify 'pancreatic cyst' findings in the free text of electronic medical