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 cross-over 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

Table 1 List of keywords used to define the 'pancreatic cyst' concept by the natural language processing software

Concept	Regular expression
Pancreatic cyst	(?i)(pancreatic cyst(s)? | cyst(s)?(of| in)? the pancreas | pancreatic cystic)
Pancreatic pseudocyst	(?i)(pseudo\s?cyst(s)?)
Mucinous cyst	(?i)(mucinous cyst(ic | ts) neoplasm | mucinous cystadenoma | intraductal papillary mucinous | \b(MCN)\b | \b(MCA)\b | \b(IPMN)|\b(IPMT)\b)
Serous cyst	(?i)(serous cyst(ic | s | adenoma) | \b(SCA)\b)
Retention cyst	(?i)(retention cyst(ic | s))
Cystic neuroendocrine tumour	(?i)(cystic neuroendocrine tumor | cystic neuroendocrine | neuroendocrine cyst(ic|ts)|islet cell cyst tumor|cystic islet cell tumor)
Cystic degeneration cancer	(?i)(cystic degeneration cancer | cystic degeneration | degeneration cyst(ic | s))
Duct ectasia	(?i)(duct(al) ectasia | ectasia of the(pancreatic)? duct | ectasic duct)
Duct dilation	(?i)(pancreatic duct(al) dila(ta)tion | dila(ta)tion of the(pancreatic)? duct | dilated(pancreatic)? duct)

IPMN, intraductal papillary mucinous neoplasm; IPMT, intraductal papillary mucinous tumour; MCA, mucinous cystadenoma; MCN, mucinous cystic neoplasm; SCA, serous cystadenoma.

records. The system input was a text file containing all the reports for all the patients. Therefore, the first step was to separate each report using a report separator that identified the beginning and end of a report. In the next step a Metadata annotator was built to extract Meta information from each report such as medical record number, report type, number, name, date and body of report. In the next section, a sentence detector was used to identify each sentence within the report's body. Once the sentences were identified, the algorithm parses each sentence for at least one of the keywords. A study on negation has shown that most clinical observations are negated in narrative clinical reports, as physicians often record if a condition is absent or ruled out.[19] We developed a novel negation algorithm called DEEPEN (Dependency parser negation),[20] that uses dependency parser[21]and a set of extensive rules built on top of it to determine whether concepts were affirmed or negated. Dependency is a binary asymmetric relation between a token (word or other group of characters) and its dependents in one sentence.

The initial set of regular expressions based on the concepts reported in Table 1 was applied to the training set. The false positives and false negative cases were analysed and required modification. This process was repeated several times until we reached the target precision and recall on the training set. At the final step the algorithm was applied to the test set for evaluation.

Validation

The validation was performed manually by a group of physician experts in pancreatology who were blinded to the results of the NLP system. All cases were individually validated by review of medical reports. Findings from the algorithm were compared to the manual review, which was considered the 'Gold Standard'. Any discrepancy was re-evaluated by the entire team. The validation allowed evaluation of the performance characteristics of our NLP 'pancreatic cyst' identification system (true positives, true negatives, false negatives, false positives, sensitivity and specificity). By pointing out the algorithm issues leading to false positives and false negatives, the manual validation allowed improvement of the initial computer algorithm accuracy. The algorithm was updated every month to refine it and get closer to the gold standard, and the evaluation provided in this study reflects the performance of the final version of the algorithm (final version that the team re-run at once on every monthly batch of patients during the testing phase). No machine learning and Bayesian methodology were used in the development of the identification algorithm. However, DEEPEN used dependency parser which itself is a probabilistic model that finds relations between words within a sentence.

Statistical analysis

Data compilation was performed using Microsoft Excel 2011® (Redmond, WA, USA). NLP programming was conducted using Java. Descriptive statistics included numbers and percentages.

True positives were defined as patients positively identified by both NLP algorithm and manual review. True negatives were defined as patients with a concept negated by both NLP algorithm and manual review. False positives were defined as patients positively identified by the NLP algorithm but not confirmed by manual review. Conversely, false negatives were patients with a concept negated by NLP algorithm but positively identified on manual review.

Sensitivity and specificity were calculated using accepted definitions. The sensitivity of our identification system was defined as the proportion of patients with 'pancreatic cyst' and/or 'duct dilation' correctly identified by the NLP system compared with the actual number of patients that manual review confirmed as containing the concept of 'pancreatic cyst' (true positives and false negatives). Similarly, the specificity was defined as the proportion of patients without a 'pancreatic cyst' or 'ductal dilation' correctly identified and excluded by our system compared with the total number of patients that manual review confirmed as not having the concept of a 'pancreatic cyst' (true negatives and false positives). The results were reported according to the Standards for Reporting of Diagnostic Accuracy (STARD) recommendations.[22]

Results

Identification

There was no limitation in the type of records analysed, and both outpatients and inpatients were included in our population. The specific inpatient service or outpatient department/division where patient care was provided at the time of diagnosis of a pancreatic cyst was not one of the features extracted during the automated

Table 2 Monthly data results (number) for pancreatic cyst identification

Month	Clinical records	Patients	Identified patients[a]	True positives	True negatives	False positives	False negatives
March	97535	7950	227	98	128	1	0
April	78451	6419	199	93	106	0	0
May	70101	6036	186	83	102	1	0
June	78110	7514	165	79	85	1	0
July	81991	7390	196	97	97	1	1
August	79072	7534	197	86	110	1	0
September	80973	7826	186	78	104	4	0

[a]'Identified patients' represents patients with at least one mention of a pancreatic cyst or pancreatic ductal dilation in their electronic medical record.

screening process. A total of 566 233 reports, including 50 669 unique patients, were identified during the 7-month period. Of them, 1359 had at least one mention of a 'pancreatic cyst'. Three of those identified patients were further excluded ($n = 1356$) because they were detected twice (one patient identified in June had previously been identified in March and two patients identified in August and September had previously been identified in July and August, respectively). When considering imaging studies alone, a 'pancreatic cyst' was originally mentioned on a CT in 84.9%, ultrasonography in 9.5% and MRI in 5.6% of patients. The computer algorithm identified 623 positive (patients with pancreatic cyst or pancreatic ductal dilation) patients (Table 2). Manual (physician expert) review identified 615 positive patients (nine and one patients were found to be false positives and false negative, respectively). This resulted in a calculated prevalence of pancreatic cysts of 1.2%. The false positives and false negatives reports from the final query are provided and explained in Table 3. The most common cause of false positives was a complex sentence structure that caused the NLP process to fail. As the negation algorithm (DEEPEN) was not based on words location within a sentence, but on the sentence structure and suspected relation between words, even simple errors to the eye can have led to system errors (false positives) (Table 3). They usually included sentences with multiple negated findings where the algorithm did not identify an association between the negation term and the 'pancreatic cyst' concept. There was only 1 false negative in 7 months and it was due to a dictation error.

System performance

Over the 7-month period, the mean sensitivity of the NLP algorithm for identification of a pancreatic cyst and/or ductal dilation was 99.85% (range 98.98–100). Similarly, the mean specificity was 98.8% (range 96.3–100) (Fig. 1). The analysis was pursued beyond the pilot study period and the results are shown in the dotted line on Fig. 1.

Implementation

Our system was implemented on the hospital (Wishard) server to obtain real-time tracking. It was run daily, and the process was executed every night. Depending on the number of daily records, it took 2 to 3 h to complete the run. The medical record numbers of patients identified as true positives by the algorithm were then recorded in a database. An intermediate step of manual triage was then required, mainly to exclude patients with a benign lesion (pseudocyst and serous cysts). An application to alert the primary care provider and/or the ordering physician is currently being developed and should be installed soon. This represents the second phase of our project. Once the alert system is implemented, clinical impact and change in practice will be analysed.

Discussion

Patient care and clinical research up until the present are largely based on retrospective or prospective collection of data into databases, which is often done manually. The increased utilization of EMR by medical centres has created new patient care and clinical research possibilities. Through automated tracking of patient data, EMR increases the research scope (data volume, time) and statistical power while decreasing the required manpower utilization. The present pilot study, over a 7-month period, demonstrates that it is feasible and inexpensive to automate the identification of patients with pancreatic cyst(s) and/or pancreatic ductal dilation using natural language processing (NLP). Our algorithm allowed tracking of those patients with high sensitivity (99.9%) and specificity (98.8%). Although manual review remains an important part of the study, patient capture is easier, faster and more thorough when employing a NLP algorithm. This has been demonstrated in previously published work from our team.[23]

Current strategies for early diagnosis of pancreatic cancer have focused on serum biomarkers. The most commonly used biomarker is serum carbohydrate antigen 19-9 (CA19-9). Its use as a screening tool in the general population, however, would be suboptimal because of its low sensitivity (median 79%, range 70–90%) and specificity (median 82%, range 68–91%).[24] Small studies have suggested that the use of a combination of biomarkers instead of an individual biomarker may improve sensitivity and specificity.[25] Given the high genetic heterogeneity of pancreatic adenocarcinoma,[26] such may require a large number of biomarkers and is not likely to be routinely utilized soon. Simi-

Table 3 Description and explanation of the false positives ($n = 9$) and false negative ($n = 1$) from the final query

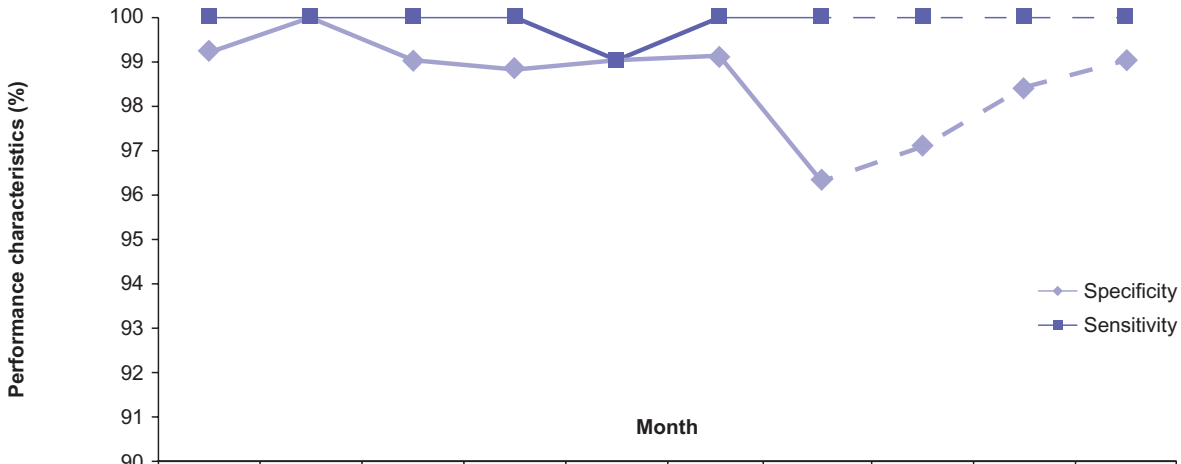
False positives ($n = 9$)	
Sentence in report	**Explanation**
Specific MRCP sequences were not included in this study however there is concern as was noted on prior CT scan for pancreatic divisum however no significant **pancreatic ductal dilation** is noted.	Complex sentence structure: multiple negation terms in the same sentence
There is homogenous enhancement of the pancreas without surrounding inflammatory changes or **pancreatic ductal dilatation**.	Complex sentence structure: algorithm did not identify an association between 'without' and 'pancreatic ductal dilatation'
Linear calcific density measuring 10 x 3 mm (axial image 43) is present within the pancreatic body, unchanged since prior exam, unlikely to be present within the pancreatic duct given no **pancreatic ductal dilatation**.	Complex sentence structure: sentence structure that causes algorithm pre-processing to fail.
Abdominal CT scan on 11/94 showed a large pancreas, no abscess, **pseudocyst** or phlegmon and she retained some contrast in the gallbladder.	Complex sentence structure: confusing sentence formation.
There is a vague area of decreased attenuation seen within the body of the pancreas no evidence of **pancreatic ductal dilation** or peripancreatic inflammation.	Missing punctuation: missing of a period after 'pancreas'
No evidence of complicating features or **pseudocyst** formation.	Complex sentence structure
Peripancreatic inflammatory changes with no appreciated pancreatic mass, **pseudocyst** or calcifications.	Complex sentence structure: algorithm did not identify an association between 'no' and 'pseudocyst'.
Transcribed by – PSC Transcription Date – < Date> RADIOLOGY IMPRESSION CT Abdomen and Pelvis: Acute pancreatitis, improved compared to prior examination of May 2005, without evidence of necrosis, **pseudocyst**, or abscess.	Punctuation error: presence of special characters such as colons gives a different structure to this sentence
Previously seen **pseudocyst** in the neck of the pancreas is no longer appreciated.	Complex sentence structure: algorithm identified an association between 'no' and 'appreciated' but not with 'pseudocyst'
False negatives ($n = 1$)	
Sentence in report	**Explanation**
There is a fluid collection near the fundus of the stomach, without definite wall indicates **pseudocyst.**	Complex sentence structure and dictation error: one word missing before 'indicates'

larly, cross-sectional imaging studies have low performance characteristics for screening pancreatic cancer in the general population.[27] Beyond the identification and tracking of patients with pancreatic cysts, our system sets ground for improved pancreatic cancer screening. First and foremost, our system identifies patients with pancreatic cysts with a very high accuracy. Pancreatic cysts, especially mucinous cysts, are well-established precancerous lesions. As they have the potential to develop into invasive adenocarcinoma in a median of 5 years (range 2–20),[3,4] tracking them closely for clinically relevant changes may represent a window of opportunity for pancreatic cancer prevention and early detection. Preliminary data showed that the system could accurately detect patients with premalignant cysts (mucinous cysts), thus ensuring adequate management. However, the results were scarce and not well refined at the time of this paper, and were not included in the manuscript. Second, once patients with pancreatic cysts are correctly identified, screening this 'at risk' subpopulation for pancreatic cancer may be more feasible because the pretest probability is increased, thus compensating for the suboptimal sensitivity/specificity of currently available biomarkers/imaging studies.

As this algorithm is adaptable, it can be incorporated into any hospital electronic system to help capture patients with pancreatic cysts. After this pilot study, we plan to extend the system to the entire Indiana Network for Patient Care (INPC). The INPC contains the digital medical data of >95% of Indiana's hospital and medical institutions as well as a large number of hospitals in other states bordering Indiana. We anticipate the use of this programme as a template for other regional health information organizations (e.g. Boston, Utah, Stanford and Vanderbilt). The ultimate goal is to move towards establishment of a national pancreatic cyst registry. This programme may lead to a more organized national initiative for pancreatic cancer prevention and early detection, and optimal education of both healthcare providers and patients on current management and screening resources available.

Our study has limitations. It is a 7-month pilot study from a single patients network (INPC) and a single institution (Wishard Memorial Hospital). Further testing of our algorithm on other medical centre data is required to confirm its generalizability. Although NLP seems to be an inexpensive system, we have not conducted a cost-effectiveness analysis. The basic engine for NLP is 90% adaptable to a new environment. The implementation cost in a new hospital system would be associated with time/cost for obtaining the EMR text data, translating it into a format that our system can read (if not already in a readable format) and creating an alert programme compatible with the hospital system.

In spite of a high performance in identifying 'pancreatic cysts', recurrent and persistent errors have been identified, especially false-positives cases in complex structured sentences with multiple negated findings. The newly developed negation algorithm (DEEPEN) analysed sentences as a whole and was based on relations between words instead of differential location within one sentence. As this problem is rooted in the conception of the



	March	April	May	June	July	August	September	October	November	December
Specificity	99.2	100	99	98.8	99	99.1	96.3	97.1	98.4	99
Sensitivity	100	100	100	100	99	100	100	100	100	100

Figure 1 System performance characteristics (sensitivity and specificity) over the study period for pancreatic cyst identification

negation algorithm itself and was therefore not fixable, and as only five false positives (no impact on sensitivity) were identified by this error, we decided to keep using this very refined algorithm. Although 'pseudocyst' is a benign condition that alone does not require screening for pancreatic cancer, we included it in the final list of 'pancreatic cyst' concepts to be thorough and not miss patients with potentially other types of cysts. Our reasoning was based on manual review of multiple cross-sectional imaging studies reporting a pseudocyst in spite of the absence of clinical and radiological evidence of pancreatitis. Similarly, hypothetical terms, such as 'may represent', were considered affirmed to avoid missing patients with a potentially premalignant condition.

Incidentally discovered pancreatic cysts may be inconsequential. The data currently available on the natural history of pancreatic cysts and their malignant potential have some inherent selection/referral bias, and thus the percentage of truly consequential cysts might be lower in a general population screened by an automated process. The present study is a pilot series that aimed to confirm feasibility of an automated process using Natural Language Processing for pancreatic cyst screening. It accomplished this with high sensitivity and specificity. To analyse the potential for individual lives and cost-effectiveness, further studies including public health cost analyses would be needed to compare the cost of a preventive strategy with follow-up examinations versus the cost of pancreatic cancer treatment.

In conclusion, NLP is an effective, inexpensive and highly accurate tool to automatically identify patients with pancreatic cysts. This computerized system can help screen and track patients 'at risk' of pancreatic cancer. Because of its adaptability, we believe this system can grow in three ways. First, the system can be expanded regionally and ultimately nationally, to build a platform for a national pancreatic cyst registry, one arm of a pancreatic cancer early detection initiative. Second, the NLP algorithm used in the present study can be adapted to identify patients at risk of pancreatic cancer as a result of inherited conditions, e.g. a family history of pancreatic cancer. This is already underway with plans to pilot this study at a central Indiana hospital. Finally, the NLP algorithm used in the present study can be modified and applied to identify patients with other (i.e. non-pancreatic) precancerous lesions or conditions.

Conflicts of interest

None declared.

References

1. Howlader N, Noone AM, Krapcho M, Krapcho M, Garshell J, Neyman N *et al*. (2013) *SEER Cancer Statistics Review, 1975-2010*. Bethesda, MD: National Cancer Institute. http://seer.cancer.gov/csr/1975_2010/. Based on November 2012 SEER data submission posted to the SEER website, April. Available from: http://seer.cancer.gov/statfacts/html/pancreas.html (last accessed 1 March 2014).
2. Bruno MJ. (2013) Early diagnosis of pancreatic cancer; looking for a needle in a haystack? *Gut* 62:955–956.
3. Tanaka M, Chari S, Adsay V, Fernandez-del Castillo C, Falconi M, Shimizu M *et al*. (2006) International consensus guidelines for management of intraductal papillary mucinous neoplasms and mucinous cystic neoplasms of the pancreas. *Pancreatology* 6:17–32.
4. Tanaka M, Fernandez del Castillo C, Adsay V, Chari S, Falconi M, Jang JY *et al*. (2012) International consensus guidelines 2012 for the management of IPMN and MCN of the pancreas. *Pancreatology* 12:183–197.
5. Hruban RH, Canto M, Goggins M, Schulinck R, Klein AP. (2010) Update on familial pancreatic cancer. *Adv Surg* 44:293–311.
6. The Sol Goldman Pancreatic Cancer Research Center at Johns Hopkins Hospital. The National Familial Pancreas Tumor Registry (NFPTR). Available at: http://pathology.jhu.edu/pancreas/nfptr/index.php (last accessed 22 April 2014).

7. Laffan TA, Horton KM, Klein AP, Berlanstein B, Siegelman SS, Kawamoto S *et al*. (2008) Prevalence of unsuspected pancreatic cysts on MDCT. *Am J Roentgenol* 191:802–807.
8. Zhang XM, Mitchell DG, Dohke M, Holland GA, Parker L. (2002) Pancreatic cysts: depiction on single-shot fast spin-echo MR images. *Radiology* 223:547–553.
9. Kimura W, Nagai H, Kuroda A, Muto T, Esaki Y. (1995) Analysis of small cystic lesions of the pancreas. *Int J Pancreatol* 18:197–206.
10. Jones MJ, Buchanan AS, Neal CP, Dennison AR, Metcalfe MS, Garcea G. (2013) Imaging of indeterminate pancreatic cystic lesions. A systematic review. *Pancreatology* 13:436–442.
11. Berland LL, Silverman SG, Gore RM, Mayo-Smith WW, Megibow AJ, Yee J *et al*. (2010) Managing incidental findings on abdominal CT: white paper of the ACR incidental findings committee. *J Am Coll Radiol* 7:754–773.
12. Procacci C, Biasiutti C, Carbognin G, Accordini S, Bicego E, Guarise A *et al*. (1999) Characterization of cystic tumors of the pancreas: CT accuracy. *J Comput Assist Tomogr* 23:906–912.
13. Carbonell JG, Hayes PJ. (1992) Natural language understanding. In: Shapiro SC, ed. *Encyclopedia of Artificial Intelligence*. New York, NY: Wiley & Sons, pp. 997–1016.
14. Nadkarni PM, Ohno-Machado L, Chapman WW. (2011) Natural language processing: an introduction. *J Am Med Inform Assoc* 18:544–551.
15. Chapman WW, Gundlapalli AV, South BR, Dowling JN. (2011) Natural language processing for biosurveillance. In: Zeng D, Chen H, Castillo-Chavez C, Lober WB, Thurmond M, eds. *Infectious Disease Informatics and Biosurveillance. Integrated Series in Information Systems Vol. 27*. New York: Springer, pp. 279–310.
16. Biondich PG, Paul G, Grannis SJ. (2004) The Indiana network for patient care: an integrated clinical information system informed by over thirty years of experience. *J Public Health Manag Pract* 10:S81–S86.
17. Unified Medical Language System (UMLS) on US National Library of Medicine (NIH) website. Available at: http://www.nlm.nih.gov/research/umls/ (last accessed 1 April 2014).
18. Ferrucci D, Lally A. (2004) UIMA: an architectural approach to unstructured information processing in the corporate research environment. *Nat Lang Eng* 10:327–348.
19. Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG. (2001) Evaluation of negation phrases in narrative clinical reports. *Proc AMIA Symp* 105–109.
20. Mehrabi S, Schmidt CM, Waters JA, Beesley C, Krishnan A, Kesterson J *et al*. (2013) An efficient pancreatic cyst identification methodology using natural language processing. *Stud Health Technol Inform* 192:822–826.
21. De Marneffe MC, Manning CD. Stanford typed dependencies manual. 2008. Available at: http://nlp.stanford.edu/software/dependenciesmanual.pdf (last accessed 1 April 2013).
22. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM *et al*. (2003) Towards complete and accurate reporting of studies of diagnosis accuracy: the STARD initiative. *BMJ* 326: 41–44.
23. Al-Haddad MA, Friedlin J, Kesterson J, Waters JA, Aguilar-Saavedra JR, Schmidt CM. (2010) Natural language processing for the development of a clinical registry: a validation study in intraductal papillary mucinous neoplasms. *HPB* 12:688–695.
24. Goonetilleke KS, Siriwardena AK. (2007) Systematic review of carbohydrate antigen (CA19-9) as a biochemical marker in the diagnosis of pancreatic cancer. *Eur J Surg Oncol* 33:266–270.
25. Firpo MA, Gay DZ, Granger SR, Scaife CL, DiSario JA, Boucher KM *et al*. (2009) Improved diagnosis of pancreatic adenocarcinoma using haptoglobin and serum amyloid A in a panel screen. *World J Surg* 33:716–722.
26. Jones S, Zhang X, Parsons DW, Lin JC, Leary RJ, Angenendt P *et al*. (2008) Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. *Science* 321:1801–1806.
27. Sahani DV, Shah ZK, Catalano OA, Boland GW, Brugge WR. (2008) Radiology of pancreatic adenocarcinoma: current status of imaging. *J Gastroenterol Hepatol* 23:23–33.