Comparing Methods for Identifying Pancreatic Cancer Patients Using Electronic Data Sources

Jeff Friedlin, DO,1,2 Marc Overhage MD, PhD1,2
Mohammed A Al-Haddad, MD,2 Joshua A Waters, MD,2 J. Juan R Aguilar-Saavedra, MD,2 Joe Kesterson, MA1 Max Schmidt, MD, PhD, MBA2
1Regenstrief Institute, Inc, 2Indiana University School of Medicine, Indianapolis, IN.

Abstract
We sought to determine the accuracy of two electronic methods of identifying pancreatic cancer in a cohort of pancreatic cyst patients, and to examine the reasons for identification failure. We used the International Classification of Diseases, 9th Edition (ICD-9) codes and natural language processing (NLP) technology to identify pancreatic cancer in these patients. We compared both methods to a human-validated gold-standard surgical database. Both ICD-9 codes and NLP technology achieved high sensitivity for identifying pancreatic cancer in these patients. We compared both methods to a human-validated gold-standard surgical database. Both ICD-9 codes and NLP technology achieved high sensitivity for identifying pancreatic cancer, but the ICD-9 code method achieved markedly lower specificity and PPV compared to the NLP method. The NLP method required only slightly greater expenditures of time and effort compared to the ICD-9 code method. We identified several variables influencing the accuracy of ICD-9 codes to identify cancer patients including: the identification algorithm, kind of cancer to be identified, presence of other conditions similar to cancer, and presence of conditions that are precancerous.

Introduction
Investigators performing epidemiological studies frequently rely on administrative claims data to identify cases of specific diseases such as cancer. The claims data most frequently used in this manner are the International Classification of Diseases, 9th Edition (ICD-9) codes. The advantages to using ICD-9 codes to identify patients are that they are widely available and require significantly lower expenditures of time and effort compared to human chart review, self-report surveys, clinical trials etc. They have been used in research to identify cancer cases involving data not otherwise captured by cancer registries like the Surveillance, Epidemiology, and End Results (SEER) such as the incidence of brain metastases1, and the cost burden of cancer2. However, administrative claims databases collect data for the purposes of reimbursement and not clinical care, and their accuracy in identifying cancer patients varies. Eichler et al1 achieved a sensitivity of 100% and a positive predictive value (PPV) of 91% when using ICD-9 codes to identify lung cancer patients with central nervous system (CNS) metastases. However, Baldi et al3 achieved sensitivities ranging from 72%-81% and PPVs ranging from 79%-93% when using ICD-9 codes to identify patients with lung, colorectal and breast cancer. Freeman et al4, achieved both a sensitivity/specificity of greater than 90%, but only a PPV of about 70%, when using ICD-9 codes to identify breast cancer patients.

Unfortunately, if a diagnosis code is inaccurate, research based upon its assignment may be flawed. Some research has been conducted into the sources of errors in ICD-9 coding generally5, but not specifically as they relate to coding of cancer diagnoses. Additionally, little research has been conducted to investigate what are the viable alternatives for identifying patients in circumstances where ICD-9 codes prove inaccurate. This knowledge is important because it would not only help researchers predict circumstances when ICD-9 codes are less likely to be accurate, but also inform them of precise and feasible alternatives. One alternative method of identifying patient populations is through natural language processing (NLP) technology of medical reports. NLP has been used to identify patients with pneumonia6, obesity and comorbidities7, and clinical trial eligibility status8.

Researchers at the Indiana School of Medicine (IUSM) are studying patients with intraductal papillary mucinous neoplasms (IPMN). IPMNs are cysts that form in the pancreas and increase a patient’s risk of developing pancreatic cancer9. Little is known about the natural history of IPMN and the factors that increase the risk of these cysts becoming cancer. As part of this research, surgeons at IUSM have created a detailed registry of IPMN patients that includes the pancreatic cancer status of these patients.

Using this rich source of human-validated patient data, we sought to determine how reliably a diagnosis of pancreatic cancer could be concluded in these IPMN patients based upon two automated, electronic methods: 1.) existence of an ICD-9 code of pancreatic cancer in the patient record and 2.) use of NLP technology to identify pancreatic cancer in text reports. We also attempted to identify variables that might influence the accuracy of ICD-9 codes to identify cancer patients.

Methods
All patients in this study are part of the Indiana Network for Patient Care (INPC)10, a local health information exchange (HIE) that has been operating
for nearly 11 years. INPC participants include all five of the major hospital and health care systems in Indianapolis, as well as several large physician practices, two independent commercial laboratories, public health agencies, payors, and the Indiana State Medicaid program. Surgeons at IUSM have created a large pancreatic cyst patient registry (hereafter referred to as the surgical database) containing data for adult patients who have pathologically confirmed IPMN. Started in 1986, it includes patients treated at a large IUSM general surgery practice. The surgical database contains patient demographics, and other information including date of diagnosis, IPMN characteristics (such as cyst size, location), co-morbid conditions, pancreatic cancer status, etc. For patients with pancreatic cancer, it includes date of cancer diagnosis, tumor characteristics, cancer stage and treatment, as well as death status. The surgical database is updated through chart review of patient records by surgeons of the practice. As of January 2009, the pancreatic cyst surgical database contained 211 patients.

To evaluate the accuracy of the ICD-9 method for identifying pancreatic cancer in IPMN patients, we obtained the medical record numbers of all 211 patients in the surgical database and queried the INPC to identify patients with at least one occurrence of the ICD-9 diagnosis code for pancreatic cancer (157.*). This query was performed in January 2009. Each patient could have more than one occurrence of the code, and we placed no date restrictions on when the code appeared in the record.

To evaluate the accuracy of the NLP method for identifying pancreatic cancer in IPMN patients, we again used the medical record numbers of the 211 patients in the surgical database and queried the INPC to collect all medical text reports for these patients. The query was performed in January 2009, and we excluded all reports dated before 1986. We applied the NLP system to identify mentions of pancreatic cancer in any and all types of reports, not exclusively surgical pathology reports. None of these reports contained ICD-9 codes or terms.

<table>
<thead>
<tr>
<th>Concept</th>
<th>Context</th>
</tr>
</thead>
<tbody>
<tr>
<td>impression stage II (t3n0m0) adenocarcinoma of the pancreas.</td>
<td>positive</td>
</tr>
<tr>
<td>patient history: 64-year-old female whipple for gallbladder and pancreatic cancer</td>
<td>history</td>
</tr>
<tr>
<td>he has no concerning features for cancer of the pancreas</td>
<td>negated</td>
</tr>
<tr>
<td>cyst head of pancreas with strong family history of pancreas cancer</td>
<td>familyHx</td>
</tr>
</tbody>
</table>

**Figure 1.** Examples of Concept-Context couplets produced by REX

### NLP System

We used the Regenstrief EXtraction tool (REX) to process patient text reports. REX is an NLP software tool developed at the Regenstrief Institute in Indianapolis, IN, in 2006 and has been described previously. Briefly, REX is a rule-based NLP system written in Java that has successfully extracted patient data and concepts from microbiology reports, admission notes, and radiology reports. REX’s primary output for a given report is a series of concept-context couplets that are produced for all found mentions of a target concept within a report. An example of REX output for pancreatic cancer is shown in Figure 1. To identify concept-context couplets, REX uses a combination of regular expressions and algorithms to first detect where in the text keywords or phrases related to a concept are found. It then examines ‘windows’ of words before and after the concept phrase to determine concept context (i.e. positive, negated, historical, family history, related etc). The impetus for its development was the need by Regenstrief researchers for accurate NLP technology that did not take weeks to train and develop. As such, REX is not designed to identify every patient concept present in a report; instead, its main use and function is the rapid deployment of NLP technology for specific, targeted NLP tasks. It allows a user to quickly enhance and/or customize the NLP algorithms for particular use cases through modification of a knowledge base external to the REX program itself. These changes can be made using a simple word processor and since no modification of REX code is required, no programming experience is necessary (although knowledge of regular expressions is needed).

To identify keywords and phrases that express the concept of pancreatic cancer, we first performed a literature review of biomedical literature pertinent to pancreatic cancer. We then examined the Unified Medical Language System (UMLS) to identify synonyms and alternative phrases for the concept of pancreatic cancer. Also, we collected a training set of 300 INPC medical text reports (100 surgical pathology reports, 100 discharge summaries, and 100 clinic notes/consults) that contained the keywords ‘pancrea’ and ‘carcinoma’ and/or ‘cancer’. This training set was used to identify additional words/phrases clinicians use to express pancreatic cancer, as well as to test REX performance. We ensured the training set was unique from our evaluation set by excluding all records from patients whose medical record numbers matched those of the 211 patients in the pancreatic cyst surgical database.

### Statistical Analysis

We compared both the ICD-9 code and NLP methods of identifying pancreatic cancer patients to

---

**AMIA 2010 Symposium Proceedings Page**
the gold-standard information contained in the pancreatic cyst surgical database in order to calculate sensitivity, specificity, and positive predictive value (PPV) of each method. We defined sensitivity as the proportion of patients with pancreatic cancer in the surgical database who were identified as having pancreatic cancer. We defined specificity as the proportion of patients without pancreatic cancer in the surgical database who were identified as not having pancreatic cancer. PPV was defined as the proportion of patients identified as having pancreatic cancer whose diagnosis was confirmed by the surgical database.

This study was approved by the Indiana University Institutional Review Board and conducted in compliance with their regulations.

<table>
<thead>
<tr>
<th></th>
<th>Surgical Database</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>ICD-9</strong></td>
<td>Positives</td>
</tr>
<tr>
<td>Positives</td>
<td>52</td>
</tr>
<tr>
<td>Negatives</td>
<td>3</td>
</tr>
<tr>
<td>Sens 95% Spec 46%</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Surgical Database</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>NLP</strong></td>
<td>Positives</td>
</tr>
<tr>
<td>Positives</td>
<td>48</td>
</tr>
<tr>
<td>Negatives</td>
<td>7</td>
</tr>
<tr>
<td>Sens 87% Spec 94%</td>
<td></td>
</tr>
</tbody>
</table>

Figure 2. Accuracy of identifying pancreatic cancer patients by ICD-9 code and NLP compared to gold standard. (N=211)

Results

There were 211 patients in the surgical pancreatic cyst database. Of these, 55 (26%) had pathologically confirmed pancreatic cancer (156 had no pancreatic cancer).

Figure 2 displays the results of the two methods in identifying pancreatic cancer patients compared to the gold standard. The ICD-9 method was slightly more sensitive than the NLP method, but at a cost of much lower specificity and PPV. Of the 55 pancreatic cancer patients, we identified 52 using ICD-9 codes (sensitivity of 95%). However, the ICD-9 code method misclassified 84 (54%) of the 156 pancreatic cancer negative patients as having pancreatic cancer (specificity of 46%, and PPV of 38%). We reviewed the records of these 84 falsely identified pancreatic cancer patients and found that 65% had records with mentions of pancreatic cancer in text reports. Typical was a radiology report with the phrase: “Indication: history of pancreatic cancer” despite surgical pathology reports that only mentioned IPMN.

For the NLP method evaluation, we collected a total of 7366 text reports for the 211 pancreatic cyst patients. The number of text reports obtained per patient ranged from 1 to 249 (median 51). There were 337 unique types of reports.

As shown in Figure 2, of the 55 pancreatic cancer patients, NLP identified 48 (sensitivity 87%). The NLP method incorrectly identified 9 of the 156 pancreatic cancer negative patients as having pancreatic cancer (specificity 94%, and PPV 84%).

We further compared the two methods of identifying pancreatic cancer in this cohort by analyzing the overlap of errors committed by each method. As shown in Figure 3, of the 84 false positive patients identified by ICD-9 code, only 5 (6%) were also incorrectly identified positive for pancreatic cancer by NLP. Of the 9 false positive cancer errors made by NLP, 5 (56%) were also incorrectly identified by ICD-9 codes. Of the three false negative errors committed by the ICD-9 code method, all were determined by NLP to be positive pancreatic cancer patients. Of the seven false negative errors committed by NLP, all were positive pancreatic cancer patients by ICD-9 code.

We saw empirical evidence suggesting that IPMN was interpreted as pancreatic cancer by clinicians, although further study is needed to determine if this hypothesis is correct. For example, we observed discharge summaries containing a discharge diagnosis of pancreatic cancer, as well as consultation reports containing a patient history of pancreatic cancer, despite pathology reports that only mentioned IPMN.

We analyzed the reports of the nine pancreatic cancer patients that NLP incorrectly identified as

84 ICD-9 false positive patients
5 (6%) NLP Positive
79 (94%) NLP Negative

9 NLP false positive patients
5 (56%) ICD-9 Positive
4 (44%) ICD-9 Negative

Figure 3 Analysis of the false positive errors committed by the ICD-9 code and NLP methods.
having pancreatic cancer (false positives). Each patient’s medical record contained numerous positive mentions of pancreatic cancer even though their surgical pathology reports were negative for malignancy and in fact the patient did not have pancreatic cancer. For example, one hospitalized patient had a pathology report that stated “Final Diagnosis: intraductal papillary-mucinous tumor, adenoma” (a non-malignant tumor), but the discharge summary for this patient stated: “Discharge Diagnosis: 1. Pancreatic cancer”. Another patient had an anesthesia operative report that stated “patient diagnosed with pancreatic cancer last July” even though the surgical pathology report from this date clearly stated the specimen was positive for IPMN but negative for malignancy. We initially programmed REX to interpret phrases such as “history of pancreatic cancer” and “Indication: pancreatic cancer” as positive for pancreatic cancer. However, we discovered after reviewing the training set that these phrases frequently are not indicative of the diagnosis of pancreatic cancer for the reasons mentioned above. Consequently, REX classifies these phrases as ‘related to’ but not positive for pancreatic cancer, thus preventing large numbers of false positive errors.

We reviewed the reports of the seven pancreatic cancer patients that REX missed as having pancreatic cancer (false negatives). For two of these patients, the surgical pathology reports indicating pancreatic cancer were not included in the set of 7366 text reports collected for this project, and the remaining reports for these patients contained no positive mentions of pancreatic cancer. The remaining five patients all had unique phrasal patterns for pancreatic cancer that were not recognized by REX such as: “malignant cells arising from an adenocarcinoma arising from an intraductal papillary mucinous tumor”.

Modifying REX for this project was quick and straightforward. To perform the literature review, examine the UMLS, process the training set, and perform error analysis and knowledge base revision took approximately 20 man hours of effort. Processing the entire 7000 document test set took approximately 30 minutes.

**Discussion**

Both ICD-9 codes and NLP technology achieved high sensitivity for identifying pancreatic cancer (the ICD-9 code method was slightly more sensitive), but we found major differences in specificity and PPV for the two methods. The ICD-9 code method achieved significantly lower specificity and PPV compared to the NLP method.

To our knowledge, ours is the first study investigating the accuracy of using ICD-9 codes to identify patients with pancreatic cancer. Using ICD-9 codes to identify pancreatic cancer patients was less accurate in our study compared to other studies involving other cancers. We analyzed the causes of low precision attained by ICD-9 codes in our study and identified several variables that likely influence the accuracy of ICD-9 codes to identify cancer patients generally. These variables include the identification algorithm used, the kind of cancer to be identified, the presence of other conditions with similarities to cancer, the presence of conditions that are precancerous, and the types of diagnostic tests used to detect the cancer. Studies achieving greater accuracy with ICD-9 codes used algorithms consisting of either combinations of ICD-9 diagnosis and procedure codes (specific to the cancer) or multiple instances of ICD-9 diagnosis codes to identify cancer patients, rather than a single instance of an ICD-9 diagnosis code we used in our study. It is likely we would increase our specificity and PPV if our criteria to identify patients consisted of multiple instances of diagnosis codes or combinations of diagnosis and procedure codes, but this may have lowered our sensitivity as well. Further study is needed to investigate this hypothesis.

It is likely that cancer type plays a role in the accuracy of using ICD-9 codes to identify cancer patients. Evidence suggests that the existence of discrete procedures specific for a cancer location or type leads to better performance of ICD-9 diagnosis codes in detecting cancer patients. A “Whipple” procedure is a surgery generally reserved for pancreatic cancer, and this may have played a role in the higher sensitivities of ICD-9 codes in our study compared to other studies.

Perhaps the greatest influence on the accuracy of ICD-9 codes to identify pancreatic cancer patients relates to our study cohort and the nature of pancreatic cysts themselves. All patients in our study had pancreatic cysts which are a neoplastic process with some of the same characteristics as cancer, including lesion size, location, and histology. Also similar to cancers, these ‘tumors’, as they are sometimes referred to in clinical reports, are frequently excised and/or biopsied, with pathology reports produced describing the lesions. We hypothesize that these similarities to cancer may lead to confusion among clinicians and coders resulting in over-coding of pancreatic cancer in these cyst patients. Additionally, endoscopists may assign the pancreatic cancer code after an endoscopy but before there is histiologic proof of cancer. This phenomenon has been confirmed in other studies.

Another likely contributor to over-coding of pancreatic cancer in these patients is that IPMN is a precancerous lesion, which could cause ambiguity
and confusion in the diagnosis of these patients. Also a likely etiology of over-coding of pancreatic cancer in our study relates to the ordering of diagnostic tests or procedures. In previous unpublished research, we discovered that ICD-9 cancer codes are occasionally entered by some providers as the reason for ordering a diagnostic test (i.e. rule out pancreatic cancer or suspect pancreatic cancer). Since there is no ICD-9 code for “suspicion of pancreatic cancer” these patients are often coded simply as pancreatic cancer. The advent of ICD-10 codes may help prevent this in the future. The ICD-10 code Z03.1 (‘an observation of suspected malignant neoplasm’) could be used when ordering a diagnostic test for a patient suspected of cancer. Time will tell though whether clinicians will routinely use this code in these circumstances.

The ease and rapidity with which REX was deployed for this project is significant because the principal advantages to using ICD-9 codes to identify patients is not increased accuracy, but rather their wide availability and the lower expenditure of time and effort required, compared to human chart review. Using REX to identify pancreatic cancer patients resulted in increased accuracy and only slightly greater expenditures of time and effort compared to the traditional ICD-9 code method.

There are limitations to our study. Although we used data from a large and geographically diverse HIE, our data nonetheless originated from a single patient network. Our results may not be applicable to HIEs where billing procedures and/or clinical documentation practices differ. Our patient cohort has the potential for selection bias in that all patients had IPMN. Our results may not have been the same if we applied these methods to identify pancreatic cancer patients in a more general patient population. The ICD-9 code method may be more accurate in identifying pancreatic cancer patients in individuals who do not have cysts in the pancreas.

References