Radiology: Artificial Intelligence

ORIGINAL RESEARCH

Automated Identification and Measurement Extraction of Pancreatic Cystic Lesions from Free-Text Radiology Reports Using Natural Language Processing

Rikiya Yamashita, MD, PhD • Kristen Bird, MD • Philip Yue-Cheng Cheung, MD, MEng • Johannes Hugo Decker, MD, PhD • Marta Nicole Flory, MD • Daniel Goff, MD, PhD • Linda Nayeli Morimoto, MD • Andy Shon, MD • Andrew Louis Wentland, MD, PhD • Daniel L. Rubin, MD, MS • Terry S. Desser, MD

From the Departments of Biomedical Data Science (R.Y., D.L.R.) and Radiology (K.B., P.Y.C.C., J.H.D., M.N.F., D.G., L.N.M., A.S., A.L.W., D.L.R., T.S.D.), Stanford University School of Medicine, 300 Pasteur Dr, Stanford, CA 94305. Received April 6, 2021; revision requested May 11; revision received October 26; accepted December 2. **Address correspondence to** T.S.D. (e-mail: *tsdesser@stanford.edu*).

Authors declared no funding for this work.

Conflicts of interest are listed at the end of this article.

See also commentary by Horii in this issue.

Radiology: Artificial Intelligence 2022; 4(2):e210092 • https://doi.org/10.1148/ryai.210092 • Content codes: AI CT GI MR

Purpose: To automatically identify a cohort of patients with pancreatic cystic lesions (PCLs) and extract PCL measurements from historical CT and MRI reports using natural language processing (NLP) and a question answering system.

Materials and Methods: Institutional review board approval was obtained for this retrospective Health Insurance Portability and Accountability Act–compliant study, and the requirement to obtain informed consent was waived. A cohort of free-text CT and MRI reports generated between January 1991 and July 2019 that covered the pancreatic region were identified. A PCL identification model was developed by modifying a rule-based information extraction model; measurement extraction was performed using a state-of-the-art question answering system. The system's performance was evaluated against radiologists' annotations.

Results: For this study, 430 426 free-text radiology reports from 199 783 unique patients were identified. The NLP model for identifying PCL was applied to 1000 test samples. The interobserver agreement between the model and two radiologists was almost perfect (Fleiss $\kappa = 0.951$), and the false-positive rate and true-positive rate were 3.0% and 98.2%, respectively, against consensus of radiologists' annotations as ground truths. The overall accuracy and Lin concordance correlation coefficient for measurement extraction were 0.958 and 0.874, respectively, against radiologists' annotations as ground truths.

Conclusion: An NLP-based system was developed that identifies patients with PCLs and extracts measurements from a large single-institution archive of free-text radiology reports. This approach may prove valuable to study the natural history and potential risks of PCLs and can be applied to many other use cases.

Supplemental material is available for this article.

The widespread use of abdominal CT and MRI for the investigation of upper abdominal complaints results in frequent incidental detection of small (<2.5 cm) pancreatic cystic lesions (PCLs). Despite years of experience with these lesions at imaging and extensive study, however, the natural history of small PCLs remains uncertain, and an indolent versus aggressive course for an otherwise simple-appearing cyst cannot be predicted with certainty. In a 2013 study, the estimated overall prevalence of PCLs in a population of adults in the United States (age range, 40–84 years) was 2.5%, with cysts detected at 2.2% of upper abdominal CT examinations and at 19.6% of MRI examinations (1).

To assist radiologists in providing consistent and rational recommendations when a cyst is detected at imaging, the American College of Radiology (ACR) developed guidelines for management of PCLs detected incidentally at pancreatic imaging (2). The recommendations are based on the initial size of the cyst and the rate of growth at serial imaging. Assessment of the validity of these recommendations is ongoing.

Robust research studies of PCLs require large sample sizes, because malignant transformation is rare in otherwise simple-appearing small cystic lesions (3). Cohort size in research studies is constrained by the limited availability of experts' time necessary to identify suitable cases through manual review of free-text documents such as radiology reports. However, a tremendous number of free-text radiology reports are stored in digital format in electronic medical record systems, and these can be mined and leveraged for many research and application development purposes. Furthermore, recent advances in natural language processing (NLP) have enabled analysis of large amounts of unstructured text data for identifying concepts, patterns, topics, keywords, and other attributes.

Abbreviations

ACR = American College of Radiology, BioBERT = Bidirectional Encoder Representations from Transformers for Biomedical Text Mining, IPMN = intraductal papillary mucinous neoplasm, IQR = interquartile range, NLP = natural language processing, PCL = pancreatic cystic lesion, PDAC = pancreatic ductal adenocarcinoma

Summary

Automated algorithms were developed using natural language processing and a question answering system to identify patients with pancreatic cystic lesions and extract lesion measurements from a 28-year archive of free-text CT and MRI radiology reports.

Key Points

- An automated natural language processing system for identifying pancreatic cystic lesions (PCLs) from free-text radiology reports showed almost perfect interobserver agreement with radiologists (Fleiss κ = 0.951), and the false-positive rate and true-positive rate were 3.0% and 98.2%, respectively, against a consensus of radiologists' annotations as ground truths.
- A second question answering system–based algorithm for measurement extraction of PCLs achieved an overall accuracy of 0.958 and Lin concordance correlation coefficient of 0.874 against radiologists' annotations as ground truths.

Keywords

Informatics, Abdomen/GI, Pancreas, Cysts, Computer Applications-General (Informatics), Named Entity Recognition

Automated cohort selection and information extraction from a large corpus of radiology reports can greatly facilitate research on PCLs. For cohort selection, three previous studies successfully applied NLP to identify PCLs from within radiology reports (4–6), with two studies (4,5) using in-house rule-based systems and one study (6) using a combination of commercial software and a string-search algorithm. One shortcoming of these studies, however, is that the algorithms used are not publicly available. With respect to extracting the information necessary for lesion characterization, cyst measurement and interval growth are the key parameters necessary for managing incidental PCLs per the recommendations of the ACR Incidental Findings Committee, as noted above (2). To our knowledge, no studies have been published in which NLP was used to extract measurements of PCLs specifically from radiology reports. Bozkurt et al (7) developed an automated measurement extraction system for CT and MRI reports using a hybrid NLP algorithm, but their model extracts measurements of any sort of abnormality in reports; that is, it extracts measurements not only of PCLs but also, when applied to the pancreas, other lesions as well, such as solid tumors and pancreatic ducts.

A retrospective study published by Pandey et al (8) in 2019 evaluated the size categories of the ACR management guidelines in 390 patients referred to their institution with a known clinical diagnosis of an incidentally detected PCL. Though they did not specify how they identified the patient cohort in their study, it is presumable that patients were identified from the clinical history or study indication field in the radiology report. Cyst analysis and measurements were performed by manual review of all images and thus were time intensive for radiologists to perform. Our study aims to explore the feasibility of conducting a similar but wholly automated assessment of the natural history of PCLs by deploying NLP methods to extract the information necessary to provide management recommendations from radiology reports only. Thus, we sought to automatically identify patients whose radiology reports described PCLs (ie, not simply those with a known diagnosis of an incidental PCL) and to extract measurements from historical CT and MRI reports using NLP techniques.

Materials and Methods

Study Sample

This Health Insurance Portability and Accountability Act–compliant retrospective study was approved by the institutional review board, and the need for patient informed consent was waived. Our institution has dedicated examination codes for pancreas-specific examinations as well as general abdominal examinations. We identified a cohort of free-text radiology reports for CT and MRI examinations generated between January 1991 and July 2019 that covered the pancreatic region through a database query using 204 relevant examination codes (see Appendix E1 [supplement]).

Identifying Patients with Pancreatic Cysts

An NLP model for identifying PCLs was developed by modifying CheXpert labeler (9), an automated rule-based labeler that was initially developed for chest radiography, to extract observations from free-text radiology reports. The labeler was set up in three distinct stages: mention extraction, mention classification, and mention aggregation. First, the labeler extracts mentions of observations from the free-text radiology reports. A list of phrases for mention extraction (see Appendix E2 [supplement]) was curated by a radiologist (R.Y.) with 12 years of experience in body imaging by reviewing 204 randomly sampled reports that were subsequently excluded from the rest of the analysis (Fig 1). Second, the extracted mentions of observations are classified as negative (eg, "no evidence of pancreatic cystic lesion"), uncertain (eg, "low-attenuation may represent small pancreatic cystic lesion"), or positive (eg, "hypodensity representing small pancreatic cystic lesion"). Finally, we use the classification for mentions of observations to arrive at a final label for the presence or absence of a PCL. The model's performance for identifying reports with PCLs was assessed through a reader study using 1000 reports from 1000 patients selected via stratified random sampling. We recruited eight radiologists (K.B., P.Y.C.C., J.H.D., M.N.F., D.G., L.N.M., A.S., A.L.W.) with 3–10 years' experience each, and each radiologist independently annotated 250 reports as to the presence or absence of a PCL. Thus, each report was annotated twice by two independent radiologists. We assessed agreement among the radiologists, as well as between the algorithm and the radiologists (see statistical analysis section below). To assess the model performance, we established consensus ground truths (the presence or absence of a PCL in each report) on the basis of the radiologists' responses obtained through the reader study, wherein uncertain responses and disagreements between the

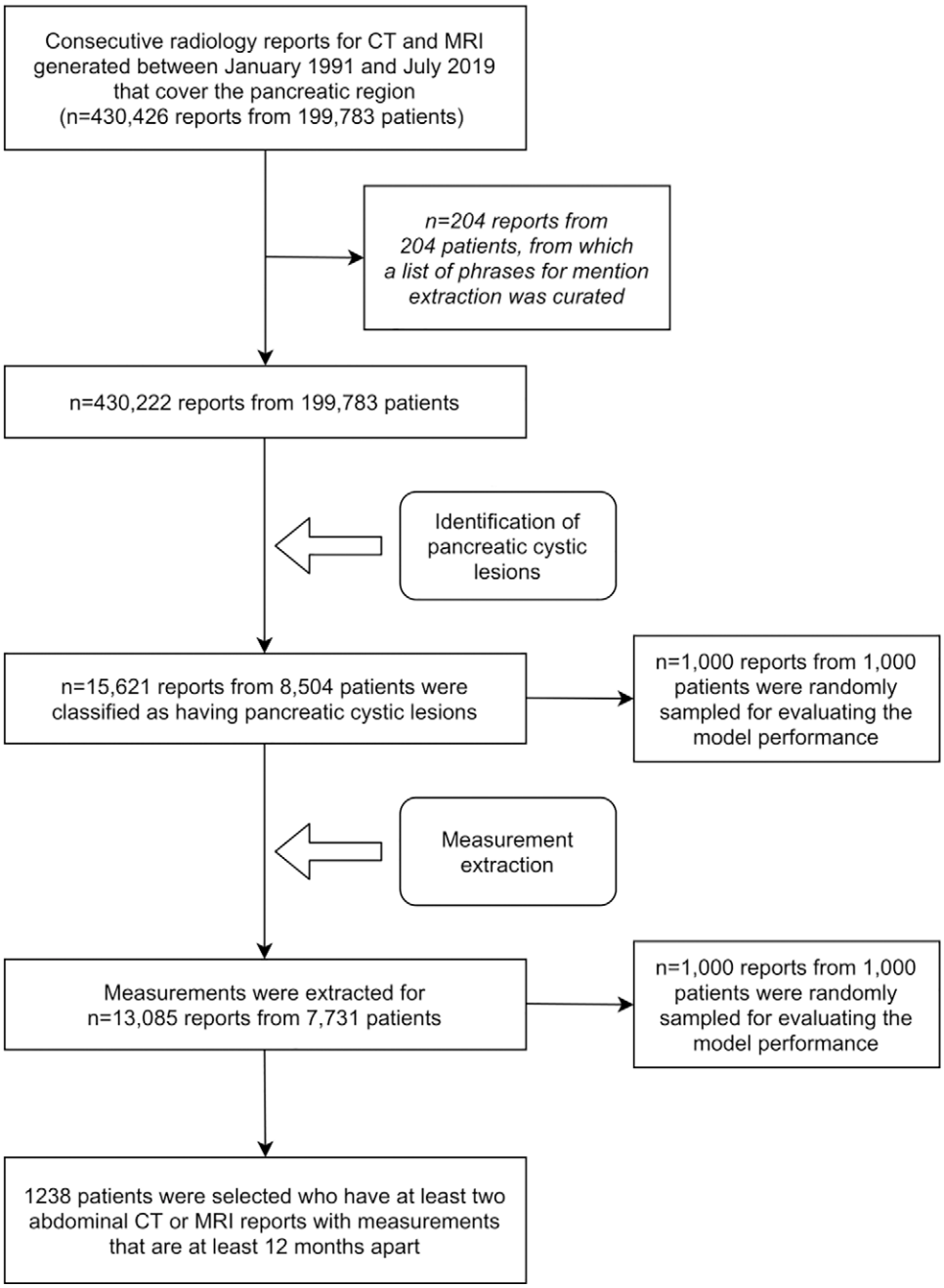

Figure 1: Study flowchart.

two radiologists were resolved by a senior radiologist (T.S.D.) with 30 years of experience in body imaging. The radiologists at our institution have followed the cyst measurement method proposed in the ACR white paper since its publication in 2017 (2); however, the PCL measurement method was not standardized for the reports generated before publication of the ACR white paper. To assess the effect of the ACR white paper, we evaluated our model performance on reports generated before and after white paper publication separately.

Extracting Measurements for Pancreatic Cysts

As a next step, we developed an NLP-based approach for measurement extraction of PCLs. We used a Bidirectional Encoder Representations from Transformers for Biomedical Text Mining (BioBERT)–based question answering model, which was first pretrained on general domain corpora (English Wikipedia and BooksCorpus) and biomedical domain corpora (PubMed abstracts and PubMed Central full-text articles) (10) and then was subsequently fine-tuned on the Stanford Question Answering Dataset (version 2.0) (11). We selected the BioBERT-based model because BioBERT is one of the state-of-the-art models on various NLP tasks in the biomedical domain, including named entity recognition, relation extraction, and question answering (10). The specific model we used in this study was downloaded from the NLP model repository HuggingFace (12), and we did not perform any further fine-tuning on our dataset. We fed radiology reports and a collection of questions into the model as inputs, and the model output relevant texts or subsentences of the reports that would best answer the questions. To increase the accuracy of the measurement extraction, before being fed into the model, sentences in the "Findings" subsection that were relevant to the pancreas were selected using the open-source software Python (version 3.6.10; Python Software Foundation; *https://www.python.org/*); 3.35 sentences were extracted on average. A list of questions we used is shown in Table 1. Of the model outputs, we recorded the single largest diameter at each report, which was used for the rest of the analysis. To assess the performance of measurement extraction, 1000 reports from 1000 patients were randomly selected, and measurements were manually extracted by a radiologist (R.Y.) with 12 years of experience in body imaging. As with the lesion identification, we also assessed the measurement extraction performance on reports generated before and after white paper publication separately. Additional methodological details can be found in the source code (see code availability section at the end of this article).

As in the previous work by Pandey et al (8), we selected patients with PCLs who had at least two CT or MRI reports with

Table 1: List of Questions for Measurement Extraction for Pancreatic Cystic Lesions

Prefix	Lesion	Suffix
What is the size of the	cyst	in the pancreas
	cystic lesion	in the pancreatic head
	cystic focus	in the pancreatic neck
	low density	in the pancreatic body
	low attenuation	in the pancreatic tail
	hypodensity	in the uncinate process of the pancreas
	hypoattenuation	
	IPMN	
	IPMT	
	intraductal papillary mucinous neoplasm	
	intraductal papillary mucinous tumor	
	T2 bright lesion	
	T2 bright focus	
	T2 hyperintense lesion	
	T2 hyperintense focus	

Note.—Each question is a combination of prefix, lesion, and suffix. IPMN = intraductal papillary mucinous neoplasm, IPMT = intraductal papillary mucinous tumor.

measurements from studies performed at least 12 months apart. Among such patients whose pathology reports were available through the Stanford Research Repository, we extracted histopathologic diagnosis for any resected pancreatic abnormality. For extracting histopathologic diagnosis, we only included pathology reports of surgically resected specimens and did not include endoscopic US–guided fine-needle aspiration or biopsy reports. Pathology reports were reviewed manually. Radiology reports generated after the production of pathology reports were excluded. The single largest diameter at baseline and at the last available follow-up were recorded for all patients, as the radiology reports often contain only the single largest diameter of the largest PCL. For each cyst, growth was defined on the basis of the recommendations of the white paper of the ACR Incidental Findings Committee (2). For groups with baseline size of less than 5 mm, of 5–14 mm, of 15–25 mm, or of more than 25 mm, an increase or decrease from baseline size of 100% or more, 50% or more, 20% or more, and 20% or more, respectively, was considered a change in size.

Statistical Analysis

To assess the interobserver variability of PCL identification in our study, using the same method as the original CheXpert-labeler study (9), we selected 1000 reports from 1000 patients via stratified random sampling, which consist of 250 samples each for positively classified CT reports, negatively classified CT reports, positively classified MRI reports, and negatively classified MRI reports. Cohen κ and Fleiss κ were calculated for agreement between two and three observers, respectively. We considered κ greater than 0.81 as denoting almost perfect agreement and values of 0.61–0.80, 0.41–0.60, 0.21–0.40, 0.0–0.20, and less than 0.0 as denoting substantial, moderate, fair, slight, and poor agreement, respectively (13).

To assess the model performance of PCL identification, we computed accuracy, sensitivity, specificity, positive predictive value, and negative predictive value using the ground truths generated via the reader study. The 95% CIs were calculated using the Clopper-Pearson interval. To evaluate the model performance for measurement extraction, we computed accuracy and Lin concordance correlation coefficient (14) between measurements extracted by the model and the radiologist. For measurements, accuracy was defined as the proportion of reports in which the maximum measurements extracted by the model and the radiologist's annotation were exactly the same. A two-tailed α criterion of .05 was used for statistical significance. Statistical analyses were performed in Python (version 3.6.10) and R software (version 3.6.3; The R Foundation).

Results

Study Sample

We identified 430 426 free-text radiology reports (378 061 CT and 52 365 MRI examinations) from 199 783 unique patients through a database query with 204 examination codes between January 1991 and July 2019. As shown in the study flowchart (Fig 1), a sample of 204 reports was used to generate phrases for mention extraction, and the studies were then excluded from subsequent analysis. Cases were retrospectively sorted by examination code, and 9193 of 430 426 reports were found to have pancreatic lesion–specific structured text.

Identifying Patients with Pancreatic Cysts

Of the remaining 430 222 reports from 199 783 patients, 15 621 reports (3.63%) from 8504 patients (4.26%) were classified by the algorithm as positive for PCLs (10 281 of 377 955

CT reports [2.72%] and 5340 of 52267 MRI reports [10.2%]) (Table 2). Of the 8504 patients positive for PCL, age data were available from the reports of 5558 patients; the median age was 70.0 years (interquartile range [IQR], 49.0–86.0 years). Data on sex were available in reports for 5595 patients (3218 [57.5%] female, 2377 [42.5%] male).

Among the 8504 patients positive for PCL, subtypes of the PCL were mentioned in reports for 6994 patients. Of these 6994 patients, the reports for 6341 (90.7%), 238 (3.4%), 384 (5.5%), 15 (0.2%), and 16 (0.2%) had phrases suggestive of intraductal papillary mucinous neoplasms (IPMNs), serous cystic neoplasms, mucinous cystic neoplasms, solid pseudopapillary neoplasms, and lymphoepithelial cysts, respectively.

For the reader study with 1000 samples selected via stratified random sampling, interobserver agreement between two radiologists was almost perfect (Cohen κ = 0.968); moreover, agreement between the NLP algorithm and two observers was also almost perfect (Fleiss κ = 0.951). With the ground truths generated via the reader study, the false-positive rate and true-positive rate for PCL identification were 3.0% and 98.2%, respectively, for the whole case cohort of CT and MRI reports; 3.1% and 98.8%, respectively, for CT only; and 2.8% and 97.6%, respectively, for MRI only (Table 3). Among reports generated before and after ACR white paper publication, the model achieved higher performance on the reports generated after white paper publication (Table 3).

Table 2: Numbers of Reports and Patients Positive for Pancreatic Cystic Lesions

Variable	Reports	Patients
Positive for PCL	15621 (3.63)	8504 (4.26)
Suggestive of IPMN	13154 (3.06)	7321 (3.66)
Follow-up duration (y)		
1	…	1496 (0.75)
2	…	742 (0.37)
3	…	506 (0.25)
4	…	358 (0.18)
5	…	250 (0.13)

Note.—Data are numbers with percentages in parentheses. Percentages based on total number of reports (*n* = 430222) and patients (*n* = 199783). IPMN = intraductal papillary mucinous neoplasm, PCL = pancreatic cystic lesions.

Extracting Measurements for Pancreatic Cysts

Of the 15621 positively classified reports by the PCL identification model, our measurement extraction model identified measurements for PCL in 13085 reports (83.8%). Examples of measurement extraction by the question answering system are shown in Figure 2. For the 1000 randomly sampled reports and manually collected ground truth labels, the accuracy for extracting the largest measurements of the PCLs was 0.958, and the Lin concordance correlation coefficient was 0.874 (Fig 3). The accuracy and concordance correlation coefficient of the measurement extraction were 0.954 and 0.774, respectively, for the reports generated before white paper publication, and 0.962 and 0.894, respectively, for the reports generated after white paper publication. No measurements were mentioned in the remaining 2536 reports. An error analysis demonstrated that the measurement extraction failed in 42 of 1000 reports for the following reasons: *(a)* the lesions were not PCLs but were described as either "nonenhancing," "low-attenuation," or "hypodense" (*n* = 13); *(b)* there was a lesion with solid and cystic characteristics (*n* = 1); *(c)* our model was confused by prior measurements (eg, "31 × 33 mm low attenuation cyst at the pancreatic head/uncinate process has decreased in size and now measures approximately 15 × 11 mm" and "previously described 6-mm hypodense cystic lesion in the pancreatic neck is not well seen, likely due to motion artifact") (*n* = 8); *(d)* the measurements were noted in parentheses (*n* = 3); *(e)* there was a cystic dilatation of main pancreatic duct (*n* = 2); *(f)* the lesions were suspicious for pseudocysts (*n* = 1); and *(g)* the PCL measurements were mentioned in the reports but the algorithm did not extract them for non–human-interpretable reasons (*n* = 14).

On the basis of the results of measurement extraction, we identified 1238 patients who had at least two abdominal CT or MRI reports with measurements available and whose studies were performed at least 12 months apart. The median follow-up duration was 31.4 months (IQR, 19.5–56.1 months; range, 12.0–260.9 months). Histopathologic diagnoses for pancreatic abnormalities were identified and obtained for 31 patients (Table 4) who underwent surgical resection after a median

Table 3: Cyst Identification Performance on Consensus Ground Truth by Radiologists

Metric	CT + MRI (*n* = 1000)	CT (*n* = 500)	MRI (*n* = 500)	CT + MRI (pre-ACRWP, *n* = 697)	CT+MRI (post-ACRWP, *n* = 303)
Accuracy	97.6 (96.5, 98.5)	97.8 (96.1, 98.9)	97.4 (95.6, 98.6)	97.0 (95.4, 98.1)	99.0 (97.1, 99.8)
Sensitivity	98.2 (96.6, 99.2)	98.8 (96.5, 99.7)	97.6 (94.8, 99.1)	97.5 (95.2, 98.9)	99.4 (96.8, 100.0)
Specificity	97.0 (95.2, 98.3)	96.9 (93.9, 98.6)	97.2 (94.3, 98.9)	96.5 (94.1, 98.1)	98.5 (94.6, 99.8)
PPV	97.0 (95.1, 98.3)	96.8 (93.8, 98.6)	97.2 (94.3, 98.9)	96.0 (93.3, 97.9)	98.8 (95.9, 99.9)
NPV	98.2 (96.6, 99.2)	98.8 (96.5, 99.8)	97.6 (94.8, 99.1)	97.8 (95.8, 99.1)	99.2 (95.8, 100.0)

Note.—Data are percentages with 95% CIs in parentheses. ACRWP = American College of Radiology white paper, NPV = negative predictive value, PPV = positive predictive value.

Sample report	NLP model (mm)	Ground truth (mm)
pancreas: low-density 7 mm cystic focus in the uncinate process and is a small cystic focus adjacent to the tail the pancreas measuring 7 mm is also seen. pancreatic duct is borderline dilated in the tail but otherwise normal in caliber. no solid pancreatic masses are seen. a hypervascular 1cm mass is seen in the tail of the pancreas and is favored to represent a splenule.	7	7
pancreas: adjacent to the pancreaticojejunostomy there is a horseshoe shaped 4.8 x 2.5 cm fluid collection, likely post surgical. there is an adjacent jackson-pratt drain which extends along the body/tail of the pancreas and comes in close contact to this small fluid collection. the pancreas enhances symmetrically, and there are a few stable small cystic lesions in the tail of the pancreas measuring up to 1 cm the spleen and bilateral adrenal glands are unremarkable.	10	10
pancreas: redemonstration of a 7 mm hyperenhancing lesion in the head of the pancreas, unchanged dating back to 2/2/2017 and without uptake on pet, favored to represent a small neuroendocrine tumor. punctate hypoattenuating lesion in the pancreatic head is stable, likely a small sidebranch ipmn. mild prominence of the pancreatic duct measuring 2-3 mm has been slowly increasing from february 2017.	n/a	n/a

Figure 2: Examples of measurement extraction by the question answering system. The system identified measurements of pancreatic cystic lesions (highlighted in orange), and it successfully ignored measurements for prior studies, as well as measurements for noncystic abnormalities (highlighted in green). NPL = natural language processing.

follow-up of 31.8 months (IQR, 17.7–67.8 months; range, 13.5–260.9 months). Of the 31 patients who underwent surgical resection for their pancreatic abnormalities, pathologic analysis revealed nine (29.0%) invasive pancreatic ductal adenocarcinomas (PDACs), 15 (48.4%) IPMNs, four (12.9%) serous cystic neoplasms, two (6.5%) mucinous cystic neoplasms without invasive carcinoma or high-grade dysplasia, and one (3.2%) cystic pancreatic neuroendocrine tumor (Table 4). Of the nine PDACs, two (22.2%) arose from a gastric-type IPMN (one combined IPMN, one branch duct IPMN), and the other seven were de novo. Of the nine PCLs in patients with pathologic analysis–proven PDAC, six (66.7%) presented growth at follow-up according to the ACR growth criteria, whereas of the 15 PCLs in patients with pathologic analysis–proven IPMN, eight (53.3%) presented growth at follow-up. Of the 15 IPMNs, three (20.0%) had pathologic analysis–proven mural nodules, of which one presented growth at follow-up and the other two remained stable.

The characteristics of PCLs at baseline and their interval growth patterns are summarized in Table 5. At baseline, 1238 cysts had a median cyst size of 10.0 mm (IQR, 6–15 mm). At the last follow-up, the cysts had a median size of 11.0 mm (IQR, 7–17 mm). By ACR size category, the 1238 PCLs measured as follows: less than 5 mm, 140 (11.3%); 5–14 mm, 742 (59.9%); 15–25 mm, 259 (20.9%); and greater than 25 mm, 97 (7.8%). Among the PCLs demonstrating growth according to ACR 2017 criteria (2), the median baseline cyst size was 9.0 mm (IQR, 6–16 mm).

A total of 270 cysts (21.8%) increased in size (median follow-up, 44.4 months), 114 (9.2%) decreased in size (median follow-up, 31.3 months), and 854 (69.0%) remained stable (median follow-up, 28.9 months). In the baseline size groups of less than 5 mm, 5–14 mm, 15–25 mm, and greater than 25 mm, size increase was noted in 34 (24.3%), 157 (21.2%), 59 (22.8%), and 20 (20.6%) cysts, respectively (Table 5). The bar chart in Figure 4, which is modeled on Pandey et al (8), shows the proportion of PCLs showing size change in the baseline size categories as defined by the ACR white paper recommendations (2).

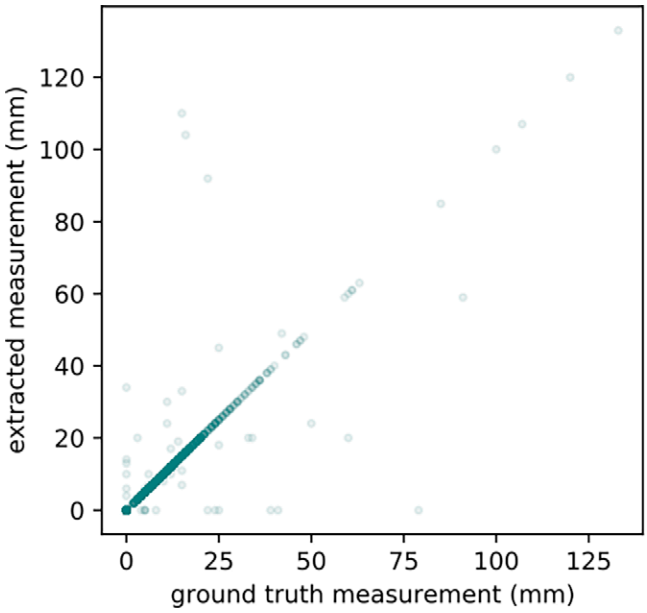

Figure 3: Agreement of measurement extraction between ground truth and natural language processing model. Scatterplot shows agreement of measurement extraction between a radiologist and our question answering–based system; 95.8% of the reports were aligned on the diagonal line, which is almost perfect agreement (Lin concordance correlation coefficient = 0.874).

Discussion

Despite decades of experience with CT imaging and MRI of incidental PCLs, definitive risk stratification strategies have not yet been established. The recently developed management recommendations of the ACR for incidental PCLs specify timelines for serial follow-up of lesions on the basis of baseline cyst size and interval growth (2). These management recommendations have been informed largely by medical society consensus papers, meta-analyses, and observational studies, because conclusive evidence is still lacking (3). Given the number of years that CT and MRI have been in use, vast archives of evidence to inform decision-making about PCLs are potentially available, but this evidence is currently buried within free-text reports. Robust NLP tools provide a potential opportunity to mine

Table 4: Histopathologic Diagnoses and Their Association with Cyst Size Categories and Interval Growth Patterns in Patients with Pancreatic Cystic Lesions

Histopathologic Finding	<5 mm	5–14 mm			15–25 mm			≥25 mm			
	All	Stable	Enlarged	Decreased	Stable	Enlarged	Decreased	Stable	Enlarged	Decreased	Total*
IPMN	0	4 (26.7)	3 (20.0)	0	1 (6.7)	4 (26.7)	1 (6.7)	1 (6.7)	1 (6.7)	0	15 (100)
PDAC	0	2 (22.2)	3 (33.3)	0	1 (11.1)	1 (11.1)	0	0	2 (22.2)	0	9 (100)
Arising in IPMN	…	1	1	…	…	…	…	…	…	…	2
SCN	0	0	0	0	0	1 (25.0)	1 (25.0)	1 (25.0)	1 (25.0)	0	4 (100)
MCN	0	0	1 (50.0)	0	0	0	0	1 (50.0)	0	0	2 (100)
Cystic PNET	0	0	0	0	1 (100)	0	0	0	0	0	1 (100)
Total	0	6 (19.4)	7 (22.6)	0	3 (9.7)	6 (19.4)	2 (6.5)	3 (9.7)	4 (12.9)	0	31 (100)

Note.—Data are numbers with percentages in parentheses. IPMN = intraductal papillary mucinous neoplasm, MCN = mucinous cystic neoplasm, PDAC = pancreatic ductal adenocarcinoma, PNET = pancreatic neuroendocrine tumor, SCN = serous cystic neoplasm.
* Totals may not equal 100 owing to rounding.

Table 5: Characteristics of Pancreatic Cystic Lesions at Baseline and Their Interval Growth Patterns

Characteristic	Overall	Size Change Category		
		Stable	Enlarged	Decreased
No. of cysts	1238 (100)	854 (69.0)	270 (21.8)	114 (9.2)
Baseline size category				
<5 mm	140 (100)	106 (75.7)	34 (24.3)	0 (0.0)
5–14 mm	742 (100)	570 (76.8)	157 (21.2)	15 (2.0)
15–25 mm	259 (100)	135 (52.1)	59 (22.8)	65 (25.1)
≥25 mm	97 (100)	43 (44.3)	20 (20.6)	34 (35.1)
Cyst size (mm)*				
At baseline	10.0 (6.0–15.0)	9.0 (6.0–13.0)	9.0 (6.0–16.0)	19.0 (15.0–25.75)
At last follow-up	11.0 (7.0–17.0)	9.75 (6.0–14.0)	18.0 (12.0–28.0)	12.0 (8.0–14.0)
Imaging follow-up duration (mo)*	31.4 (19.5–56.1)	28.9 (18.7–50.0)	44.4 (23.7–74.3)	31.3 (18.6–56.8)

Note.—Except where otherwise noted, data are numbers with percentages in parentheses.
* Data are medians with interquartile range in parentheses.

these historical archives across multiple institutions to inform evidence-based recommendations based on natural history as assessed on serial imaging examinations.

With the aforementioned factors as motivation, we sought to develop tools to extract the parameters that underscore the ACR guidelines. Our NLP system has two stages. First, the system identified patients with PCLs from more than 430 000 free-text radiology reports. We showed that the system achieved almost perfect agreement with radiologists in detecting the presence of PCLs in the free text of radiology reports (Fleiss κ = 0.951). Our system identified PCLs in 2.72% of abdominal CT reports and 10.2% of abdominal MRI reports, which is comparable to previous studies in which PCLs were identified at 2.4% of CT examinations (15) and 19.6% of MRI examinations (16).

Second, our system extracted measurements for the PCLs from the free-text reports. Regarding the single largest measurement in each examination, our system achieved excellent accuracy of 0.958 and a Lin concordance correlation coefficient of 0.874 against manual annotations by a radiologist as ground truths.

Our PCL identification model is based on the CheXpert-labeler (9), which was developed to extract structured labels for a large chest radiograph dataset. Although the CheXpert-labeler was originally intended to be applied to chest radiography reports, we showed that our approach successfully transformed it into a PCL detector for CT and MRI reports. One drawback of this rule-based approach is that it requires a list of phrases for mention extraction per each use case, which requires experts' inputs. As presented in Appendix E2 (supplement), however, the phrases in the list for mention extraction are not medical site–specific; that is, they consist only of general terms. Therefore, it is reasonable to consider that our approach could easily be generalized to other medical sites as well as a different set of radiologists. Additionally, the advantage of such a rule-based approach is that it does not require the manually labeled large datasets that are typically required for supervised machine learning models.

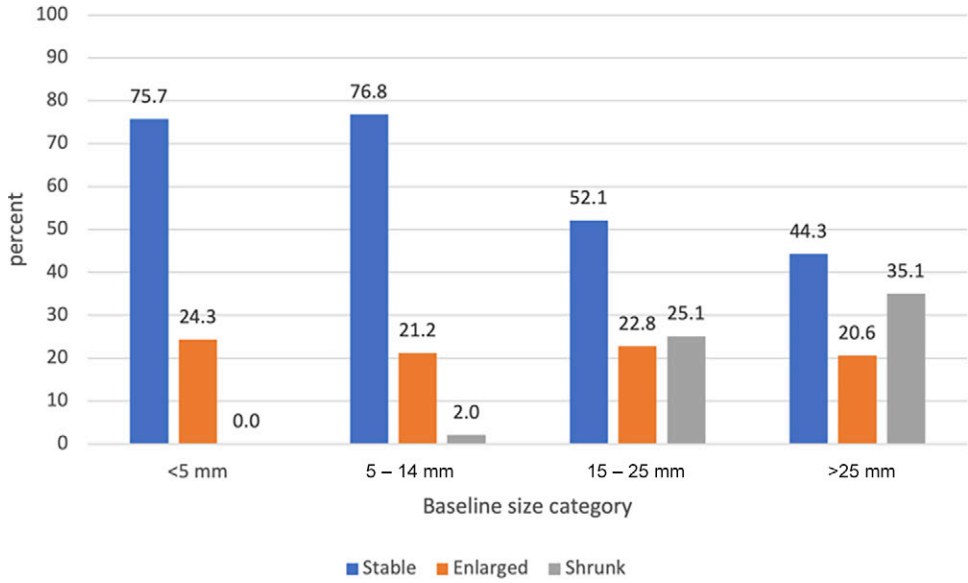

Figure 4: Bar chart shows the proportion of pancreatic cystic lesions with interval growth patterns in the baseline size categories.

For measurement extraction, we pioneered the use of a question answering system; this approach differs from that of previous studies that applied rule-based and/or supervised machine learning–based approaches (7,17). Although we applied the question answering system in an unsupervised manner—that is, the model has never been directly trained for measurement extraction on radiology reports—our results demonstrated the efficacy of such a system for extracting measurements from free-text radiology reports. Our approach successfully ignored prior measurements and measurements for noncystic lesions as shown in Figure 2, which has been challenging for prior NLP approaches, such as the hybrid NLP algorithm that integrates rule-based modules and conditional random field model as developed by Bozkurt et al (7). One possible explanation is that our approach is based on bidirectional encoding that was obtained by the Transformer architecture (10,18,19) that enables the model to understand contexts in free-text documents. It is possible that performance might be improved by directly pretraining the measurement extraction model on a large collection of unlabeled radiology reports; this should be investigated in future studies.

One advantage of our approach is its extensibility or generalizability. Because our NLP pipeline does not require any training datasets or ground truth annotations (ie, it requires only a list of phrases for lesion identification and a set of questions for measurement extraction), our approach can be readily generalized to other use cases, such as measurement extraction for liver and kidney cysts or lesions.

With the extracted measurements, we successfully analyzed interval growth patterns of PCLs on the basis of the ACR size categories and growth criteria. Our results showed that seven of nine PDACs were found in patients with PCLs measuring 5–25 mm, of which three PCLs remained stable throughout the follow-up period. This result suggests that the ACR baseline size group and interval growth may not be sufficient for risk stratification of small PCLs (<25 mm).

Our study was subject to some limitations. First, because the radiology reports were generated through clinical practice and collected from the database, lesion measurements were performed by numerous radiologists and the measurement methods were not standardized (20), especially for the reports generated before publication of the ACR white paper. Since publication of the ACR white paper in 2017 (2), radiologists at our institution have followed the cyst measurement method proposed therein. Nevertheless, measurement variability could potentially have effects on the downstream assessment of growth pattern; this issue was unavoidable because of the retrospective nature of our study. A possible solution is to collect a larger number of reports (eg, from other institutions) so that the effect of variation in measurements can be averaged out. Second, there is no guarantee that the largest measurements for the baseline and the last follow-up examinations were extracted from the same cystic lesion. Therefore, the growth of the cysts may not be reliable, especially when the index cyst shrinks or smaller cysts grow rapidly during the follow-up period. It is necessary to integrate imaging data and cross-reference the cystic lesions to address this limitation, which is outside the scope of the present study. Third, there is no guarantee that the histopathologic diagnoses extracted from pathology reports correspond to the index lesions in the radiology reports in any patient who had more than one PCL. Fourth, the specific type of PCL, such as IPMN or serous cystic neoplasm, was not validated by comparing the text with the radiologists' annotations; therefore, those results should be interpreted with caution. Finally, our PCL identification model does not consider the overlapping nature of entity mentions. Integrating approaches to address this problem, such as conditional random fields (21), hypergraphs (22), and neural networks (23,24), into our approach could improve the performance of PCL identification.

In conclusion, we presented an NLP-based system that identifies patients with PCLs and extracts the largest measurements

of the cysts from a large single-institution archive of free-text radiology reports. Our approach may prove valuable to study the natural history and potential risks of PCLs, and this approach can be applied to many other use cases because it does not require huge, annotated datasets to train models. Further work is needed to explore the extensibility of our NLP model to other large institutional radiology report databases.

Author contributions: Guarantors of integrity of entire study, **R.Y., J.H.D., D.G., T.S.D.**; study concepts/study design or data acquisition or data analysis/interpretation, all authors; manuscript drafting or manuscript revision for important intellectual content, all authors; approval of final version of submitted manuscript, all authors; agrees to ensure any questions related to the work are appropriately resolved, all authors; literature research, **R.Y., T.S.D.**; clinical studies, **K.B., P.Y.C.C., J.H.D., M.N.F., D.G., A.S., A.L.W., T.S.D.**; statistical analysis, **R.Y.**; and manuscript editing, **R.Y., P.Y.C.C., J.H.D., M.N.F., A.L.W., D.L.R., T.S.D.**

Disclosures of conflicts of interest: **R.Y.** No relevant relationships. **K.B.** No relevant relationships. **P.Y.C.C.** No relevant relationships. **J.H.D.** No relevant relationships. **M.N.F.** No relevant relationships. **D.G.** No relevant relationships. **L.N.M.** No relevant relationships. **A.S.** No relevant relationships. **A.L.W.** No relevant relationships. **D.L.R.** Associate editor of *Radiology: Artificial Intelligence*. **T.S.D.** No relevant relationships.

Code availability: The implementation of CheXpert-labeler is available at *https://github.com/stanfordmlgroup/chexpert-labeler*. A list of phrases for mention extraction to modify CheXpert-labeler is available in Supplementary.txt (Appendix E2 [supplement]). The code for measurement extraction is available at *https://github.com/rikiyay/measurement_extraction_from_radiology_reports*. The BioBERT-based question answering model is available at *https://huggingface.co/ktrapeznikov/biobert_v1.1_pubmed_squad_v2*.

References

1. Gardner TB, Glass LM, Smith KD, et al. Pancreatic cyst prevalence and the risk of mucin-producing adenocarcinoma in US adults. Am J Gastroenterol 2013; 108:1546-50.
2. Megibow AJ, Baker ME, Morgan DE, et al. Management of Incidental Pancreatic Cysts: A White Paper of the ACR Incidental Findings Committee. J Am Coll Radiol 2017;14(7):911–923.
3. Moayyedi P, Weinberg DS, Schünemann H, Chak A. Management of pancreatic cysts in an evidence-based world. Gastroenterology 2015;148(4):692–695.
4. Roch AM, Mehrabi S, Krishnan A, et al. Automated pancreatic cyst screening using natural language processing: a new tool in the early detection of pancreatic cancer. HPB (Oxford) 2015;17(5):447–453.
5. Xie F, Chen Q, Zhou Y, et al. Characterization of patients with advanced chronic pancreatitis using natural language processing of radiology reports. PLoS One 2020;15(8):e0236817.
6. Ip IK, Mortele KJ, Prevedello LM, Khorasani R. Focal cystic pancreatic lesions: assessing variation in radiologists' management recommendations. Radiology 2011;259(1):136–141.
7. Bozkurt S, Alkim E, Banerjee I, Rubin DL. Automated Detection of Measurements and Their Descriptors in Radiology Reports Using a Hybrid Natural Language Processing Algorithm. J Digit Imaging 2019;32(4):544–553.
8. Pandey P, Pandey A, Luo Y, et al. Follow-up of Incidentally Detected Pancreatic Cystic Neoplasms: Do Baseline MRI and CT Features Predict Cyst Growth? Radiology 2019;292(3):647–654.
9. Irvin J, Rajpurkar P, Ko M, et al. CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison. arXiv 1901.07031 [preprint] http://arxiv.org/abs/1901.07031. Posted January 21, 2019. Accessed March 8, 2021.
10. Lee J, Yoon W, Kim S, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 2020;36(4):1234–1240.
11. Rajpurkar P, Jia R, Liang P. Know What You Don't Know: Unanswerable Questions for SQuAD. arXiv 1806.03822 [preprint] http://arxiv.org/abs/1806.03822. Posted June 11, 2018. Accessed March 8, 2021.
12. Wolf T, Debut L, Sanh V, et al. HuggingFace's Transformers: State-of-the-art Natural Language Processing. arXiv 1910.03771 [preprint] http://arxiv.org/abs/1910.03771. Posted October 9, 2019. Accessed June 30, 2021.
13. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33(1):159–174.
14. Lin LI. A concordance correlation coefficient to evaluate reproducibility. Biometrics 1989;45(1):255–268.
15. Zhang XM, Mitchell DG, Dohke M, Holland GA, Parker L. Pancreatic cysts: depiction on single-shot fast spin-echo MR images. Radiology 2002;223(2):547–553.
16. de Jong K, Nio CY, Hermans JJ, et al. High prevalence of pancreatic cysts detected by screening magnetic resonance imaging examinations. Clin Gastroenterol Hepatol 2010;8(9):806–811.
17. Sevenster M, Buurman J, Liu P, Peters JF, Chang PJ. Natural Language Processing Techniques for Extracting and Categorizing Finding Measurements in Narrative Radiology Reports. Appl Clin Inform 2015;6(3):600–610.
18. Vaswani A, Shazeer N, Parmar N, et al. Attention Is All You Need. arXiv 1706.03762 [preprint]. http://arxiv.org/abs/1706.03762. Posted June 12, 2017. Accessed March 8, 2021.
19. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv 1810.04805 [preprint] http://arxiv.org/abs/1810.04805. Posted October 11, 2018. Accessed March 8, 2021.
20. McErlean A, Panicek DM, Zabor EC, et al. Intra- and interobserver variability in CT measurements in oncology. Radiology 2013;269(2):451–459.
21. Alex B, Haddow B, Grover C. Recognising nested named entities in biomedical text. In: Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing. Stroudsburg, PA: Association for Computational Linguistics, 2007; 65–72.
22. Lu W, Roth D. Joint Mention Extraction and Classification with Mention Hypergraphs. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA: Association for Computational Linguistics, 2015; 857–867.
23. Shibuya T, Hovy E. Nested Named Entity Recognition via Second-best Sequence Learning and Decoding. arXiv 1909.02250 [preprint] http://arxiv.org/abs/1909.02250. Posted September 5, 2019. Accessed July 1, 2021.
24. Dadas S, Protasiewicz J. A Bidirectional Iterative Algorithm for Nested Named Entity Recognition. IEEE Access 2020;8:135091–135102.