

# Measurement of pancreatic cystic lesions on magnetic resonance imaging: efficacy of standards in reducing inter-observer variability

Dell P. Dunn,<sup>1,2</sup> Olga R. Brook,<sup>1</sup> Alexander Brook,<sup>1</sup> Giselle Revah,<sup>1,3</sup>  
Sumayya Jawadi,<sup>1,4</sup> Maryellen Sun,<sup>1</sup> Karen S. Lee,<sup>1</sup> Koenraad J. Mortele<sup>1</sup>

<sup>1</sup>Department of Radiology, Beth Israel Deaconess Medical Center, 330 Brookline Avenue, Ansin 224, Boston, MA 02115, USA

<sup>2</sup>Department of Radiology, David Grant Medical Center, Travis AFB, CA, USA

<sup>3</sup>Medical Imaging Department, Ottawa Hospital, Ottawa, ON, Canada

<sup>4</sup>Department of Radiology, St. Vincents Hospital, Worcester, MA, USA

## Abstract

**Purpose:** The purpose of this study is to assess inter-observer variability in the measurement of pancreatic cystic lesions with MRI and to determine the impact of measurement standards.

**Materials and methods:** In this IRB-approved, HIPAA-compliant study with waiver of informed consent, 144 MRI examinations, containing pancreatic cystic lesions measuring between 5 and 35 mm, were reviewed independently by two radiology attendings and two abdominal imaging fellows. Measurements were repeated by the same reviewers 12 weeks later after the introduction of measurement standards. Results were analyzed using within-subject standard deviation, intraclass correlation coefficient, and kappa.

**Results:** Prior to standardization, the within-subject standard deviation, showing measurement variability in each cyst, was 4.0 mm, which was reduced to 3.3 mm after introduction of measurement standards ( $p < 0.01$ ). Overall inter-observer agreement, kappa, improved from 0.59 to 0.65 ( $p = 0.04$ ). The frequency of all four reviewers agreeing on size category increased from 51% to 60%. The intraclass correlation coefficient increased from 0.81 to 0.86.

**Conclusions:** There is significant and frequent inter-observer variability in the measurement of pancreatic cystic lesions with MRI which could affect clinical management. Implementation of measurement standards reduces measurement variability and aids in preventing

erroneous reporting of growth and potentially unwarranted changes in management.

**Key words:** IPMN—Pancreatic cysts—MRI—Pancreas—Variability

Incidental focal pancreatic cystic lesions (FPCL) are present in up to 20% of all abdominal MRI examinations and the incidence increases to 40% in patients older than 70 years of age [1]. The large majority of these incidental cysts are benign side-branch intraductal pancreatic mucinous neoplasms (IPMN); however, a fraction can be dysplastic or malignant [2–4].

Several management guidelines have been developed for incidental pancreatic cystic lesions [5–7]. These guidelines employ a combination of single length size cutoffs, growth on follow-up exams, presence of symptoms, and imaging characteristics to direct management decisions including frequency of follow-up imaging, invasive testing such as endoscopic ultrasound (EUS) guided fine needle aspiration and resection. None of these guidelines, however, include standards for the measurement of pancreatic cystic lesions nor do they define what constitutes growth on follow-up.

An additional challenge lies in the configuration of FPCLs—they are often pleomorphic in shape, clustered, or multiple which can make consistent measurements on serial follow-up exams difficult (Fig. 1). Furthermore, when interpreting MRI with magnetic resonance cholangiopancreatography (MRCP), the radiologist must choose between axial or coronal, T1- or T2-weighted, MRCP or post-contrast sequences to measure

the lesion (Fig. 2). Significant variability has been reported when measuring FPCLs between different imaging modalities [8]. These factors and the lack of measurement standards result in a potential for significant inter-observer variability in the measurement of pancreatic cystic lesions.

Therefore, the purpose of this study was to quantify inter-observer variability in the measurement of FCPLs on MRI and whether this variability could affect clinical management. We then introduced measurement standards and assessed their impact on measurement variability.

## Materials and methods

### Subjects

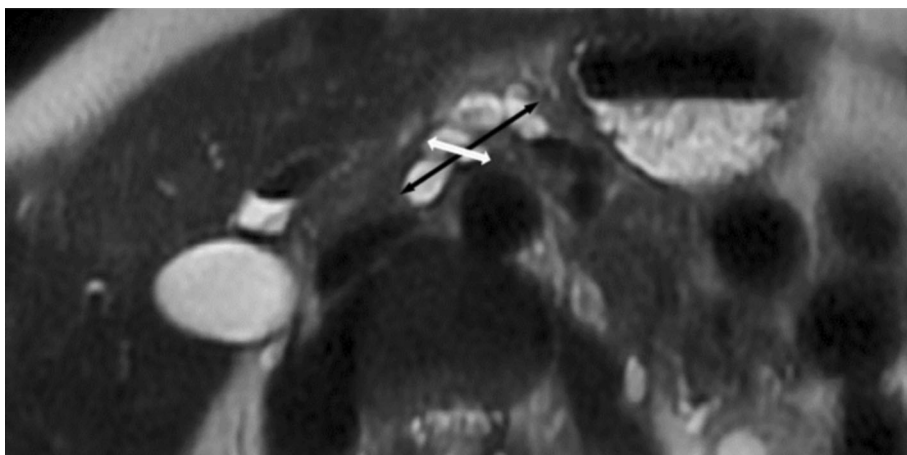
In this prospective, IRB-approved, HIPAA-compliant study, with waiver of informed consent, 144 MRI examinations containing at least one FPCL were selected from a database of over 700 MRIs with FPCLs acquired between 2002 and 2011. These exams were randomly selected from six size groups that bracket management decision points (10, 20, 30 mm) in the Sendai and ACR criteria [5–7]. The size of the cysts was determined independently at the time of inclusion in the larger database and this measurement otherwise did not contribute to the study. Power analysis required 24 exams in each size group (5–10, 11–15, 16–20, 21–25, 26–30 and 31–35 mm). Factors such as lesion morphology, imaging appearance, and patient characteristics were not considered in exam selection. Personally identifiable information (PII), previous measurements, and prior reports were stripped from the exams and they were uploaded onto the hospital picture archiving and communication system (PACS) (GE Medical, Waukesha, WI) with a unique study identifier.

### Image acquisition

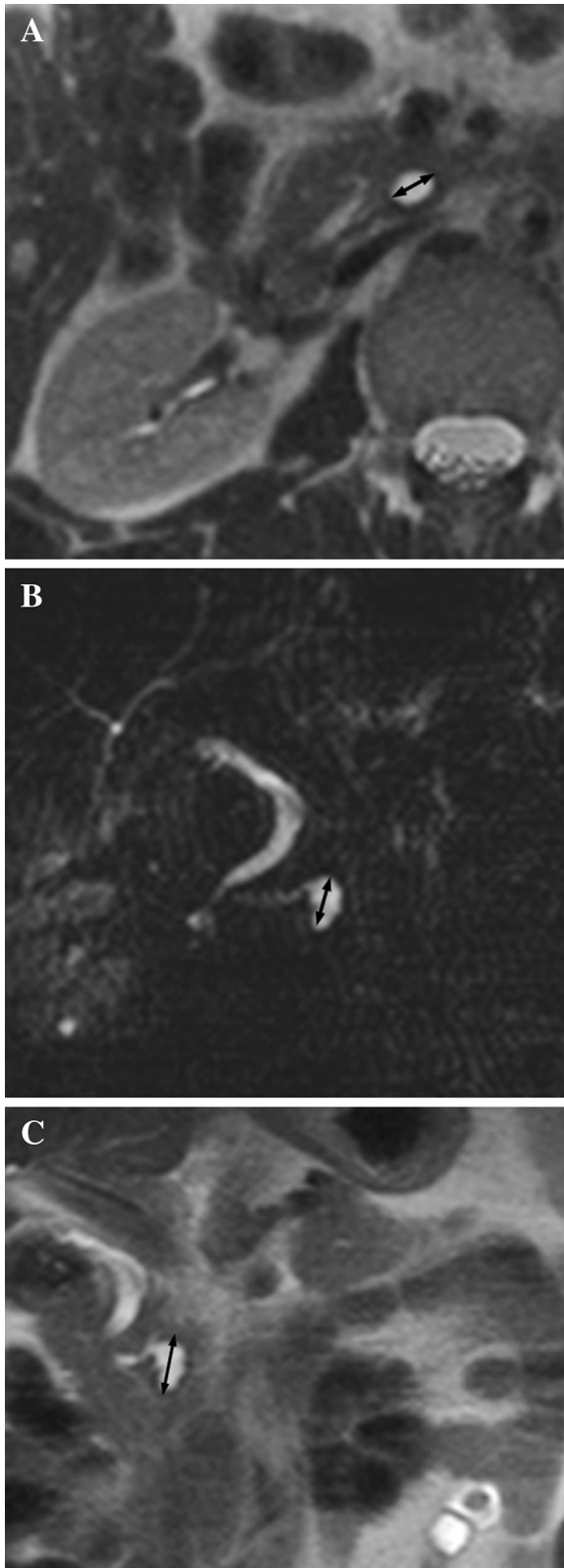
The only criterion for inclusion of an exam in the study was the presence of a FPCL. As such, MRI exams included in the study reflect common clinical practice with exams acquired on multiple magnets from multiple vendors at 1.5 and 3.0 Tesla field strengths in a tertiary care hospital, community hospital, and imported from outside institutions. Protocols were variable depending on the clinical indication, location of study, and date of the study. Use of oral and intravenous contrast and MRCP sequences in the study exams was also variable.

### Image analysis

Four reviewers, two attending abdominal imagers, specializing in body MRI, each with 5 years of post-fellowship experience (M.S. and K.S.L.) and two abdominal imaging fellows (G.R. and S.J.) measured each FPCL twice—once according to their regular clinical practice and again 12 weeks later, after the introduction of, and training with, measurement standards. Greatest single length measurement was used according to the RECIST criteria for both measurements [9]. The interval between measurement sessions was intended to reduce recall bias. Greatest single length measurement, series number, and image number were recorded on an electronic worksheet that indicated the anatomic location of each cystic lesion in the pancreas but without image number or previously measured size (i.e., posterior aspect of the uncinate process). Prior measurements before standards introduction were not available during the second round of measurements. Each reviewer performed measurements independent and separate of other reviewers and was blinded to any previous measurements. All exams series and images were available to the reviewers.



**Fig. 1.** A 67-year-old asymptomatic female with clustered cystic lesions in the pancreas. Measured size of this cluster of cysts varies dramatically depending on whether the largest individual cyst is measured (*white arrow*) or if the cluster is measured as a single lesion (*black arrow*).



◀**Fig. 2.** A 61-year-old male with an incidentally discovered FPCL depicting variable size measurement depending on image selection and measurement technique. **A–C**, axial T2 weighted image (**A**) measured the inner wall distance as 0.8 cm. Coronal MIP image from a 3D MRCP sequence (**B**) measured the inner wall distance as 1.6 cm. Coronal T2 weighted image (**C**) measured the outer wall distance as 2.0 cm.

for measurement and using the same measurement technique—and the standards were designed to accomplish these goals. Measurement standards (Fig. 3) were introduced in a group training session where examples of the standardized measurement approach were shown and questions could be asked and discussed as a group to ensure identical training for each reviewer.

### *Statistical analysis*

Within-subject standard deviation was used as the principal measure of variability. Variability was also assessed using intraclass correlation coefficient (ICC) for overall absolute agreement. Sub-groupings for statistical analysis were based on median measured size. Overall and pairwise Kappa ( $\kappa$ ) was used to assess agreement on management category. Agreement values were interpreted according to the following scale: 0.01–0.20, slight agreement; 0.21–0.40, fair agreement; 0.61–0.80, substantial agreement; 0.41–0.60, moderate agreement; and 0.81–1.0, excellent agreement [11]. Lesions where two reviewer measurements differed by more than 10 mm or 50% were considered at risk for erroneous reporting of growth.

Z test,  $\chi^2$  test, and McNemar's test were used for comparisons. The level of statistical significance was set at  $p = 0.05$ . Statistical analysis was carried out using Matlab (Mathworks, Natick, MA).

## **Results**

### *Size measurements*

Figure 4 shows the distribution of all 144 measurements by each reviewer before (Fig. 4A) and after (Fig. 4B) the introduction of measurement standards. There is initially a systematic bias towards smaller measurements seen in one of the readers (Fellow 2, cyan, Fig. 4A) that disappears after introduction of standards (Fellow 2, cyan, Fig. 4B). Clear outliers are seen in both datasets which importantly impact the resulting data.

The mean within-subject standard deviation (SW), a measurement of the variability for each lesion measured, decreased in all subgroups after introduction of standard measurements, improving from 4.0 to 3.3 mm (Table 1,  $p < 0.01$ ). SW was noted to increase with size of the lesion being measured (Fig. 5).

### *Measurement standards*

We identified two intermediate objectives to reduce variability between reviewers—choosing the same image

Measurement Standards
1. Measure the longest outer-wall to outer-wall dimension on a single coronal T2W image
2. Axial images or MIP/slab MRCP images can be used for localization
3. If cystic lesion is not well seen on coronal T2WI, then measurement may be obtained from an axial image or slab/MIP MRCP image in that order
4. A neck or linear connection to the pancreatic duct is not to be included

Fig. 3. Measurement standards for pancreatic cysts on MRI provided to reviewers in a group training after the first round of measurements (MIP = maximal intensity projection).

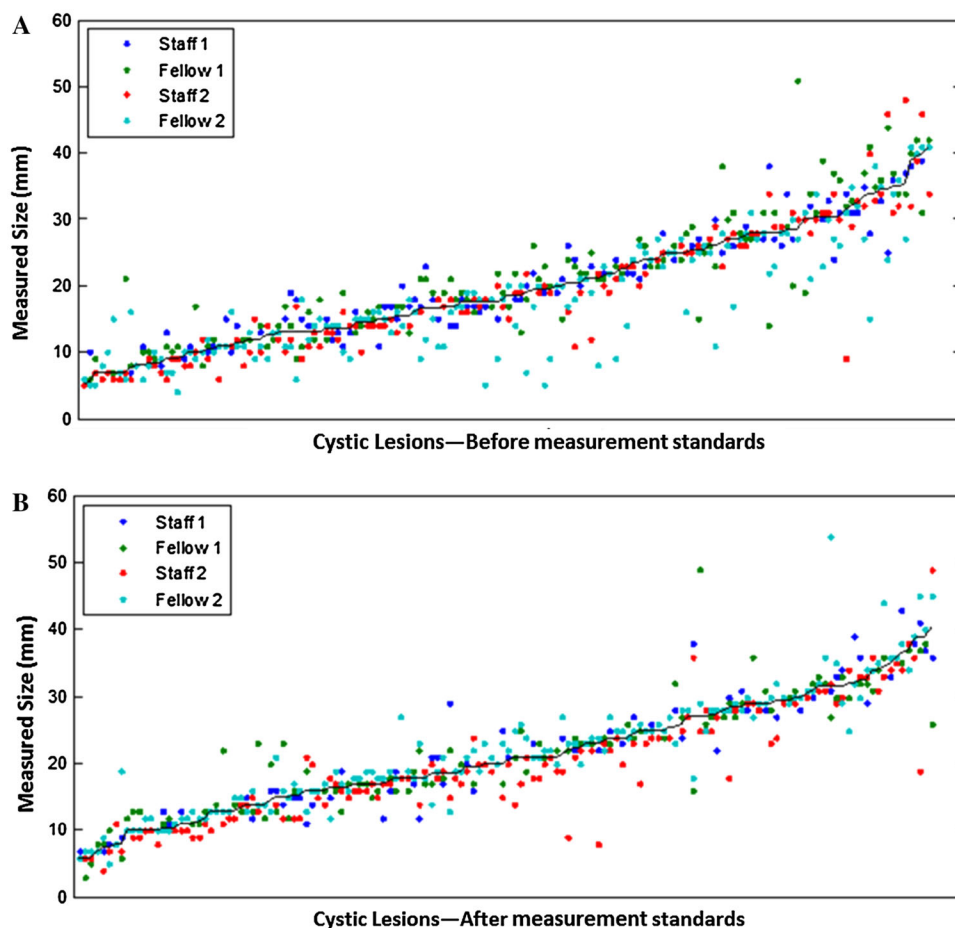


Fig. 4. **A–B**, Distribution of all measurements, color-coded by reviewer, for all 144 FPCLs before (**A**) and after (**B**) the implementation of measurement standards. The cysts are

sorted along the x-axis by the median of the 4 measurements, which is also shown as *thin black line*.

The measure of absolute size agreement, the ICC, increased from 0.81 to 0.86 (excellent agreement), though this result did not reach statistical significance ( $p = 0.20$ ).

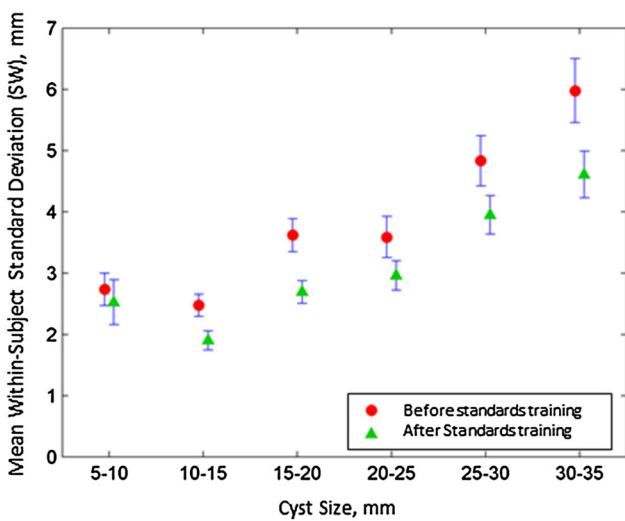
### Management categories

Agreement between reviewers on management category (<10, 10–19, 20–29, ≥30 mm) as measured by  $\kappa$  was

**Table 1.** Variability for measurements of each FPCL before and after introduction of standards as measured by the within-subject standard deviation (SW) and its standard error

Measurement variability by cyst size and location				
Size (mm)	Round 1-before standards		Round 2-after standards	
	# cysts	SW (mm)	# cysts	SW (mm)
5–9	18	2.73 ± 0.26	8	2.52 ± 0.36
10–14	32	2.47 ± 0.18	25	1.90 ± 0.16
15–19	30	3.62 ± 0.27	35	2.69 ± 0.19
20–24	19	3.58 ± 0.34	26	2.96 ± 0.24
25–29	23	4.83 ± 0.41	26	3.95 ± 0.32
30+	22	5.98 ± 0.52	24	4.61 ± 0.38
Location				
Head/neck	85	4.42 ± 0.20	85	3.75 ± 0.17
Body/tail	59	3.26 ± 0.17	59	2.41 ± 0.13
Overall	144	3.99 ± 0.14	144	3.27 ± 0.11

*p* < 0.01



**Fig. 5.** Variability of FPCLs bracketed by 95% confidence intervals based on lesion size before and after the implementation of measurement standards.

**Table 2.** Overall and pairwise Fleiss  $\kappa$  comparisons between all reader pairs

Kappa ( $\kappa$ ) comparison for assignment to size-based management category	Prior	After standards	
	to standards	implementation	
Pairwise comparison			
Attending 1 and fellow 1	0.66	0.69	
Attending 1 and attending 2	0.69	0.67	
Attending 1 and fellow 2	0.56	0.78	
Fellow 1 and attending 2	0.60	0.58	
Fellow 1 and fellow 2	0.46	0.70	
Attending 2 and fellow 2	0.56	0.61	
All readers	0.59	0.67	<i>p</i> 0.006

0.59 ± 0.02 (moderate agreement) for measurements before introduction of standards and 0.67 ± 0.02 (substantial agreement) afterwards (*p* < 0.01). All 4 readers agree on the size category in 73 cases (51%) initially and

in 86 cases (60%) after training (*p* = 0.06). Of the lesions with at least one measurement greater than or equal to 30 mm, a clinically significant cutoff, there was agreement on management category in 14/38 (37%) cases before training and 20/38 (53%) afterwards (*p* = 0.08).

Pairwise comparisons are seen in Table 2. There is initially highest agreement ( $\kappa$  = 0.69) on size category between the attending readers and least agreement ( $\kappa$  = 0.46) between the two fellows. Following introduction of standards, there was no difference in agreement between more ( $\kappa$  = 0.67) and less experienced readers ( $\kappa$  = 0.70).

### Image selection

Table 3 shows the effectiveness of standards in encouraging readers to take measurement from the same sequences. All readers took measurements from the same series in 26/144 (18%) of cases initially and 105/144 (73%) after introduction of standards (*p* < 0.01). Table 4 shows pairwise comparison between readers choosing almost the same image, meaning an image on the same series within one slice of the other. Paired reviewers choose the same image (±1 slice) in an average of 57 out of 144 or 40% of cases before, and 114 out of 144 or 79% of cases after standards implementation (*p* < 0.01).

Exams were not selected based on location of the FPCL within the pancreas; however, Table 1 shows variability was greater in the measurements of lesions in the head, neck, and uncinate process of the pancreas (SW = 4.42 ± 0.20) than in the body and tail (SW = 3.26 ± 0.17, *p* < 0.01). This difference persisted after introduction of measurement standards (SW = 3.75 ± 0.17 and 2.41 ± 0.13, *p* < 0.01).

Increase in size of a lesion ≥10 mm or ≥50% on follow-up has been used to define growth of a FPCL [10] and could trigger a change in management. Initially, in 62 of the 144 lesions (43.1%) at least two of the four readers had measurement discrepancies which met these

**Table 3.** Number of different MR image series used by reviewers for the 144 measurements before and after introduction of measurement standards

Effect of standards on image selection				
Number of series used	Round 1-before standards		Round 2-after standards	
	<i>N</i>	%	<i>N</i>	%
1	26	18.1	105	72.9
2	58	40.3	34	23.6
3	53	36.8	5	3.5
4	7	4.9	0	0

**Table 4.** Frequency of readers using the same image  $\pm 1$  slice measuring 144 pancreatic cysts

Frequency of 2 reviewers choosing the same image $\pm 1$ slice for measurement				
Readers	Prior to standards implementation		After standards implementation	
	<i>N</i>	%	<i>N</i>	%
Staff 1 and fellow 1	60	42%	112	78%
Staff 1 and staff 2	54	36%	117	81%
Staff 1 and fellow 2	46	32%	114	79%
Fellow 1 and staff 2	62	43%	115	80%
Fellow 1 and fellow 2	47	33%	111	77%
Staff 2 and fellow 2	73	51%	116	81%

criteria. After introduction of standards, only 30 of the 144 (20.8%) measured lesions met these criteria for change, a 51.6% reduction ( $p < 0.01$ ).

## Discussion

Quality, performance management, and process improvement have become a focus of radiology departments and national meetings over the last decade with the goal of improving value, consistency, and safety. One key area of quality improvement is identifying and eliminating variability in the diagnosis and management of disease and the consistent implementation of practice guidelines.

Significant variability has been shown when measuring FPCLs using differing modalities such as computed tomography (CT), MRI, and EUS [12]. There is also significant variability in the management recommendations made by interpreting radiologists [13, 14]. Inter-observer agreement using MRI in identifying and categorizing concerning features has been found to be especially low [15]. Do et al. also studied inter-observer variability in identification of concerning features and single length measurement specifically of IPMNs with CT [16]. To the best of our knowledge, however, inter-observer variability in FPCL size measurements on MRI has not been previously studied.

## Measurement variability

The results of this study indicate that significant inter-observer variability exists in the measurement of focal pancreatic cystic lesions on MRI and the variability (as represented by the ICC) for FPCLs in the first phase of this study (ICC = 0.81) is similar to that reported for CT measurements of IPMNs (0.85) [15] and more than the variability reported for lung nodules (0.97), lymph nodes (0.93), and liver tumors (ICC = 0.95) [16].

Hopper et al. reported increased variability in the measurement of irregular tumors in the abdomen which certainly applies to FPCLs [17]. Sources of variability in this study included factors intrinsic to these lesions, such as irregular shape or clustered cysts, location of the lesion in the head of the pancreas, the image selected for measurement, and individual measurement techniques that lead to systematic over- or under-measurement. Maimone et al. found an average of 3–4 mm absolute size difference when comparing measurements acquired from at least 2 different modalities (CT, MRI, EUS) within a 90-day window [8]. They suggested the source of the variability was the different modalities. However, in this single modality study, the reported mean inter-observer within-subject standard deviation (SW) was 3.99 mm, which is similar to the inter-modality variability Maimone et al. reported.

De Jong et al. studied the inter-observer agreement using MRI in identification and categorization of concerning features in pancreatic cysts [15]. Somewhat surprisingly, they found only poor to moderate agreement between four experienced radiologists for identifying concerning characteristics (ICC = 0.23–0.53), though the use of ICC for these categorical comparisons is questionable. Do et al. recently reported similarly limited inter-observer agreement using CT for characterization of IPMN subtype, presence of nodules, and solid components ( $\kappa = 0.28$ –0.46). They found much better agreement on continuous variables such as maximal dimension (ICC = 0.85), pancreatic duct size (ICC = 0.75), and common bile duct size (ICC = 0.61) [15]. Of note, the average size of the IPMNs in that dataset was 35 mm (11–83 mm), while the FPCLs studied in this study averaged 20 mm (4–54 mm). While agreement on absolute size measurements in our study (ICC = 0.81 and 0.86) was in a range similar to that reported by Do et al., our moderate to substantial agreement on assignment to management categories ( $\kappa = 0.59$  and 0.67) along with the results by Do et al. on feature analysis points to the difficulty in consistent MRI evaluation of these lesions by different radiologists.

Analysis of the largest measurement discrepancies in our dataset showed that readers sometimes measured a different lesion than the one intended and measured by others. While specific anatomic descriptors were pro-

vided (i.e., pancreatic body, superior to the duct), the presence of multiple cystic lesions in the pancreas or even a segment can make consistent identification and repeated measurement of the same lesion difficult. A second source of large measurement discrepancies was clustered or septated cysts. Each reviewer had to decide whether to measure the lesion as a single unit or measure the largest component.

Location of the lesions within the pancreas also contributed to variability. Measurements of lesions within the head, neck, and uncinate process of the pancreas demonstrated significantly more inter-observer variability than those in the body and tail. This difference may be attributable to the change in orientation of the pancreas (horizontal in the body/tail vs. more vertical in the head) or increased complexity of the surrounding anatomy in the region of the pancreatic head and uncinate process.

Considering the effect of experience on inter-observer variability, on initial measurements, there was less inter-observer variability between the two most experienced readers,  $\kappa = 0.69$ , than between the two least experienced readers,  $\kappa = 0.46$ . This suggests that experience does contribute to variability in cyst measurement.

Contrary to results from McErlean et al. [17] as well as the 2010 ACR Consensus Statement [7], absolute variability increased with increasing lesion size as measured by the within-subject standard deviation. This difference may be partly attributable to the difficulties involved with measuring clustered cysts.

## Clinical implications

Our data show that there is significant inter-observer variability in the measurement of FPCLs and that this variability could have a substantial impact on clinical management. All four reviewers agreed on the management size category for a given lesion only 51% of the time prior to standards training compared to 60% after training. Furthermore, the agreement between reviewers was worse for those lesions which approached the clinically important cutoff of 30 mm where guidelines recommend invasive testing [5–7].

Growth on follow-up scans is a critical image-based factor guiding management of FPCLs. The ACR statement explains that no consensus exists for what defines interval growth [7] and the revised Sendai guidelines only state follow-up interval can be lengthened if there is no change after 2 years [6]. Das et al. used a cutoff of 10 mm or 50% single length measurement increase from initial measurement to define growth [11]. In our study, measurement differences of  $\geq 10$  mm or 50% between any two of the four reviewers occurred in 62/144 exams (43.1%) before the introduction of standards.

## Effects of standards

The use of measurement standards was successful in accomplishing the intermediate goal of homogenizing image selection. The frequency of all 4 readers choosing the same series for measurement increased from 18% to 73% and reader pairs choosing the same image  $\pm 1$  slice increased from 39.5% to 79.1% after measurement standards training. Laxity in the measurement standards allowing reviewers to use an alternate plane if they found the coronal plane difficult to use likely contributed to the remaining discrepancy in image. Measurement of a nearby non-target lesion occurred regardless of the implementation of measurement standards, also contributing to discrepancies in image selection.

Measurement standards were also successful in reducing erroneous reporting of lesion change from 62/144 (43.1%) cases to 30/144 (20.8%) cases. While this rate remains high, it does indicate that measurement standards can be effective reducing unwarranted reporting of lesion change.

The use of measurement standards reduced inter-observer variation in assignment of lesions to size-based management groups as represented by the increase in  $\kappa$  ( $p < 0.01$ ). The improvement in agreement on absolute size measurement (ICC) approached significance ( $p = 0.2$ ).

The difference in  $\kappa$  observed in pairings between the two staff reviewers and the two fellow reviewers disappeared with the fellow pair increasing their agreement from  $\kappa = 0.46$  to  $\kappa = 0.70$  which was the second highest rate of agreement among the six reader pairings. The remaining inter-observer variability (ICC = 0.86) after implementation of measurement standards for FPCLs on MRI, as compared to the high agreement McErlean et al. [17] reported for CT measurements of liver masses and lymph nodes (ICC = 0.93–0.95), may be due to several factors including pleomorphic nature of FPCLs, incomplete acceptance of measurement standards by the readers, erroneous measurement of a different, non-target lesion, or clustered cystic lesions.

One of the limitations of this study is that unlike clinical practice, reviewers had no access to comparison images or prior reports. Frequently, radiologists use image references from prior reports to ensure similar measurements or simply re-measure the lesion on the comparison exam themselves. As such, the variability we report may be exaggerated compared to results in actual practice. While not the focus of this study, this limitation does highlight the utility of including an image reference in reports or saving measurement markings when interpreting studies containing FPCLs.

Another limitation of the study is the shared background of the reviewers. The two attending reviewers trained at the same institution for both residency and fellowship. This common background could have led to a

more standardized approach which could have artificially reduced variability compared to reviewers with more varied training backgrounds. Also, the 4 reviewers in this study may not be representative of radiologists elsewhere in their initial measurement approach or acceptance of standards.

The exams in the study came from many different institutions over a long period with different techniques and image quality which may affect measurement variability, but reflects our practice of reviewing outside images and old comparison studies.

## Conclusion

This study shows that there is significant inter-observer variability in the measurement of pancreatic cystic lesions which can crucially affect clinical management. Larger lesion size, location of the lesion in the pancreatic neck, head, or uncinata process, and less reader experience all increased measurement variability in this study.

The measurement standards introduced in this study were effective in homogenizing image selection, reducing unwarranted reporting of change, reducing variability attributable to reader experience, and improving agreement on management categories.

Based on these findings, we recommend that future management guidelines of pancreatic cystic lesions should include standards for measurement and define parameters that constitute growth. The measurement standards used in this study were effective and could be used until further research further validates these standards or identifies more effective standards.

## References

1. Lee KS, Sekhar A, Rofsky NM, et al. (2010) Prevalence of incidental pancreatic cysts in the adult population on MR imaging. *Am J Gastroenterol* 105:2079–2084
2. Lee CJ, Scheiman J, Anderson MA, et al. (2008) Risk of malignancy in resected cystic tumors of the pancreas < or = 3 cm in size: is it safe to observe asymptomatic patients? A multi-institutional report. *J Gastrointest Surg* 12:234–242
3. Levy P, Jouannaud V, Otoole D, et al. (2006) Natural history of intraductal papillary mucinous tumors of the pancreas: actuarial risk of malignancy. *Clin Gastroenterol Hepatol* 4:460–468
4. Kimura W, Nagai H, Kuroda A, et al. (1995) Analysis of small cystic lesions of the pancreas. *Int J Gastrointest Cancer* 18:197–206
5. Tanaka M, Chari S, Adsay V, et al. (2006) International consensus guidelines for management of intraductal papillary mucinous neoplasms and mucinous cystic neoplasms of the pancreas. *Pancreatology* 6:17–32
6. Tanaka M, Fernández-del Castillo C, Adsay V, et al. (2012) International consensus guidelines 2012 for the management of IPMN and MCN of the pancreas. *Pancreatology* 12:183–197
7. Berland LL, Silverman SG, Gore RM, et al. (2010) Managing incidental findings on abdominal CT: white paper of the ACR incidental findings committee. *J Am Coll Radiol* 7:754–773
8. Maimone S, Agrawal D, Pollack MJ, et al. (2010) Variability in measurements of pancreatic cyst size among EUS, CT, and magnetic resonance imaging modalities. *Gastrointest Endosc* 71:945–950
9. Eisenhauer EA, Therasse P, Bogaerts J, et al. (2009) New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). *Eur J Cancer* 45:228–247
10. Landis JR, Koch GG (1977) The measurement of observer agreement for categorical data. *Biometrics* 33:159–174
11. Das A, Wells CD, Nguyen CC (2008) Incidental cystic neoplasms of pancreas: what is the optimal interval of imaging surveillance? *Am J Gastroenterol* 103:1657–1662
12. Ip IK, Morteke KJ, Prevedello LM, Khorasani R (2011) Focal cystic pancreatic lesions: assessing variation in radiologists' management recommendations. *Radiology* 259:136–141
13. Macari M, Megibow AJ (2011) Focal cystic pancreatic lesions: variability in radiologists' recommendations for follow-up imaging. *Radiology* 259:20–23
14. De Jong K, Nio CY, Mearadji B, et al. (2012) Disappointing interobserver agreement among radiologists for a classifying diagnosis of pancreatic cysts using magnetic resonance imaging. *Pancreas* 41:278–282
15. Do RKG, Katz SS, Gollub MJ, et al. (2014) Interobserver agreement for detection of malignant features of intraductal papillary mucinous neoplasms of the pancreas on MDCT. *AJR* 203:973–979
16. Hopper KD, Kasales CJ, Van Slyke MA, et al. (1996) Analysis of interobserver and intraobserver variability in CT tumor measurements. *AJR* 167:851–854
17. McErlean A, Panicek DM, Zabor EC, et al. (2013) Intra- and interobserver variability in CT measurements in oncology. *Radiology* 269:451–459