Provider Specific Quality Measurement for Endoscopic Retrograde Cholangiopancreatography

Provider-specific quality measurement for ERCP using natural language processing

Timothy D. Imler[1-3], Stuart Sherman[1,2], Thomas F. Imperiale[1,2,4,5], Huiping Xu[6], Fangqian Ouyang,[6] Christopher Beesley[3], Charity Hilton[3], and Gregory A. Coté[1,7]

1. Division of Gastroenterology and Hepatology; Indiana University School of Medicine. Indianapolis, IN.
2. Department of Medicine; Indiana University School of Medicine. Indianapolis, IN.
3. Department of Biomedical Informatics; Regenstrief Institute, LLC. Indianapolis, IN.
4. Center of Innovation; Health Services Research and Development; Richard L. Roudebush VA Medical Center. Indianapolis, IN.
5. Health Services Research, Regenstrief Institute, LLC. Indianapolis, IN.
6. Department of Biostatistics; Indiana University School of Medicine. Indianapolis, IN
7. Division of Gastroenterology and Hepatology; Medical University of South Carolina. Charleston, SC.

Keywords

Endoscopic retrograde cholangiopancreatography, quality measurement, natural language processing

Abbreviations

ERCP = Endoscopic retrograde cholangiopancreatography

NLP = Natural language processing

Contributions

Study design (Imler/Cote/Imperiale/Sherman); Data collection (Imler/Hilton/Beesley); Data analysis (Imler/Cote/Xu/Ouyang/Sherman); Statistical analysis (Imler/Xu/Ouyang/Imperiale); Manuscript drafting (Imler); Critical editing (All listed authors)

Conflicts of Interest

The authors disclose that Dr. Imler has filed for provisional patent (IURTC-14098-01-US-E) for similar work (colonoscopy quality) under the name Tracking Real-time Assessment of Quality Monitoring in Endoscopy (TRAQ-ME) through the Indiana University Research and Technology Corporation (IURTC).

Correspondence

Timothy D. Imler, M.D., M.S.
Assistant Professor of Medicine; Division of Gastroenterology and Hepatology Research Scientist; Regenstrief Institute
702 Rotary Circle; Suite 225; Indianapolis, Indiana 46202
Office: (317) 274-9137; Fax: (317) 278-6870; Email: timler@iu.edu

Word Count (abstract): 276

Duplicate/Previous Publications: None

Grant Support

Provider Specific Quality Measurement for Endoscopic Retrograde Cholangiopancreatography

This work was performed at the Regenstrief Institute, Indianapolis, Indiana, and was supported in part by the American Society for Gastrointestinal Endoscopy Covidien Senior Investigator Mentoring Award (Imperiale) and the American Society for Gastrointestinal Endoscopy Career Development Award (Imler).

Provider Specific Quality Measurement for Endoscopic Retrograde Cholangiopancreatography

Background: Natural language processing (NLP) is an information retrieval technique that has been shown to accurately identify quality measures for colonoscopy. There are no systematic methods by which to track adherence to quality measures for ERCP, the highest risk endoscopic procedure widely used in practice.

Aim: Our aim was to demonstrate the feasibility of using NLP to measure adherence to ERCP quality indicators across individual providers.

Methods: ERCPs performed by 6 providers at a single institution from 2006 to 2014 were identified. Quality measures were defined using society guidelines and from expert opinion, and then extracted using a combination of NLP and data mining (eg, ICD-9 CM codes). Validation for each quality measure was performed by manual record review. Quality measures were grouped into preprocedure (5), intra-procedure (6), and postprocedure (2). NLP was evaluated using measures of precision and accuracy.

Results: A total of 23,674 ERCPs were analyzed (average patient age of 52.9 ± 17.8, 14,113 (59.6%) women). Among 13 quality measures, precision of NLP ranged from 84% to 100% with intraprocedure measures having lower precision (84% for precut sphincterotomy). Accuracy of NLP ranged from 90% to 100% with intraprocedure measures having lower accuracy (90% for pancreatic stent placement).

Conclusion: NLP in conjunction with data mining facilitates individualized tracking of ERCP providers for quality metrics without the need for manual medical record review. Incorporation of these tools across multiple centers may permit tracking of ERCP quality measures through national registries.

Background

Quality measurement of endoscopy is becoming the standard of care in the United States[1-5] and may influence choice in provider, outcomes, and reimbursement[5-7]. ERCP, the highest risk endoscopic procedure in widespread practice,[8] has not been extensively studied for individual, endoscopic-based quality measures[9]. Historically, ERCP quality has focused on provider or facility volume, with higher volumes associated with higher quality as defined by success and adverse event rates after adjusting for procedure indication[10, 11]. In 2006, the American Society for Gastrointestinal Endoscopy (ASGE) and the

Provider Specific Quality Measurement for Endoscopic Retrograde Cholangiopancreatography

American College of Gastroenterology (ACG) Task Force on Quality Endoscopy provided the first quality indicators for ERCP, many based on expert consensus[9], and these were subsequently updated in 2014[1].

Similar to colonoscopy, there are challenges in obtaining ERCP-based quality measures due to the time intensive nature of manual medical record review. Given this challenge in colonoscopy, several studies have demonstrated the feasibility of using natural language processing (NLP) to extract these measures from text documents in the medical record, with >90% accuracy for colonoscopy-specific measures[12-16]. We hypothesized that NLP, in conjunction with data mining using standard billing codes, could be used to track ERCP quality measures accurately and efficiently. If successful, NLP could be incorporated into health systems that aim to monitor ERCP quality and provide feedback to providers, administrators, and payers in an effort to show adherence to national benchmarks; if needed, these data could be used to refine local practices through quality improvement initiatives[17].

The primary aim of this study was to measure the precision and accuracy of NLP in assessing ERCP-specific quality measures. The secondary aim was to provide pilot data on variability of adherence to quality benchmarks among individual providers at a single institution.[17]

Methods

After Institutional Review Board (IRB) approval, we identified ERCP procedure reports and related clinical data from January 1, 2006 through July 25, 2014.

Data Sources

ERCP procedure reports were identified from the Indiana Network for Patient Care (INPC), [18] which is a large regional health information exchange that obtains data from healthcare organizations and payers[19, 20]. The database houses more than 4 billion pieces of clinical data with over 160 million text reports. All providers in this study work at a single center (Indiana University Health; University Hospital), a high-volume referral center with approximately 3000 ERCP procedures performed annually. All ERCP procedure reports are stored within the INPC and were created using a single point-and-click endowriter (Provation® MD; Wolters Kluwer). Clinical and payer data sources facilitate pairing ERCP reports with

procedure indications (International Classification of Diseases, Ninth Revision, Clinical Modification, ICD9-CM).

Natural Language Processing System

The Regenstrief Institute has created an Apache Unstructured Information Management Applications (UIMA™)[21] based NLP system (nDepth) that uses open-source applications for NLP processing released under the Apache license version 2.0.[22] In addition to search, more advanced NLP techniques (e.g. negation, regular expressions, and standard terminologies) are available through the system and were used as part of this study.

Selection of Quality Measures

Quality measures were identified based on the 2014 ASGE/ACG Quality Indicators for ERCP that were reviewed and endorsed by the American Society for Gastrointestinal Endoscopy (ASGE), the American College of Gastroenterology (ACG), and the American Gastroenterological Association (AGA)[1]. Measures were categorized as (1) preprocedure, (2) intraprocedure, and (3) postprocedure. Four additional measures (pre-cut sphincterotomy, cannulation of the pancreatic duct, injection of the pancreatic duct, and pancreatic duct stent placement) were added from internal discussion among the authors for exploratory analysis (Table 1). Table 1 also includes all ASGE/ACG quality indicators that were not included in this study and rationale for their exclusion. We addressed rectal indomethacin as an alternative to pancreatic stenting by introducing quality measure (QM10^) that allows for either option within the endoscopy report. Quality measures that did not require text extraction via NLP were extracted by an INPC data manager. These quality measures were extracted using ICD9-CM) codes.

Validation of Quality Measures identified by NLP

Documents listed as “ERCP” within the INPC include both radiology interpretation of fluoroscopic images from ERCP, and the ERCP procedure report itself. These were separated using NLP and validated via manual review. The ERCP procedure reports were used for analysis of quality measures. Only the ERCP procedure reports were used for NLP analysis of quality measures.

Provider Specific Quality Measurement for Endoscopic Retrograde Cholangiopancreatography

We randomly separated the above ERCP procedure reports 1:1 into training (fully available for investigators to create NLP query) and testing (blinded to investigators) sets. The training set of ERCP reports was used to create NLP algorithms in an iterative fashion. The NLP algorithms were used on the test documents (which were not reviewed during the algorithm development phase) to evaluate the ability of the system to accurately determine the expected finding. Independent training and test sets were used to avoid over-fitting the NLP algorithm. Figure 1 shows the flow of the documents and analysis for the study.

All ERCP procedure reports were analyzed as part of Metrics 1-5 as they did not require a specific indication (ie, ICD code). Metrics 6-10 required ICD9-CM identification before NLP analysis using ERCPs performed for choledocholithiasis, defined by ICD9 codes 574.3* - 574.5*; procedures were excluded when one or more of the following codes were associated: (1) spasm of sphincter of Oddi (576.5), (2) acute pancreatitis (577.1), (3) 577.2 (cyst and pseudocyst of pancreas), (4) 577.1 (chronic pancreatitis), (5) 577.8 (other specified disease of pancreas), (6) 751.7 (anomalies of pancreas), and (7) 157.* (malignant neoplasm of pancreas). We selected choledocholithiasis because this represents the most common indication for ERCP and requires selective biliary cannulation and common bile duct stone extraction. Within these procedures, rates of inadvertent pancreatic duct cannulation and injection may also denote technical proficiency in treating choledocholithiasis. NLP was not used for identification of choledocholithiasis due to the complexity of terms and the large variation within the procedure reports when discussing indication.

NLP was used to identify individual endoscopists and all quality measures listed in Table 1. For each quality measure, we randomly selected 50 individual documents from the training set for which the quality measure was identified by NLP and 50 other documents for which the quality measure was not identified (Figure 1). A single expert gastroenterologist (TDI) reviewed all documents manually (different random selection for each quality measure) from the NLP search and assessed true positives (TP, those documents that were appropriately identified by the search) and true negatives (TN, those documents without the presence of the quality measure by NLP). The precision (True positives/50) and accuracy of the NLP search for each quality measure were assessed based on the manually reviewed documents.

We estimated that 50 documents were sufficient for procedure identification given the high prevalence of the measures within the dataset as well as our previous experience with NLP validation for colonoscopy[12, 23]. A formal power calculation was not performed due to inability to know the percentage of the various outcomes a priori. In total, nearly 1,500 ERCP endoscopy reports were reviewed for the validation.

Extraction of other Quality Measures

Metric "1" (appropriateness of indication) was searched according to ICD-9 CM codes based on the multi-society recommendations for appropriate indications[1]. Appropriate codes were identified by rating the indication as highly appropriate or potentially appropriate. These codes were rated if they had more than 100 instances of being used in the first or second position of billing within 7 days of an ERCP procedure. Additional codes were added based on the known appropriate indications despite having fewer than 100 instances of being used. ICD9-CM codes for Metric 1 are listed in Appendix 1 (available online). NLP was not used for Metric 1 as the variety of potential indications was too large to validate in this study. Metric "12" for perforation related to the procedure was determined based on ICD-9 CM codes including perforation of bile duct (576.3) and perforation of intestine (569.83) within 7 days of ERCP procedure. The metric rate was calculated as the number of perforations / the number of total ERCP by provider and in aggregate. Metric "13" for significant post-sphincterotomy bleeding was determined based on ICD-9 code acute post-hemorrhagic anemia (285.1) within 7 days of procedure. The metric rate was no. of bleeding events / no. of total ERCPs by provider and in aggregate.

Two priority measures endorsed in the society guidelines were not included in this analysis: (1) technically successful placement of a bile duct stent and (2) rate of post-ERCP pancreatitis. We chose not to measure stent placement given the diversity of indications for this maneuver, and post-ERCP pancreatitis requires a more deliberate prospective study design to measure accurately.

Analysis

Statistical analysis was performed to quantify the precision and accuracy of NLP for each quality measure. ***Precision (similar to positive predictive value):*** True positives / Test outcome positives over the 50 reviewed documents. ***Accuracy:*** True positives + True negatives / Total population of 100

reviewed documents. Because the entire document set (n = 63,119 ERCP procedures) was not manually annotated for a criterion standard, a true sensitivity (reports in agreement/positive reports by manual review) could not be calculated.

After the 6 providers were identified by NLP, individualized quality metrics were extracted. For each provider, patient characteristics including age, gender, and race were summarized using mean and standard deviations for continuous variables and frequency and proportions for categorical variables. For binary quality metrics, unadjusted rates for each provider were calculated using proportions and compared using the Pearson chi-square test. For the number of ERCPs performed per year, mean and standard deviation of annual procedural volume were calculated over the study period; providers were compared using the ANOVA F-test. All statistical analyses were performed using SAS 9.4 (SAS Institute, Cary, NC).

Results

Of 63,119 documents on 15,581 patients that were indexed as an “ERCP” document, 39,440 were excluded with 100% accuracy, as keyword text search indicated that these documents were radiology reports or other nonprocedure reports. This resulted in 23,679 ERCP procedures on 13,299 patients identified by NLP (Figure 1). Validation of this methodology showed all reviewed documents to be true ERCP procedure reports written by an endoscopist (eg, not a radiology report). Of 23,679 ERCP procedures, 5 were missing patient age, gender, or race and hence were excluded as having incomplete data and not available for adjustment. The remaining 23,674 procedures comprised the final study sample used to evaluate the quality measures across the 6 providers. Each provider averaged more than 280 ERCPs per year with a range of 282.7 to 570.9. The mean age of patients was 52.9 (17.8) years; 59.6% were women and the majority (75.9%) were white (Supplementary Table 1, available online).

Precision and accuracy of NLP in measuring ERCP quality measures

Table 2 shows the precision and accuracy of NLP for each quality indicator. Precision ranged from 84% to100% with intra-procedure measures having lower values. Accuracy ranged from 90% to 100% with

intra-procedure measures having lower accuracy. We excluded seventeen documents post-hoc because the primary provider listed on the procedure note was a trainee (n=9) or rarely performed ERCP (n=8).

Across the 6 providers and after adjusting for age, sex, and race, variability in adherence to the 13 ERCP-specific quality measures evaluated is illustrated in Figure 2. For those metrics with available, society-recommended benchmark rates, all providers met or nearly met (as defined by an overlapping standard error) their targets for documentation of informed consent, appropriate indication, achievement of deep cannulation, complete extraction of common bile duct stones <1 cm, perforation rate, and hemorrhage rate. Benchmark rates for appropriate documentation of the pre-procedure history and physical examination and risk of adverse events were not met by any of the 6 providers.

Exploratory quality measures

Four additional potential quality measures were also evaluated within ERCPs performed for choledocholithiasis. Rates of unintended pancreatic duct cannulation and injection were significantly varied among the 6 providers, ranging from 15.6% to 26.0% and 8.7% to 22.6%, respectively ($P < 0.0001$ for each. The rate of precut sphincterotomy to achieve deep bile duct cannulation was <4% for all providers, ranging from 1.1% to 4.0% ($p < 0.0001$).

DISCUSSION

ERCP is a challenging procedure with high risk for adverse events and technical failure. The risks increase exponentially when the procedure is performed for non-obstructive indications (eg, sphincter of Oddi dysfunction or idiopathic acute pancreatitis). Feedback to colonoscopists on their adenoma detection rate (ADR) may or may not improve provider performance[17, 24]. It is plausible that feedback to ERCP providers on their adherence to national recommendations would also improve the quality of ERCP services provided. The primary objective of this study was to develop a feasible method to track quality metrics in ERCP. An example of a stoplight report card is shown in Figure 3. Using an existing open-source based NLP system, we extracted quality measures over an 8.5-year period and compared them across individual providers and with society guidelines. This work is the first attempt to assess ERCP

quality measures using NLP, and supports the feasibility of applying these techniques to larger datasets across multiple health-care systems.

In this study we demonstrate clinically significant variation in rate of pancreatic cannulation when not the intended target among providers, even among a highly skilled group of endoscopists at a single referral center (>9% variation). Given that one of the high risks for ERCP is development of post-ERCP pancreatitis, this may be a high-impact process quality metric even among high-volume providers of ERCP. This knowledge may guide quality improvement projects to enhance appropriate documentation and identify providers who are not meeting society-endorsed benchmarks. The majority of quality measures show little to no variation in this highly homogeneous population of high-volume providers. Although many of the measures (eg, document pre-procedure H&P) may not reflect quality as these are often done externally to the report, they are contained within the quality tracking measurements for all endoscopic procedures.

Two priority measures included in the society guideline were not included in this study: rate of bile duct stent placement and rate of post-ERCP pancreatitis. We did not include stent placement because the study population used to analyze intraprocedural quality metrics was limited to patients undergoing ERCP for choledocholithiasis, which is the most common indication for ERCP. In addition, post-ERCP pancreatitis cannot be reliably detected in a retrospective study design; given the concern for detection bias, we decided to exclude post-ERCP pancreatitis from this pilot study.

The study has several limitations. First, the sample is restricted to ERCPs performed at a regional referral center. Although this high-volume unit does not reflect the general patient population, adherence to quality measures should apply to all ERCP providers. Furthermore, this study seeks to develop and validate a feasible method for assessing ERCP quality measures, and not to report adherence to benchmarks established by society guidelines.

A second limitation relates to the methodology used to define appropriate indication for ERCP. The spectrum of indications and their appropriateness will require further study and more detailed manual record review; it is possible that lower rates observed in this pilot study could be attributable to the

definition of appropriate indication (see Appendix). Additionally, future studies will need to refine the definition using ICD-10-CM definitions.

A third limitation is that a single endoscopy software (Provation MD; Wolters Kluwer) was used during the study period. This greatly enhances the ability for text mining and natural language processing to accurately detect specific concepts (eg, 100% accuracy for providers). However, our group has shown that this technique can be applied to other institutions and accurately measure variables despite different methods for text document entry (eg, dictation and endoscopy software)[16]. We also made assumptions about ICD-9-CM coding in relation to the procedure. This can be seen with the post-sphincterotomy bleeding rate allowing for 7 days after the procedure for any event associated with a specific ICD-9-CM code (285.1). With this assumption we may pick up non-ERCP related bleeding and/or bleeding not due to sphincterotomy. Additional ICD-9-CM or ICD-10-CM coding such as 578.9 (GI bleeding) and 998.11 (hemorrhage complicating a procedure) might be used in the future to expand the identification of delayed adverse events.

We believe NLP is unlikely to provide a solution for accurate assessment of quality metrics without additional input of other methods such as data mining of discrete elements (eg, billing and procedure codes). However, given the breadth of maneuvers and outcomes specific to ERCP, NLP provides a standardized tool that could be applied across different platforms for procedure documentation. The optimal combination may be to incorporate NLP into a multifaceted approach to data extraction and summation.

Conclusion

Overall, this study demonstrates that NLP, in conjunction with data mining of ICD codes, has the ability to track some quality measures, especially intra-procedure. However, the ability to accurately extract several peri-procedural quality measures (ie, post-ERCP pancreatitis) from the electronic record is unclear and requires further study. Future studies will need to refine this methodology and apply it to a larger number of providers and institutions, with the long-term goal being accurate tracking and reporting of ERCP quality.

Acknowledgements

Dr Timothy Imler had full access to all of the data in the study and takes responsibility for the integrity of the data and accuracy of the data analysis. The authors disclose that Dr Imler has filed for provisional patent (IURTC-14098-01-US-E) for this work under the name Tracking Real-time Assessment of Quality Monitoring in Endoscopy (TRAQ-ME) through the Indiana University Research and Technology Corporation (IURTC). This work was performed at the Regenstrief Institute, Indianapolis, Indiana, and was supported in part by the American Society for Gastrointestinal Endoscopy Covidien Senior Investigator Mentoring Award (Imperiale) and the American Society for Gastrointestinal Endoscopy Career Development Award (Imler).

Ethics

This study was approved by the Institutional Review Board at Indiana University. Dr Imler has filed for provisional patent (IURTC-14098-01-US-E) for this work under the name Tracking Real-time Assessment of Quality Monitoring in Endoscopy (TRAQ-ME) through the Indiana University Research and Technology Corporation (IURTC). No other conflicts of interest are claimed by the remaining authors.

References

1. Adler DG, Lieb Ii JG, Cohen J, et al. Quality indicators for ERCP. Gastrointest Endosc 2014.
2. Rex DK, Schoenfeld PS, Cohen J, et al. Quality indicators for colonoscopy. Gastrointest Endosc 2014.
3. Park WG, Shaheen NJ, Cohen J, et al. Quality indicators for EGD. Gastrointest Endosc 2014.
4. Wani S, Wallace MB, Cohen J, et al. Quality indicators for EUS. Gastrointest Endosc 2014.
5. Corley DA, Jensen CD, Marks AR, et al. Adenoma detection rate and risk of colorectal cancer and death. N Engl J Med 2014;370:1298-306.
6. Solad Y, Wang C, Laine L, et al. Influence of colonoscopy quality measures on patients' colonoscopist selection. Am J Gastroenterol 2015;110:215-9.
7. Hewett DG, Rex DK. Improving colonoscopy quality through health-care payment reform. Am J Gastroenterol 2010;105:1925-33.
8. Committee ASoP, Anderson MA, Fisher L, et al. Complications of ERCP. Gastrointest Endosc 2012;75:467-73.
9. Colton JB, Curran CC. Quality indicators, including complications, of ERCP in a community setting: a prospective study. Gastrointest Endosc 2009;70:457-67.
10. Cote GA, Imler TD, Xu H, et al. Lower provider volume is associated with higher failure rates for endoscopic retrograde cholangiopancreatography. Med Care 2013;51:1040-7.
11. Kapral C, Duller C, Wewalka F, et al. Case volume and outcome of endoscopic retrograde cholangiopancreatography: results of a nationwide Austrian benchmarking project. Endoscopy 2008;40:625-30.
12. Imler TD, Morea J, Kahi C, et al. Natural language processing accurately categorizes findings from colonoscopy and pathology reports. Clin Gastroenterol Hepatol 2013;11:689-94.
13. Mehrotra A, Dellon ES, Schoen RE, et al. Applying a natural language processing tool to electronic health records to assess performance on colonoscopy quality measures. Gastrointest Endosc 2012;75:1233-9.e14.
14. Harkema H, Chapman WW, Saul M, et al. Developing a natural language processing application for measuring the quality of colonoscopy procedures. J Am Med Inform Assoc 2011;18 Suppl 1:i150-6.
15. Gawron AJ, Thompson WK, Keswani RN, et al. Anatomic and Advanced Adenoma Detection Rates as Quality Metrics Determined via Natural Language Processing. Am J Gastroenterol 2014;109:1844-9.
16. Imler TD, Morea J, Kahi C, et al. Multi-Center Colonoscopy Quality Measurement Utilizing Natural Language Processing. Am J Gastroenterol 2015.
17. Kahi CJ, Ballard D, Shah AS, et al. Impact of a quarterly report card on colonoscopy quality measures. Gastrointest Endosc 2013;77:925-31.
18. Regenstrief Institute L. Health Information Exchange. Volume 2014, 2014.
19. Biondich PG, Grannis SJ. The Indiana network for patient care: an integrated clinical information system informed by over thirty years of experience. J Public Health Manag Pract 2004;Suppl:S81-6.

20. McDonald CJ, Overhage JM, Barnes M, et al. The Indiana network for patient care: a working local health information infrastructure. An example of a working infrastructure collaboration that links data from five health systems and hundreds of millions of entries. Health Aff (Millwood) 2005;24:1214-20.
21. Apache.org. UIMA, 2014.
22. Apache.org. License, 2014.
23. Imler TD, Morea J, Kahi C, et al. Multi-center colonoscopy quality measurement utilizing natural language processing. Am J Gastroenterol 2015;110:543-52.
24. Corley DA, Jensen CD, Marks AR. Can we improve adenoma detection rates? A systematic review of intervention studies. Gastrointest Endosc 2011;74:656-65.

Table 1. Quality measures for ERCP and method of extraction studied as well as non-included measures from ASGE/ACG guidelines.

Metric #	Quality Indicator	Grade of Recommendation per Society Guidelines[1]	Society Guideline Priority Measure[1]	Performance Target	Extraction Method	Reason for Exclusion (if applicable)
Preprocedure						
1	Endoscopy is performed for an appropriate indication	1C+	Yes	> 80%	ICD-9 CM	
2	Informed consent is obtained and fully documented^	3	No	> 98%	NLP	
3	Pre-procedure history and directed physical examination are performed and documented	3	No	> 98%	NLP	
4	Risk for adverse events is assessed and documented before sedation is started^	3	No	> 98%	NLP	
5	Volume of ERCPs performed per year by endoscopist	1C	No	> 100	NLP	
Intraprocedure						
6	Deep cannulation of the ducts of interest is documented*	1C	Yes	> 98%	NLP and ICD-9 CM	
7	Common bile duct stones <1 cm in patients with normal bile duct anatomy are extracted successfully and documented*	1C	Yes	≥ 90%	NLP and ICD-9 CM	
8$	Pancreatic cannulation when not an intended target*	n/a	No	n/a	NLP and ICD-9 CM	

9$	Pancreatic injection when not an intended target*	n/a	No	n/a	NLP and ICD-9 CM	
10$	Pancreatic stent placement *if pancreatic duct cannulated*	n/a	No	n/a	NLP	
11$	Precut sphincterotomy for cannulation	n/a	No	n/a	NLP	
Postprocedure						
12	Perforation due to ERCP (within 7 days)	2C	No	≤ 0.2	ICD-9 CM	
13	Rate of clinically significant hemorrhage after ERCP with or without sphincterotomy (within 7 days)	1C	No	< 1	ICD-9 CM	
Not included in this study						
	Frequency with which appropriate antibiotics for ERCP are administered for settings in which they are indicated	2B	No	> 98%		Medications are not documented within the procedure notes and would be identified from pharmacy records for administration.
	Frequency with which ERCP is performed by an endoscopist who is fully trained and credentialed to perform ERCP	3	No	> 98%		All providers studied (n=6) are high-volume fully trained and credentialed physicians who perform ERCP.
	Frequency with which fluoroscopy time and radiation dose are measured and documented	2C	No	> 98%		This would be captured in alternative electronic systems and is not within the ERCP report.

	Frequency with which stent placement for biliary obstruction in patients with normal anatomy whose obstruction is below the bifurcation is successfully achieved and documented	1C	Yes	> 98%		This study is limited to patients with choledocholithiasis.
	Frequency with which a complete ERCP report that details the specific techniques performed, particular accessories used, and all intended outcomes is prepared	3	No	> 98%		All reports are created in this study using a template endoscopy software and include generalized descriptions of the accessories used.
	Frequency with which acute adverse events and hospital transfers are documented	3	No	> 98%		This would be better studied with further clinical data not available within the ERCP report.
	Rate of post-ERCP pancreatitis	1C	Yes	n/a		Rate of post-ERCP pancreatitis requires prospective evaluation for accurate capture. A retrospective assessment, as in this paper, would have underrepresented the actual incidence of post-ERCP pancreatitis due to

						detection bias.

***** For indication of choledocholithiasis (ICD-9 CM 574.3*, 574.4*, 574.5*) with exclusions of sphincter of Oddi dysfunction or pancreatic pathology by ICD-9 CM and NLP.

$ These measures were added by the co-authors as exploratory metrics.

^ Measures 2 and 4 are taken from the guidelines[1] and are differentiated by the directed statement for “informed consent” or synonyms versus “risks and benefits were discussed”.

Table 2. Validation metrics for natural language processing on ERCP quality measures with 100 documents reviewed per measure.

Measure	True Positive %	True Negative %	Testing Set Accuracy %	Testing Set Precision
ERCP Procedure Report	100	100	100	1
Correct Provider (each tested separately)	100	100	100	1
Preprocedure				
Endoscopy is performed for an appropriate indication (QM1)	n/a	n/a	n/a	n/a
Informed consent is obtained and fully documented (QM2)	100	100	100	1
Pre-procedure history and directed physical examination are performed and documented (QM3)	100	100	100	1
Risk for adverse events is assessed and documented before sedation is started (QM4)	100	100	100	1
Volume of ERCPs performed per year by endoscopist (QM5)	n/a	n/a	n/a	n/a
Intraprocedure				
Deep cannulation of the ducts of interest is documented (QM6)*	100	94	97	1
Common bile duct stones <1 cm in patients with normal bile duct anatomy are extracted successfully and documented (QM7)*	98	80	96.4	98
Pancreatic cannulation when not an intended target (QM8)*	96	100	98	96
Pancreatic injection when not an intended target (QM9)*	96	92	94	96
Pancreatic stent placement if pancreatic duct cannulated (QM10)*	92	88	90	92
Precut sphincterotomy for cannulation (QM11)	84	98	91	84
Postprocedure				

Perforation due to ERCP (within 7 days) (QM12)	n/a	n/a	n/a	n/a
Rate of clinically significant hemorrhage after ERCP with or without sphincterotomy (within 7 days) (QM13)	n/a	n/a	n/a	n/a

Precision (similar to positive predictive value): True positives / Test outcome positives over the 50 reviewed documents.

Accuracy: True positives + True negatives / Total population of 100 reviewed documents.

* Testing was done on documents with an indication of choledocholithiasis.

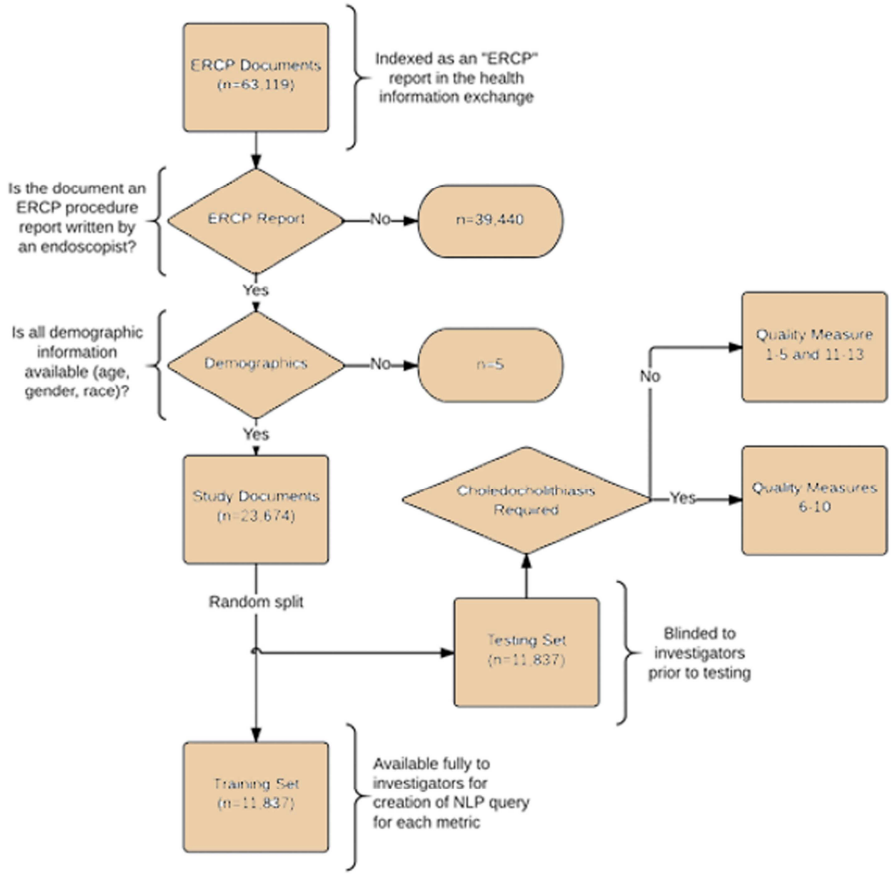
Figure 1. Study flow and methods.

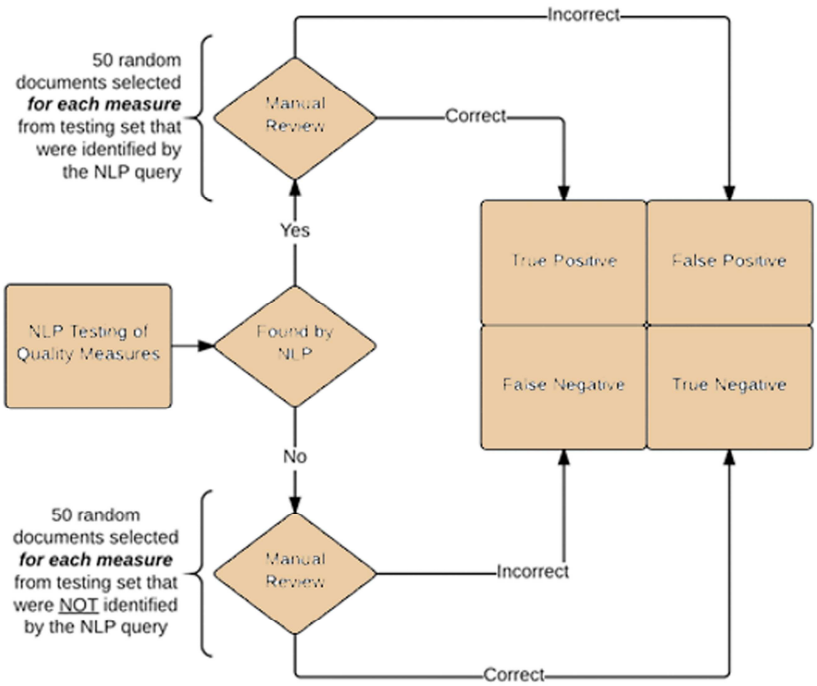
ERCP procedure reports were identified through the INPC and separated into training and testing sets (panel A). Some quality measures were assessed using all ERCP indications (measures 1-5 and 11-13); others required the indication of choledocholithiasis (measures 6-10). For each quality measure, 50 reports with and 50 reports without the measure identified were randomized extracted and reviewed manually (panel B).

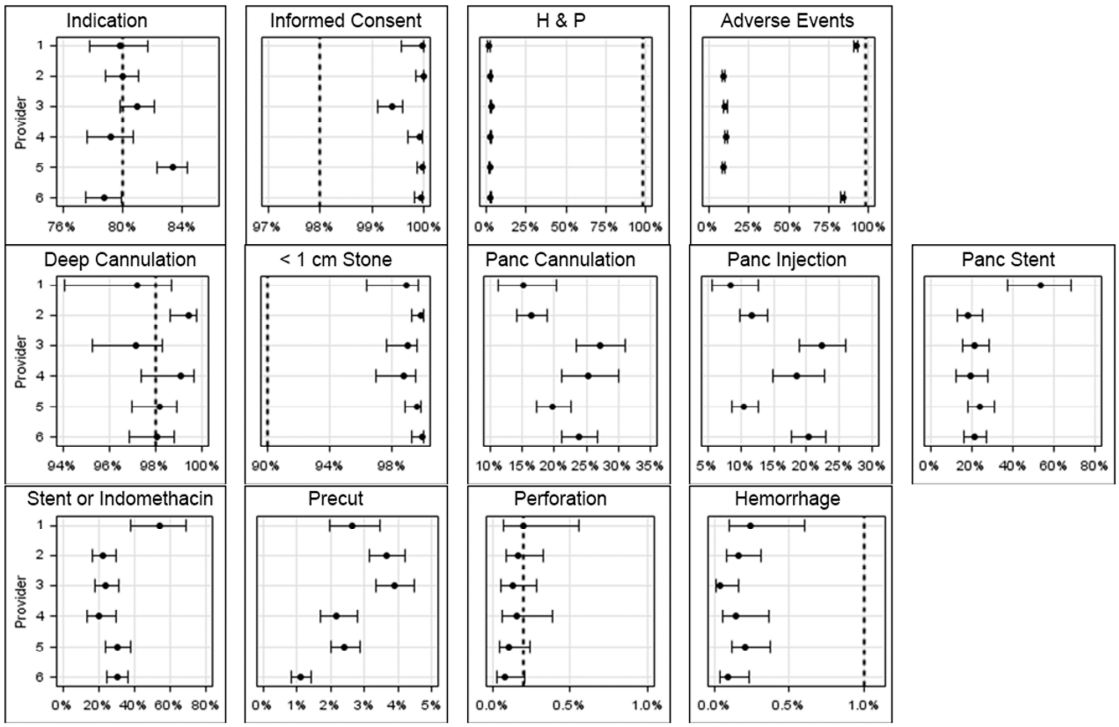
Figure 2. Adjusted provider specific quality measurements for ERCP with 95% confidence intervals. *****Dashed line represents quality measure target.

- Indication = Endoscopy is performed for an appropriate indication
- Informed Consent = Informed consent is obtained and fully documented
- H & P = Pre-procedure history and directed physical examination are performed and documented
- Adverse Events = Risk for adverse events is assessed and documented before sedation is started
- Deep Cannulation = Deep cannulation of the ducts of interest is documented
- < 1 cm Stone = Common bile duct stones <1 cm in patients with normal bile duct anatomy are extracted successfully and documented
- Panc Cannulation = Pancreatic cannulation when not an intended target
- Panc Injection = Pancreatic injection when not an intended target
- Panc Stent = Pancreatic stent placement if pancreatic duct cannulated
- Stent or Indomethacin = Pancreatic stent placement or indomethacin given if pancreatic duct cannulated
- Precut = Precut sphincterotomy for cannulation
- Perforation = Perforation due to ERCP (within 7 days)
- Hemorrhage = Rate of clinically significant hemorrhage after ERCP with or without sphincterotomy (within 7 days)

Figure 3. Example provider quality report card based on Provider 1 overall measure. Green is for being above the benchmark including 95% CI. Yellow includes the benchmark in the 95% CI. Red is below the benchmark including the 95% CI.







	Yearly		Quarterly			
	Current YTD	Previous Year	Qtr FY201x	Qtr FY201x	Qtr FY201x	Qtr FY201x
Pre-Procedure						
Frequency with which endoscopy is performed for an appropriate indication	79.5	x	x	x	x	x
Frequency with which informed consent is obtained and fully documented	100	x	x	x	x	x
Frequency with which pre-procedure history and directed physical examination are performed and documented	1.7	x	x	x	x	x
Frequency with which risk for adverse events is assessed and documented before sedation is started	92.5	x	x	x	x	x
Volume of ERCPs performed	1696	x	x	x	x	x
Intra-Procedure						
Frequency with which deep cannulation of the ducts of interest is documented	97.2	x	x	x	x	x
Frequency with which common bile duct stones <1 cm in patients with normal bile duct anatomy are extracted successfully and documented	99.1	x	x	x	x	x
Frequency of pancreatic cannulation when not an intended target	15.7	x	x	x	x	x
Frequency of pancreatic stent placement if pancreatic duct cannulated	8.8	x	x	x	x	x
Frequency of precut sphincterotomy for cannulation	2.6	x	x	x	x	x
Post-Procedure						
Frequency of perforation due to ERCP (within 7 days)	0.2	x	x	x	x	x
Rate of clinically significant hemorrhage after ERCP with or without sphincterotomy (within 7 days)	0.2	x	x	x	x	x

Appendix

ICD-9 CM Inclusion for Metric 1

Appropriate Indication codes included; calculus of bile duct (574.3*, 574.4*, 574.5*), cholangitis (576.1), obstruction of bile duct (576.2), fistula of bile duct (576.4), spasm of sphincter of Oddi (576.5), other specified disorders of biliary tract (576.8), acute pancreatitis (577.1), 577.2 (cyst and pseudocyst of pancreas), 577.1 (chronic pancreatitis), 577.8 (other specified disease of pancreas), 751.7 (anomalies of pancreas), 157.* (malignant neoplasm of pancreas), 156.* (malignant neoplasm of gallbladder and extrahepatic bile ducts), and 155.* (malignant neoplasm of liver and intrahepatic bile ducts). All codes were linked to an appropriate ERCP related CPT code within the data set. The metric rate was number of appropriately identified ERCP / number of total ERCP by provider and in aggregate.

ICD-9 CM Inclusion for Metric 6-9

Inclusion codes were calculus of bile duct (574.3*, 574.4*, 574.5*) with exclusion codes of spasm of sphincter of Oddi (576.5), acute pancreatitis (577.1), 577.2 (cyst and pseudocyst of pancreas), 577.1 (chronic pancreatitis), 577.8 (other specified disease of pancreas), 751.7 (anomalies of pancreas), and 157.* (malignant neoplasm of pancreas).

Creation of Risk Adjusted Quality Scores

The adjusted proportions for each quality measure were estimated proportions based on the logistic regression model where patients for the six providers were similar in terms of age, gender, and race. Specifically, these proportions were calculated for specified levels of the covariates that were found in the data across all providers. They could be easily calculated using the LSMEANS statement in the SAS LOGISTIC procedure with the OM option.

Appendix Table 1. Breakdown of ERCP procedures within dataset.

	Provider	# of ERCP	Per Year Rate (Std)	# of Patients	Age (Std)	Female	White
	Provider 1	1696	282.7 (113.3)	1060	53.7 (17.6)	991 (58.5%)	1266 (74.6%)
	Provider 2	5133	570.4 (194.5)	3084	51.9 (18.0)	3174 (61.8%)	3946 (76.9%)
	Provider 3	4455	495.3 (192.9)	3103	51.6 (17.0)	2812 (63.1%)	3460 (77.7%)
	Provider 4	2680	297.9 (83.7)	1804	56.3 (17.3)	1497 (55.9%)	1998 (74.6%)
	Provider 5	5138	570.9 (157.9)	2507	52.9 (18.1)	3025 (58.9%)	3920 (76.3%)
	Provider 6	4572	508.0 (137.5)	2935	52.8 (17.9)	2614 (57.3%)	3386 (74.1%)
P Value		< 0.001		Not felt to be clinically significant			