Developing a natural language processing application for measuring the quality of colonoscopy procedures

Henk Harkema,[1] Wendy W Chapman,[2] Melissa Saul,[1] Evan S Dellon,[3] Robert E Schoen,[4] Ateev Mehrotra[5]

► An additional appendix is published online only. To view this file please visit the journal online (www.jamia.org/content/18/Suppl_1.toc).

[1]Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, Pennsylvania, USA
[2]Division of Biomedical Informatics, University of California, San Diego, La Jolla, California, USA
[3]Division of Gastroenterology and Hepatology, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA
[4]Division of Gastroenterology, Hepatology and Nutrition, University of Pittsburgh, Pittsburgh, Pennsylvania, USA
[5]Division of General Internal Medicine, University of Pittsburgh, Pittsburgh, Pennsylvania, USA

Correspondence to
Dr Henk Harkema, Department of Biomedical Informatics, University of Pittsburgh, Parkvale Building M-183, 200 Meyran Avenue, Pittsburgh, PA 15260, USA; heh23@pitt.edu

Received 14 June 2011
Accepted 18 August 2011
Published Online First 21 September 2011

ABSTRACT

Objective The quality of colonoscopy procedures for colorectal cancer screening is often inadequate and varies widely among physicians. Routine measurement of quality is limited by the costs of manual review of free-text patient charts. Our goal was to develop a natural language processing (NLP) application to measure colonoscopy quality.

Materials and methods Using a set of quality measures published by physician specialty societies, we implemented an NLP engine that extracts 21 variables for 19 quality measures from free-text colonoscopy and pathology reports. We evaluated the performance of the NLP engine on a test set of 453 colonoscopy reports and 226 pathology reports, considering accuracy in extracting the values of the target variables from text, and the reliability of the outcomes of the quality measures as computed from the NLP-extracted information.

Results The average accuracy of the NLP engine over all variables was 0.89 (range: 0.62—1.0) and the average F measure over all variables was 0.74 (range: 0.49—0.89). The average agreement score, measured as Cohen's κ, between the manually established and NLP-derived outcomes of the quality measures was 0.62 (range: 0.09—0.86).

Discussion For nine of the 19 colonoscopy quality measures, the agreement score was 0.70 or above, which we consider a sufficient score for the NLP-derived outcomes of these measures to be practically useful for quality measurement.

Conclusion The use of NLP for information extraction from free-text colonoscopy and pathology reports creates opportunities for large scale, routine quality measurement, which can support quality improvement in colonoscopy care.

INTRODUCTION

Several studies have raised serious concerns about the quality of healthcare in the USA, stressing the need for continuous quality improvement.[1–3] One of the barriers to improving the quality of healthcare is the shortage of readily available measures that can be used to provide feedback to physicians.[4] To date, quality of clinical care has generally been measured using administrative claims analysis or manual record reviews.[5] Both methods have key limitations: administrative claims often lack the necessary clinical detail important to providers and medical record review is time-consuming and expensive.[5–11]

The increased use of electronic health records (EHRs) is expected to facilitate systematic, comprehensive approaches to quality measurement.[12–15] In some cases, EHRs have facilitated automatic quality measurement from structured data.[16–19] However, much of the key information in EHRs necessary for quality measurement is stored as unstructured, free text and would still require manual review.[20–22] Natural language processing (NLP) can be an efficient way of automatically extracting and structuring this information, making NLP a potentially pivotal technology for enabling quality measurement from EHR data.

In this paper we describe the design and evaluation of an NLP-based application that measures the quality of colonoscopy procedures for colorectal cancer screening from free-text data in the EHR. Colonoscopy is an ideal target for NLP-based applications, because information for measuring colonoscopy quality is not obtainable from claims data and free-text reporting via dictation is the norm.[23]

BACKGROUND

Colorectal cancer is the third most common cancer occurring in men and women and accounts for almost 10% of all cancer-related deaths in the USA.[24] It is largely preventable with regular screening and colonoscopy has become the most widely used screening modality.[25–27] However, the quality of colonoscopy reporting and performance varies widely among providers[28–31] and poor quality of colonoscopy has been associated with higher incidence of subsequent colorectal cancer.[32 33]

NLP provides a set of computational methods and techniques for automatically extracting and structuring information from free-text documents.[34–36] A wide range of NLP pipelines and components specific to clinical text has been developed since the early 1990s,[37–47] supporting applications such as pharmacovigilance, case finding and patient screening, summarization of narrative patient information, and quality measurement.[48–57]

Representative of the few applications of NLP in quality measurement, Chiang *et al* used the MedLEE NLP pipeline to process narrative discharge notes to assess attainment rates to standards of care for cardiovascular diseases.[55] D'Avolio *et al* designed an NLP application for extracting three quality-related variables from free-text, post-operative pathology reports documenting prostatectomies for the treatment of prostate cancer.[56] Pakhomov *et al* applied NLP methods to free-text clinical notes to establish whether patients had received annual foot examinations as recommended by diabetes clinical practice guidelines.[57]

Existing work regarding NLP in connection with colorectal cancer and colonoscopy is limited in

scope and quantity. Adapting a clinical information retrieval system, D'Avolio *et al* identified pathology reports that are consistent with colorectal cancer from an electronic medical record system.[58] Denny *et al* proposed an application that automatically identifies patients in need of colorectal cancer screening by detecting the timing and status of colorectal screening tests mentioned within a patient's narrative electronic clinical documentation.[59 60] Neither of these systems assesses the quality of the colonoscopy procedure itself.

Most previous work in NLP for quality measurement focuses on the description of methods that extract a limited number of quality-related variables from text, and does not use the NLP output to establish performance on specific quality measures. The NLP system presented in this paper targets a diverse and comprehensive set of variables and evaluates the usefulness of the NLP-extracted information for computing the quality of colonoscopy with regard to various measures published in the literature.

MATERIALS AND METHODS

Definition of task

Based on clinical guidelines and quality improvement targets for colonoscopy procedures published by physician specialty societies,[61 62] three clinicians with expertise in gastroenterology and internal medicine chose a set of 19 quality measures, shown in table 1, which they considered most important for colonoscopy quality. These quality measures were mapped onto 21 variables, shown in table 2, to be extracted from free-text colonoscopy and pathology reports by the NLP engine.

As detailed in Mehrotra *et al*[63] and in the online supplementary appendix to this paper, each quality measure is formulated as a fraction, where the denominator specifies the set of colonoscopy procedures that are eligible for the measure and the numerator specifies the set of colonoscopy procedures that satisfy the measure. For example, the cecal intubation measure is defined as the fraction of colonoscopies where the cecum (most proximal part of the colon) was reached, among all colonoscopies not terminated early and where the bowel preparation was adequate.

The definitions of the quality measures determine the target variables for the NLP pipeline. For the cecal intubation measure, for example, the NLP engine must establish for each report whether the cecum was observed during the procedure, whether the procedure was terminated early, and whether the preparation of the patient's bowel was adequate.

All variables are document-level variables, that is, there is one instance of each variable per report. Each colonoscopy report describes one procedure, and the pathology report linked to the colonoscopy report, if present, describes the pathology results for that procedure. A pathology report is only generated if a biopsy is taken during the colonoscopy.

Data sets

The NLP system was developed and tested on a set of operative reports for colonoscopy procedures and pathology reports retrieved from the Medical ARchival System (MARS) of the University of Pittsburgh Medical Center (UPMC).[64] To meet HIPAA guidelines and ensure patient confidentiality, all data were de-identified using an honest broker system.[65] This study was deemed to meet the criteria for exemption of informed consent by the University of Pittsburgh Institutional Review Board.

The colonoscopy reports were randomly selected from the tranche of documents in MARS documenting outpatient colonoscopies performed at any of 10 hospitals within the UPMC health system between 2007 and 2009. These reports were linked to pathology reports from MARS with matching dates. A set of 97 colonoscopy reports and 23 associated pathology reports were used for two rounds of development of the NLP pipeline; 453 colonoscopy reports and 226 pathology reports were set aside as a blind test set for the final evaluation of the NLP pipeline.

Table 1 Overview of the quality measures tracked by the NLP-based quality tool, their outcomes derived from the manual reference standard and NLP output for the test corpus of 453 colonoscopy reports and 226 pathology reports

	Ref St			NLP				
Quality measure	**E**	**P**	**%**	**E**	**P**	**%**	A_o	κ
Provide (standard) indication for procedure	453	428	94	453	373	82	0.87	0.39
Document that informed consent was obtained	426	275	65	435	222	51	0.85	0.72
Document ASA classification of physical status	426	46	11	435	42	9.7	0.95	0.82
If indication is screening, note previous colonoscopy	284	36	13	230	7	3.0	0.82	0.67
If indication is colon cancer screening and patient has IBD (UC/Crohn's), document previous colonoscopy timing	2	1	50	1	0	0.0	1.0	0.67
Document quality of preparation	453	263	58	453	228	50	0.89	0.77
Rate of procedures with adequate preparation	453	225	50	453	214	47	0.93	0.86
Track cecal intubation rate	411	406	99	436	427	98	0.89	0.24
Document cecal landmarks	411	284	69	436	283	65	0.85	0.71
Rate of detection of any adenomas	406	114	28	427	124	29	0.86	0.72
Rate of detection of large adenomas	406	10	2.5	427	21	4.9	0.86	0.33
Rate of detection of advanced adenomas	406	19	4.7	427	32	7.5	0.85	0.41
Rate of detection of polyps	406	162	40	427	121	28	0.79	0.60
Rate of detection of polyps >9 mm	406	32	7.9	427	39	9.1	0.86	0.50
Document withdrawal time	453	3	0.70	453	4	0.90	1.0	0.86
Withdrawal time ≥6 min	453	3	0.70	453	4	0.90	1.0	0.86
If indication is chronic diarrhea, obtain biopsy	22	17	77	21	15	71	0.98	0.82
Track rate of any complication	453	2	0.40	453	39	8.6	0.92	0.09
If negative study, no family history, and no UC/Crohn's, follow-up time for next procedure recommended should be 10 years	81	19	23	80	11	14	0.91	0.69

ASA, American Society of Anesthesiologists; A_o, κ, observed agreement and κ scores for NLP and reference standard partitions of report set; E, number of reports eligible for quality measure; IBD, inflammatory bowel disease; P, number of reports passing quality measure; Ref St, reference standard; UC, ulcerative colitis; %, quality measure score (P/E).

Table 2 List of target variables with brief descriptions, possible values, and their frequency of occurrence in the manually annotated test set of 453 colonoscopy reports and 226 pathology reports

Variable	Description	Values
Indication type	Indications for procedure	Conditions (up to three) selected from predefined list (551; 13)
Informed consent	Whether informed consent was obtained from patient	*Yes (289); Not mentioned (164)*
Family history	Presence of family history of colorectal cancer	*Positive (57); Negative (23); Not mentioned (373)*
Previous colonoscopy	Time since patient had last colonoscopy	Time in years (42; 13); *No previous colonoscopy (9); Not mentioned (402)*
Nursing reports	Whether physician refers to nursing reports for patient	*Yes (27); No (426)*
ASA	ASA (American Society of Anesthesiologists) classification of patient	*1 (1); 2 (41); 3 (4); 4 (0); 5 (0); Not mentioned (407)*
Preparation	Quality of bowel preparation	*Adequate (225); Not adequate (38); Not mentioned (190)*
Ileo-cecal valve	Whether ileo-cecal valve was observed	*Yes (306); No (6); Not mentioned (141)*
Appendiceal orifice	Whether appendiceal orifice was observed	*Yes (288); No (6); Not mentioned (159)*
Cecum	Whether cecum was reached	*Yes (445); No (6); Not mentioned (2)*
Procedure aborted	Whether procedure was terminated early	*Yes (6); No* (447)*
Polyp removal	Whether one or more polyps were removed	*Yes (175); No* (278)*
Biopsy	Whether biopsy was performed	*Yes (75); No* (378)*
Biopsy location†	Whether location of biopsy is documented	*Yes (57); No (18)*
Largest polyp size	Size of largest polyp found	Size in millimeters (151; 24); *No polyp found (302)*
Withdrawal time	Withdrawal time of scope	Time in minutes (3; 3); *Not mentioned (450)*
Complications	Whether any complications occurred during procedure	*Yes (2); No (312); Not mentioned (139)*
Follow-up interval	Recommended interval for follow-up colonoscopy	Interval in years (226; 22); *Not mentioned (227)*
Adenomatous‡	Whether any of the polyp specimens submitted is adenomatous	*Yes (132); No* (94)*
Largest adenoma size‡ §	Size of largest adenomatous polyp specimen submitted	Size in millimeters (132; 18)
Bad pathology‡	Whether any adenoma has villous component, high-grade dysplasia, or pathology shows invasive cancer	*Yes (16); No* (210)*

*The value *No* includes cases where the information for the variable is not mentioned in the report.
†Variable has no value if no biopsy has been performed.
‡Value of variable extracted from the pathology report (values of all other variables are extracted from the colonoscopy report).
§Variable has no value if none of the polyps submitted were found to be adenomatous.
Key to expressions in the 'Values' column: 'Type of value (Number of reports for which variable has value of given type; Number of distinct values of given type)' or '*Specific value (Number of reports for which variable has given value)*.' For example, for the variable 'Previous colonoscopy,' 42 reports stated the time in years since the patient's last colonoscopy (for a total of 13 different values), 9 reports stated that the patient had not previously had a colonoscopy, and 402 reports did not provide any information about a previous colonoscopy.

The colonoscopy reports were generated in two ways. The majority (349/453 reports in the test set) were dictated by a physician and transcribed. The remainder were produced using the Pentax report generation system. The Pentax reports consist of template-based natural language sentences, organized into sections defined by the system, where the physician typed in free text to fill in the blanks in the templates and to record additional information about the procedure. Because of their mechanical creation, the text of these reports shows less linguistic variation than the dictated reports. The pathology reports were generated using Cerner's COPATH system, which provides a combination of free-text dictation and templates.

The clinicians created a set of guidelines for annotating the colonoscopy and pathology reports with values for the target variables. This required reaching consensus on the definitions of the variables. For example, there are many ways in which the quality of the bowel preparation can be described in a report, including 'well-prepared,' 'sub-optimally cleaned,' 'fairly clean,' and 'somewhat suboptimal,' whereas the corresponding target variable only allows for two values: adequate or inadequate.

The manually annotated reports served as the reference standard for the development and testing of the NLP system. To monitor the reliability of the annotations, 35 colonoscopy reports and 10 pathology reports in the blind test set were annotated in triplicate. We report inter-annotator agreement using the average of the pairwise Cohen's κ scores for the three annotators on these reports.

System architecture

Figure 1 illustrates the general architecture of the NLP-based system for measuring colonoscopy quality. The main component is the NLP engine, which identifies the values of the target variables for each input report. These values are used to establish which reports are eligible and which reports satisfy each quality measure. The fractions of satisfactory reports versus eligible reports determine the outcomes of the quality measures.

The NLP engine has been implemented in GATE.[66] Its design is based on the Topaz architecture, which specifies a rule-based approach to indexing concepts in clinical reports and identifying properties of concepts from their contexts within the text.[67–69] Conceptually, as shown in figure 1, the NLP engine processes the colonoscopy and pathology reports in four steps.

First, in the pre-processing step, the text of a report is split into tokens, sentences, and sections. The structured header of a colonoscopy report is parsed to extract the date of the procedure. The next step uses existing biomedical vocabularies to recognize clinically relevant concepts in a report. Each sentence in the report is submitted to the MetaMap Transfer (MMTx) program (2.4.C release), which maps words and phrases to a subset of concepts in the UMLS Metathesaurus, including concepts for the semantic types Anatomical Structure, Neoplastic Process, and Sign or Symptom.[70–73] Temporal expressions and measurements of size in the text are parsed and interpreted with a set of regular expression patterns.

In the third step, the ConText algorithm is used to identify the clinical and linguistic properties of concepts.[47] Merely recognizing concepts is not sufficient for successful information extraction. For example, processing the sentences, 'He is aware of the risk of bleeding,' 'The polyp was ablated. No bleeding was noticed,' and 'Indications for procedure: rectal bleeding and colonic polyps,' the system must be able to determine that

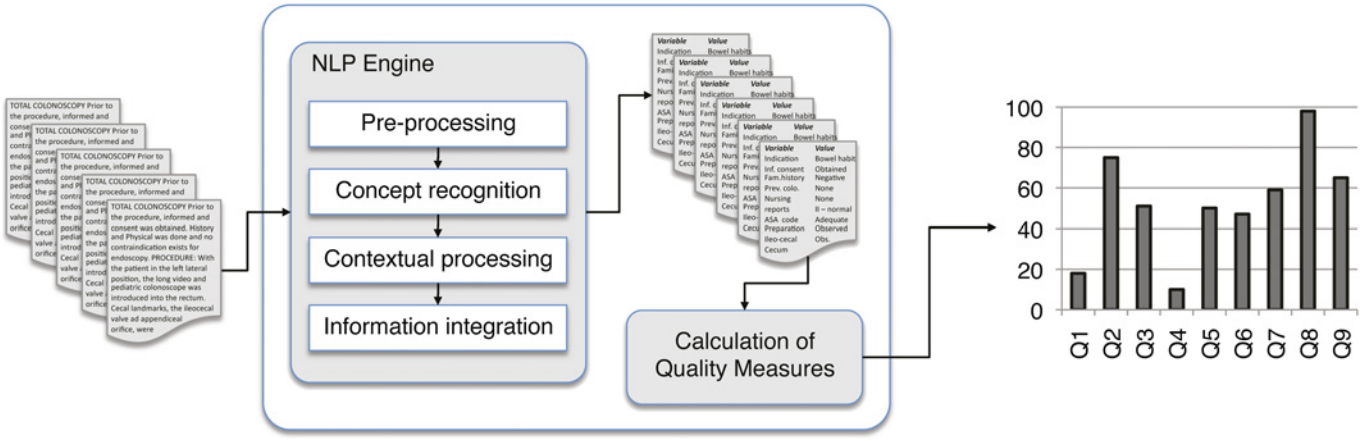

Figure 1 General architecture of the NLP-based system for measuring the quality of colonoscopy procedures from free-text clinical reports. NLP, natural language processing.

'bleeding' is a potential risk in the first sentence, a complication in the second sentence, and an indication in the third sentence. ConText operates on the premise that the presence of a property is signaled by specific keywords in the context of a concept. For example, 'pre-op diagnosis' and 'reason for procedure' signal indications for a colonoscopy. We extended ConText to recognize, among others, bowel preparations that are of adequate quality, anatomic locations where a biopsy was taken, and conditions that are indications for a colonoscopy.

The final processing step is establishing the values of the target variables. The rules used are generally simple, for example, if a concept 'Colon cancer' with contextual properties 'Directionality=Affirmed,' 'Temporality=Historical,' and 'Experiencer=Family member' has been identified within a report, the variable 'Family history' is set to 'Present.' More complex rules involve nested conditionals, considering multiple concepts of different types. The rule for the variable 'Previous colonoscopy' includes resolving and comparing temporal expressions in the text and the header of the report to determine the time elapsed since the last colonoscopy. Some of the rules target concepts found in specific report sections, for example, the 'Gross Description' section in pathology reports, which is the primary location for finding adenoma sizes. Further details about the implementation of the NLP engine are provided in the online supplementary appendix.

Evaluation

The performance of the NLP system was evaluated in two ways. The first evaluation was an intrinsic NLP assessment, where we measured the NLP engine's ability to find the correct values of each of the *variables* for the test set, using the physician annotations as the reference standard. We also compared the NLP output to a 'majority' baseline, which returns the most frequent value in the test set for each variable. The second evaluation is an extrinsic evaluation, in which we used the output of the NLP engine to calculate the outcomes of each of the *quality measures* for the test set. We compared these outcomes to those calculated from the manual annotations. The extrinsic evaluation provides a summary of the usefulness of NLP for the task of quality measurement from textual reports.

The intrinsic NLP performance results are reported using accuracy, recall, precision, and F measure for each variable. Accuracy A is defined as the fraction of the reports in the test set for which the NLP engine assigns the correct value to the variable according to the reference standard. The recall and precision scores are averaged over all values of a variable. The recall for a given variable V and value x is defined as the number of reports in the test set for which the NLP engine assigns x to V in agreement with the reference standard, as a fraction of the total number of reports for which the reference standard assigns x to V. The precision for a given variable V and value x is defined as the number of reports in the test set for which the NLP engine assigns x to V in agreement with the reference standard, as a fraction of the total number of reports in the test set for which the NLP engine assigns x to V. The average recall R_a and precision P_a are calculated as the arithmetic mean of the recall and precision scores over all values of a variable. The average F measure F_a is the harmonic mean of R_a and P_a. (Recall and precision scores for individual variable values can be found in the online supplementary appendix.)

If a variable has a skewed value distribution, that is, some values occur substantially more often than others, a relatively high accuracy score may be easily achieved when the NLP engine does well for just the frequently occurring values. However, discerning infrequent values is important for accurate calculation of the quality measures. This concern is addressed by considering the measures R_a, P_a, and F_a, which evenly weigh the performance for each value of a variable, regardless of its frequency in the report set.

For the extrinsic evaluation, we determined the outcomes of the quality measures for the reports in the test set. We compared the NLP-based outcomes against the reference standard outcomes, which are calculated from the manual annotations. Each quality measure partitions the test set into three classes: reports that are not eligible for the quality measure, reports that are eligible but do not satisfy the quality measure, and reports that satisfy the quality measure. For each quality measure, we report observed agreement and κ scores for the reference standard partition based on the manual annotations versus the partition derived from the NLP output.

RESULTS

Target variables

Table 2 presents the variables that are extracted by the NLP engine from free-text colonoscopy and pathology reports. Some variables relate to pre-procedure interactions with the patient, for example, obtaining informed consent. Other variables describe the procedure itself, for example, whether anatomical

landmarks such as the ileo-cecal valve were observed. Yet other variables relate to post-procedure care, for example, the recommended interval for a follow-up procedure.

Some variables, specifically measurements of size and time, have continuous values, whereas other variables have a fixed number of possible values. The variable 'Indication type' is special in that its value is a set consisting of up to three conditions taken from a pre-defined list of 20 possible conditions, including indications such as 'Screening for colon cancer—No history of polyps' and 'Evaluation of unexplained gastrointestinal bleeding.' As shown in table 2, very few of the variables exhibit an even distribution of values for the reports in the test set.

Completeness of documentation is an important aspect of quality. For this reason, several variables have a distinct value indicating that the necessary information was not mentioned in a report.

All but one variable, 'Bad pathology,' showed substantial to perfect inter-annotator agreement for the triply annotated reports, with average pairwise κ scores between the three annotators ranging from 0.75 to 1.0. For the binary variable 'Bad pathology,' one of the annotators disagreed with the other two on just one report, resulting in two pairwise κ scores of 0.0. Combined with a pairwise score of 1.0 for the two other annotators, the averaged κ score was 0.33. Some pairwise κ scores for the variables 'Cecum,' 'Procedure aborted,' 'Withdrawal time,' and 'Biopsy location' could not be calculated, because both annotators in the pair were in complete agreement and used one value of the variable for all reports, yielding an infinite κ score.

Intrinsic evaluation

Table 3 shows the accuracy A, average recall R_a, average precision P_a, and average F measure F_a for each target variable as extracted by the NLP engine from the test set of colonoscopy and pathology reports. The table also shows the performance of the majority baseline system, which for each variable returns the value that occurs with the greatest frequency in the test set.

As shown in table 3, the accuracy scores for the NLP engine range from 0.62, for the variable 'Complications,' to 1.0, for the variable 'Withdrawal time.' The average accuracy over all variables is 0.89. The average F measures range from 0.49, for 'Procedure aborted,' to 0.98, for 'Adenomatous.' The average F measure over all variables is 0.74.

For a given variable, the F measure is generally lower than the accuracy, indicating that the NLP engine tends to perform better for more frequently occurring values. For example, for the variable 'Cecum,' the NLP engine failed to extract the infrequent values 'No' and 'Not mentioned' with a reasonable degree of recall or precision. Also, the average precision scores are generally higher than the average recall scores; the NLP rules we formulated are precise, but may fail to generalize to text patterns that were not seen in the development set.

As shown in table 3, the NLP engine largely outperforms the majority baseline system. For variables with a very skewed value distribution, the difference in accuracy between the NLP engine and the baseline system is small, but for most of these cases the NLP engine achieves better average F measures. For the variables 'Cecum,' 'Procedure aborted,' and 'Biopsy location,' the results show that the rules in the NLP engine evidently did not encode any meaningful knowledge that would allow it to perform better than the baseline approach. For 'Biopsy location,' the NLP engine assigned 'Yes' to all reports in the test set, imitating the baseline system.

Table 3 Evaluation results for the automatic extraction of the target variables from the test corpus of 453 colonoscopy reports and 226 pathology reports and comparison with the majority baseline results, measured as accuracy A, average recall R_a, average precision P_a, and average F measure F_a

	NLP engine				Baseline	
Variable	**A**	**R_a**	**P_a**	**F_a**	**A**	**F_a**
Indication type	0.74	0.68	0.85	0.76	0.28	0.11
Informed consent	0.87	0.90	0.87	0.88	0.64	0.56
Family history	0.96	0.86	0.90	0.88	0.82	0.47
Previous colonoscopy	0.92	0.45	0.88	0.60	0.89	0.12
Nursing reports	0.97	0.79	0.93	0.86	0.94	0.65
ASA	0.98	0.61	0.98	0.75	0.90	0.38
Preparation	0.87	0.69	0.78	0.73	0.50	0.40
Ileo-cecal valve	0.91	0.62	0.89	0.73	0.68	0.45
Appendiceal orifice	0.89	0.61	0.88	0.72	0.64	0.44
Cecum	0.97	0.44	0.61	0.51	0.98	0.50
Procedure aborted	0.98	0.50	0.49	0.49	0.99	0.66
Polyp removal	0.95	0.94	0.95	0.94	0.61	0.55
Biopsy	0.78	0.86	0.71	0.78	0.83	0.62
Biopsy location*	0.76	0.50	0.76	0.60	0.76	0.60
The largest polyp size	0.86	0.63	0.60	0.61	0.67	0.08
Withdrawal time	1.0	1.0	0.88	0.93	0.99	0.39
Complications	0.62	0.77	0.49	0.60	0.69	0.44
Follow-up interval	0.87	0.44	0.74	0.55	0.50	0.08
Adenomatous†	0.98	0.98	0.99	0.98	0.58	0.54
Largest adenoma size† ‡	0.80	0.76	0.58	0.66	0.21	0.09
Bad pathology†	0.98	0.96	0.91	0.94	0.93	0.65
Average	0.89	0.71	0.79	0.74	0.72	0.42

*Variable only evaluated for the 74 reports for which 'Biopsy=Yes' in both the reference standard and the NLP output.
†Value of variable extracted from pathology report (values of all other variables are extracted from colonoscopy report).
‡Variable only evaluated for the 132 reports for which 'Adenomatous=Yes' in both the reference standard and the NLP output.
For variables in bold, there is a statistically significant difference in accuracy between the NLP engine and the baseline system (McNemar's test, $p<0.01$).
ASA, American Society of Anesthesiologists classification of physical status.

Extrinsic evaluation

Table 1 shows the outcomes of the quality measures for the colonoscopies in the test set based on the NLP output and the manual annotations for these reports, as well as the agreement scores between the NLP-based and manually derived partitions into non-eligible, eligible and non-passing, and passing reports. The κ scores range from 0.86, for the measures 'Rate of procedures with adequate preparation,' 'Document withdrawal time,' and 'Withdrawal time $\geq$6 min,' to 0.09, for the measure 'Track rate of any complication.' The average κ score for all quality measures is 0.62.

Although the observed agreement in table 1 is generally good to very good, only nine of the 19 quality measures reach a κ score larger than 0.70. Clearly, the quality measures with low κ scores rely on variables for which the NLP engine attained a relatively low F_a score.

The κ score for the quality measure 'Track rate of any complication' is particularly low. This is because the variable 'Complications' as extracted by the NLP engine overestimates the occurrence of complications in a situation where the reference standard contains only two reports mentioning a complication. For the quality measure 'Track cecal intubation rate,' there is substantial disagreement between the NLP output and the manual categorization as to how many reports are not eligible (42 vs 17) and how many reports are eligible but do not pass (5 vs 9), leading to a low κ score. Since the majority of reports are, in fact, eligible and pass this quality measure, the manual and NLP-derived outcomes are comparable, despite the low κ score.

DISCUSSION

Findings

We demonstrated that the information required for computing a set of colonoscopy quality measures is amenable to automatic extraction by NLP from free-text colonoscopy and pathology reports. The NLP engine generally performed well on extracting the values of the 21 necessary variables, with an average accuracy of 0.89 and average F measure of 0.74. For some key variable values, in particular infrequently occurring values, the recall or precision of the NLP engine was inadequate. This is reflected in our extrinsic evaluation, where we compared the manual and NLP-based outcomes of the quality measure: κ scores ranged from 0.09 to 0.86 across all measures. Nine of the quality measures achieved a score of 0.70 or above, which we consider a sufficient score for the NLP-derived outcomes for these measures to be practically useful for quality reporting.[63]

Error analysis

Inspection of the output of the NLP engine for the reports in the test set revealed three broad patterns of error. First, the NLP engine did not recognize all mentions of relevant conditions, concepts, and other terms in the reports. For example, it missed the term 'heme-positive stool,' which would have been mapped onto the indication 'Evaluation of unexplained gastrointestinal bleeding.' Furthermore, the adapted version of the ConText algorithm did not include certain atypical trigger terms, such as 'suboptimal' and 'mediocre,' which, in the right context, signal inadequate bowel preparations.

Second, the set of regular expressions used by our version of ConText to detect the contextual properties of concepts was incomplete. The low precision score for complications was primarily due to the NLP engine's difficulties establishing whether a mention of the concept 'Bleeding' in the text of a report referred to a potential risk, an indication, or a complication of the colonoscopy procedure. ConText's rules restrict the context of a concept to the sentence it appears in, whereas descriptions of potential risks and indications may span several sentences or an entire section, potentially putting the trigger term outside the sentence containing the concept it modifies. Also, the negation rules taken from the original version of ConText were developed for detecting the absence of patient findings and symptoms. These rules did not transfer well to detecting cases where the cecum and cecal landmarks were not reached or not observed.

Third, finding the values for some of the variables required inferencing that went beyond the rules that were implemented in the information integration component of the NLP pipeline. For example, one of the reports stated that the scope could only be advanced as far as the hepatic flexure, from which one can infer, using knowledge about the anatomy of the colon, that the cecum was not reached and that the ileo-cecal valve and appendiceal orifice were not observed. Similarly, if the presence of a polyp is mentioned as part of the patient's history, then, in most cases, the indication for the procedure is 'Screening/surveillance among those with a history of polyps,' even when screening or surveillance is not explicitly listed as an indication in the report.

Further development of the NLP engine will address these issues by expanding the rule base and introducing statistical approaches for some of the target variables.

Limitations

This study has several key limitations. The NLP engine was developed and evaluated using reports from 10 different hospitals within a single health system. Therefore, further evaluation is necessary to assess the generalizability to reports from other institutions. The presence of a small number of template-based reports in our data set may have limited linguistic variation.

Some of the quality measures in the published guidelines were reformulated or eliminated because they rely on information from sources that were not available in the context of this project, such as progress notes originating prior to the colonoscopy. Other measures were dropped because they would require co-reference resolution of multiple mentions of polyps and biopsies within the text of a report, which we did not address in this study. Certain indications and complications, for example, perforation of the colon, were extremely rare or non-existent in our report set. Without representative examples, it is difficult to develop NLP rules for the extraction of these cases.

CONCLUSION

The results reported in this paper raise the possibility that our NLP pipeline, with further refinement and development, can be used for routine quality measurement on a substantially larger scale. Because the method is automated, large numbers of reports can be quickly processed, enabling quality measurement at both the level of individual physicians as well as groups of providers and hospitals. In a separate study, we have applied the NLP tool to a set of over 25 000 colonoscopy reports and associated pathology reports from 10 hospitals in the UPMC health system, showing a wide range of variation in quality across hospitals and physicians.[63] If such an analysis could be done regularly across a large number of systems, our NLP pipeline would be useful for physicians and hospitals in suggesting a focus for quality improvement efforts, as well as for patients making decisions as to where to obtain care.

Acknowledgments The authors would like to thank Drs F Bishehsari, E S Dellon, and A Mehrotra for annotating the clinical reports used in this study.

Funding This project was supported through a pilot grant from the RAND-University of Pittsburgh Health Institute (RUPHI), a formal collaboration between the RAND Corporation, RAND Health, and the University of Pittsburgh Schools of the Health Sciences. Further support was provided by grant UL1 RR024153 from the National Center for Research Resources (NCRR), a component of the National Institutes of Health (NIH), and NIH Roadmap for Medical Research, grant R01 LM009427-01 from the NIH, and grant KL2 RR024154 from the NCCR.

Competing interests None.

Ethics approval The University of Pittsburgh Institutional Review Board approved this study.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1. **McGlynn EA,** Asch SM, Adams J, *et al*. The quality of health care delivered to adults in the United States. *N Engl J Med* 2003;**348**:2635—45.
2. **Chassin MR,** Galvin RW. The urgent need to improve health care quality. Institute of Medicine National Roundtable on Health Care Quality. *JAMA* 1998;**280**:1000—5.
3. *Crossing the Quality Chasm: A New Health System for the 21st Century*. Washington, DC: Institute of Medicine of The National Academies, 2001.
4. *Performance Measurement: Accelerating Improvement*. Washington, DC: Institute of Medicine of The National Academies, 2005.
5. **Diamond CC,** Rask KJ, Kohler SA. Use of paper medical records versus administrative data for measuring and improving health care quality: are we still searching for a gold standard? *Dis Manag* 2001;**4**:121—30.
6. **Iezzoni LI.** Assessing quality using administrative data. *Ann Intern Med* 1997;**127**:666—47.
7. **Zhan C,** Miller MR. Administrative data based patient safety research: a critical review. *Qual Saf Health Care* 2003;**12**(Suppl 2):ii58—63.
8. **Pawlson LG,** Scholle SH, Powers A. Comparison of administrative-only versus administrative plus chart review data for reporting HEDIS Hybrid Measures. *Am J Manag Care* 2007;**13**:553—8.
9. **Lieberman DA.** Pitfalls of using administrative data for research. *Dig Dis Sci* 2010;**55**:1721—5.
10. **Tierney WM,** Overhage JM, McDonald CJ. Toward electronic medical records that improve care. *Ann Intern Med* 1995;**122**:725—6.
11. **Eddy DM.** Performance measurement: problems and solutions. *Health Aff (Millwood)* 1998;**17**:7—25.

12. **Bates DW,** Gawande AA. Improving Safety with Information Technology. *N Engl J Med* 2003;**348**:2526—34.
13. *Patient Safety: Achieving a New Standard for Care*. Washington, DC: Institute of Medicine of The National Academies, 2004.
14. **Chaudhry B,** Wang J, Wu S, *et al*. Systematic review: impact of health information technology on quality, efficiency, and costs of medical care. *Ann Intern Med* 2006;**144**:742—52.
15. **Vogt TM,** Ickin M, Ahmed F, *et al*. The prevention index: using technology to improve quality assessment. *Health Serv Res* 2004;**39**:511—30.
16. **Persell SD,** Kho AN, Thompson JA, *et al*. Improving hypertension quality measurement using electronic health records. *Med Care* 2009;**47**:388—94.
17. **Tang PC,** Ralston M, Arrigotti MF, *et al*. Comparison of methodologies for calculating quality measures based on administrative data versus clinical data from an electronic health record system: implications for performance measures. *J Am Med Inform Assoc* 2007;**14**:10—15.
18. **O'Toole MF,** Kmetik KS, Bossley H, *et al*. Electronic health record systems: the vehicle for implementing performance measures. *Am Heart Hosp J* 2005;**3**:88—93.
19. **Baron JR.** Quality improvement with an electronic health record: achievable, but not automatic. *Ann Intern Med* 2007;**147**:549—662.
20. **Linder JA,** Kaleba EO, Kmetik KS. Using electronic health records to measure physician performance for acute conditions in primary care: empirical evaluation of the community-acquired pneumonia clinical quality measure set. *Med Care* 2009;**47**:208—16.
21. **Maddocks H,** Marshall JN, Stewart M, *et al*. Quality of congestive heart failure care: assessing measurement of care using electronic medical records. *Can Fam Physician* 2010;**56**:e432—7.
22. **Baker DW,** Persell SD, Thompson JA, *et al*. Automated review of electronic health records to assess quality of care for outpatients with heart failure. *Ann Intern Med* 2007;**146**:270—7.
23. **Waye JD,** Rex DK, Williams CB. *Colonoscopy: Principles and Practice*. Malden, MA: John Wiley and Sons, 2009.
24. **Jemal A,** Siegel R, Xu J, *et al*. Cancer statistics, 2010. *CA Cancer J Clin* 2010;**60**:277—300.
25. **Winawer SJ,** Zauber AG, Ho MN, *et al*. Prevention of colorectal cancer by colonoscopic polypectomy. The National Polyp Study Workgroup. *N Engl J Med* 1993;**329**:1977—81.
26. **Levin B,** Lieberman DA, McFarland B, *et al*. Screening and surveillance for the early detection of colorectal cancer and adenomatous polyps, 2008: a joint guideline from the American Cancer Society, the US Multi-Society Task Force on Colorectal Cancer, and the American College of Radiology. *Gastroenterology* 2008;**134**:1570—95.
27. **Phillips KA,** Liang SY, Ladabaum U, *et al*. Trends in colonoscopy for colorectal cancer screening. *Med Care* 2007;**45**:160—7.
28. **Kaminski MF,** Regula J. Colorectal cancer screening by colonoscopy—current issues. *Digestion* 2007;**76**:20—5.
29. **Chen S,** Rex DK. Endoscopist is comparable to age and gender as predictor of adenomas at colonoscopy. *Am J Gastroenterol* 2005;**100**:S393.
30. **Lieberman DA,** Faigel DO, Logan JR, *et al*. Assessment of the quality of colonoscopy reports: results from a multicenter consortium. *Gastrointest Endosc* 2009;**69**:645—53.
31. **Rex DK,** Bond JH, Winawer S, *et al*. Quality in the technical performance of colonoscopy and the continuous quality improvement process for colonoscopy: recommendations of the U.S. Multi-Society Task Force on Colorectal Cancer. *Am J Gastroenterol* 2002;**97**:1296—308.
32. **Kaminski MF,** Regula J, Kraszewska E, *et al*. Quality indicators for colonoscopy and the risk of interval cancer. *N Engl J Med* 2010;**362**:1795—803.
33. **Barclay RL,** Vicari JJ, Doughty AS, *et al*. Colonoscopic withdrawal times and adenoma detection during screening colonoscopy. *N Engl J Med* 2006;**355**:2533—41.
34. **Jurafsky D,** Martin JH. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. 2nd edn. Upper Saddle River, NJ: Pearson-Prentice Hall, 2009.
35. **Indurkhya N,** Damerau FJ. *Handbook of Natural Language Processing*. 2nd edn. Boca Raton, FL: CRC Press, 2010.
36. **Allen J.** *Natural Language Understanding*. 2nd edn. Redwood City, CA: Benjamin/Cummings, 1995.
37. **Friedman C.** *A Broad-Coverage Natural Language Processing System*. Proc AMIA Symp 2000:270—4.
38. **Hazlehurst B,** Frost HR, Sittig DF, *et al*. MediClass: A system for detecting and classifying encounter-based clinical events in any electronic medical record. *J Am Med Inform Assoc* 2005;**12**:517—29.
39. **Christensen LM,** Harkema H, Irwin J, *et al*. ONYX: a system for the semantic analysis of clinical text. *Proc BioNLP 2009*. Stroudsburg, PA: Association for Computational Linguistics, 2009:19—27.
40. **Savova GK,** Masanz JJ, Ogren PV, *et al*. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *J Am Med Inform Assoc* 2010;**17**:507—13.
41. **Crowley RS,** Castine M, Mitchell K, *et al*. caTIES: a grid based system for coding and retrieval of surgical pathology reports and tissue specimens in support of translational research. *J Am Med Inform Assoc* 2010;**17**:253—64.
42. **Zou Q,** Chu WW, Morioka C, *et al*. IndexFinder: a method of extracting key concepts from clinical texts for indexing. *AMIA Annu Symp Proc* 2003:763—7.
43. **Li D,** Savova GK, Schuler KK. Conditional Random Fields and Support Vector Machines for Disorder Named Entity Recognition in Clinical Texts. *Proc BioNLP 2008*. Stroudsburg, PA: Association for Computational Linguistics, 2008:94—5.
44. **Denny JC,** Spickard A, Johnson KB, *et al*. Evaluation of a method to identify and categorize section headers in clinical documents. *J Am Med Inform Assoc* 2009;**16**:806—15.
45. **Roberts A,** Gaizauskas R, Hepple M. Extracting clinical relationships from patient narratives. *Proc BioNLP 2008*. Stroudsburg, PA: Association for Computational Linguistics, 2008:10—18.
46. **Zhou L,** Melton GB, Parsons S, *et al*. A temporal constraint structure for extracting temporal information from clinical narrative. *J Biomed Inform* 2006;**39**:424—39.
47. **Harkema H,** Dowling JW, Thornblade T, *et al*. ConText: an algorithm for determining negation, experiencer, and temporal status from clinical reports. *J Biomed Inform* 2009;**42**:839—51.
48. **Wang X,** Hripcsak G, Markatou M, *et al*. Active computerized pharmacovigilance using natural language processing, statistics, and electronic health records: a feasibility study. *J Am Med Inform Assoc* 2009;**16**:328—37.
49. **Aramaki E,** Miura Y, Tonoike M, *et al*. Extraction of adverse drug effects from clinical records. *Stud Health Technol Inform* 2010;**160**:739—43.
50. **Pakhomov S,** Weston SA, Jacobsen SJ, *et al*. electronic medical records for clinical research: application to the identification of heart failure. *Am J Manag Care* 2007;**13** (6 Pt 1):281—8.
51. **Al-Haddad MA,** Friedlin J, Kesterson J, *et al*. Natural language processing for the development of a clinical registry: a validation study in intraductal papillary mucinous neoplasms. *HPB (Oxford)* 2010;**12**:688—95.
52. **Li L,** Chase HS, Patel CO, *et al*. Comparing ICD9-encoded diagnoses and NLP-processed discharge summaries for clinical trials pre-screening: a case study. *AMIA Anuu Symp Proc* 2008:404—8.
53. **Liu H,** Friedman C. CliniViewer: a tool for viewing electronic medical records based on natural language processing and XML. *Stud Health Technol Inform* 2004;**107**:639—43.
54. **Van Vleck TT,** Elhadad N. Corpus-based problem selection for EHR note summarization. *AMIA Annu Symp Proc* 2010:817—21.
55. **Chiang JH,** Lin JW, Yang CW. Automated evaluation of electronic discharge notes to assess quality of care for cardiovascular diseases using Medical Language Extraction and Encoding System (MedLEE). *J Am Med Inform Assoc* 2010;**17**:245—52.
56. **D'Avolio LW,** Litwin MS, Rogers SO Jr, *et al*. Facilitating clinical outcomes assessment through the automated identification of quality measures for prostate cancer surgery. *J Am Med Inform Assoc* 2008;**15**:341—8.
57. **Pakhomov S,** Bjornsen S, Hanson P, *et al*. Quality performance measurement using the text of electronic medical records. *Med Decis Making* 2008;**28**:462—70.
58. **D'Avolio LW,** Nguyen TM, Farwell WR, *et al*. Evaluation of a generalizable approach to clinical information retrieval using the Automated Retrieval Console (ARC). *J Am Med Inform Assoc* 2010;**17**:375—83.
59. **Denny JC,** Choma NN, Peterson JF, *et al*. Natural Language Processing Improves Identification of Colorectal Cancer Testing in the Electronic Medical Record. *Med Decis Making*. Published Online First: 10 March 2011. doi:10.1177/0272989X11400418.
60. **Denny JC,** Peterson JF, Choma NN, *et al*. Extracting timing and status descriptors for colonoscopy testing from electronic medical records. *J Am Med Inform Assoc* 2010;**17**:383—8.
61. **Rex DK,** Petrini JL, Baron TH, *et al*. Quality indicators for colonoscopy. *Gastrointest Endosc* 2006;**63**(4 Suppl):S16—28.
62. **Lieberman D,** Nadel M, Smith RA, *et al*. Standardized colonoscopy reporting and data system: report of the quality assurance task group of the national colorectal cancer Roundtable. *Gastrointest Endosc* 2007;**65**:757—66.
63. **Mehrotra A,** Dellon ES, Schoen R, *et al*. Applying a natural language processing tool to electronic health records to measure the quality of colonoscopy procedures. (under review).
64. **Yount RJ,** Vries JK, Councill CD. The Medical Archival System: An Information Retrieval System Based on Distributed Parallel Processing. *Inf Proc Manag* 1991;**27**:379—89.
65. **Dhir R,** Patel AA, Winters S, *et al*. A Multidisciplinary Approach to Honest Broker Services for Tissue Banks and Clinical Data: A Pragmatic and Practical Model. *Cancer*. 2008;**113**(7):1705—15.
66. **Cunningham H,** Maynard D, Bontcheva K, *et al*. GATE: a framework and graphical development environment for robust NLP tools and applications. *Proc 40th Anniversary Meeting of the Association for Computational Linguistics (ACL)*. Philadelphia, PA: 2002:168—75.
67. **Chapman W,** Conway M, Dowling J, *et al*. Challenges in adapting an NLP system for real-time surveillance. *9th Annual Conference of the International Society for Disease Surveillance*. Park City, UT, 2010.
68. **Chapman W,** Harkema H. Identifying respiratory-related clinical conditions from ED reports with Topaz. *Clin Med Res* 2010;**8**:53.
69. **Chu D.** *Clinical Feature Extraction from Emergency Department Reports for Biosurveillance [Master's Thesis]*. Pittsburgh: University of Pittsburgh, 2007.
70. **Lindberg DA,** Humphreys BL, McCray AT. The unified medical language system. *Methods Inf Med* 1993;**32**:281—91.
71. *Unified Medical Language System (UMLS) Documentation*. Bethesda, MD: National Library of Medicine. http://www.nlm.nih.gov/research/umls/documentation.html.
72. **Aronson AR.** Effective mapping of biomedical text to the UMLS Metathesaurus: The MetaMap program. *Proc AMIA Symp* 2001:17—21.
73. *MetaMap Transfer (MMTx)*. Bethesda, MD: National Library of Medicine. http://mmtx.nlm.nih.gov/MMTx/.