Published in final edited form as:

Clin Gastroenterol Hepatol. 2013 June ; 11(6): 689–694. doi:10.1016/j.cgh.2012.11.035.

NATURAL LANGUAGE PROCESSING ACCURATELY CATEGORIZES FINDINGS FROM COLONOSCOPY AND PATHOLOGY REPORTS

Timothy D. Imler, M.D.^{1,2}, Justin Morea, D.O.², Charles Kahi, M.D. M.S. F.A.C.G¹, and
Thomas F. Imperiale, M.D.^{1,2,3}

¹Division of Gastroenterology and Hepatology, Department of Medicine, Indiana University
School of Medicine

²Regenstrief Institute, Inc.

³Center of Excellence for Implementation of Evidence-based Practice, Richard L. Roudebush VA
Medical Center (3), Indianapolis, IN.

Abstract

Background & Aims—Little is known about the ability of natural language processing (NLP) to extract meaningful information from free text gastroenterology reports for secondary use.

Methods—We randomly selected 500 linked colonoscopy and pathology reports from 10,798 non-surveillance colonoscopies to train and test the NLP system. Using annotation by gastroenterologists as the reference standard, we assessed the accuracy of an open-source NLP engine that processed and extracted clinically relevant concepts. The primary outcome was the

© 2013 The American Gastroenterological Association. Published by Elsevier Inc. All rights reserved.

Correspondence: Timothy D. Imler, M.D. Division of Gastroenterology; Indiana University School of Medicine 1050 Wishard Blvd., Suite 4100 Indianapolis, IN 46202 timler@iupui.edu.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Conflicts of Interest: None

Contributions: Study design (Imler/Morea/Imperiale)

Data collection (Imler, Morea/Imperiale/Kahi)

Data analysis (Imler/Morea)

Manuscript drafting (Imler)

Critical editing (Imler/Morea/Imperiale/Kahi)

No conflicts of interest are reported from the authors on the content of this study.

Ethics This study was approved by the Institutional Review Board at Indiana University and the VA R&D Department.

Definitions

$$\begin{aligned} recall &= \frac{\text{reports in agreement}}{\text{positive reports by gold standard}} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}} \\ precision &= \frac{\text{reports in agreement}}{\text{positive reports by NLP}} = \frac{\text{true positives}}{\text{true positives} + \text{false positives}} \\ accuracy &= \frac{\text{true positives} + \text{false positives} + \text{true negatives} + \text{false negatives}}{\text{true positives} + \text{false positives} + \text{true negatives} + \text{false negatives}} \\ f - measure &= 2 \times \frac{precision \times recall}{precision + recall} \end{aligned}$$

highest level of pathology. Secondary outcomes were: location of the most advanced lesion, largest size of an adenoma removed, and number of adenomas removed.

Results—The NLP system identified the highest level of pathology with 98% accuracy, compared with triplicate annotation by gastroenterologists (the standard). Accuracy values for location, size, and number were 97%, 96%, and 84%, respectively.

Conclusions—The NLP can extract specific meaningful concepts with 98% accuracy. It might be developed as a method to further quantify specific quality metrics.

Keywords

adenoma detection rate; colon cancer screening; software; computerized; natural language procession; colonoscopy; gastroenterology; adenoma detection rate; medical informatics

Background

By 2002 more than 14.2 million colonoscopies were being performed annually in the United States with significant room for expansion¹. Abundant recent evidence has shown the protective effect of lower endoscopy for decreased mortality from colorectal cancer²⁻⁴ and colonoscopy has become the preferred test for colorectal cancer screening in the United States with multiple guideline recommendations supporting its use⁵⁻⁸.

While generally seen as a safe and effective manner for detecting and removing premalignant lesions, studies have been limited by the manner of non-specific data extraction. Many of these studies are derived from exhaustive manual review^{9, 10} or from billing codes such as International Classification of Diseases, 9th edition (ICD-9) and Current Procedural Terminology (CPT)^{11, 12}. These standards lack the ability to accurately portray the complex and detailed clinical information (concepts) being investigated and are beginning to be replaced with newer more robust terminology standards of classification^{13, 14}.

The pathologic record that is generated from a colonoscopy is generally disconnected from the procedure reports and does not allow for easy extraction of complex definitions, such as advanced adenoma. Linked concepts require the expert review of both the colonoscopy and pathology report. Manual review that is necessary to generate a large enough sample for research or quality tracking is costly and fraught with potential for annotator bias and fatigue^{15, 16}.

Medical informatics is the field that studies effective uses of biomedical data, information, and knowledge for scientific inquiry¹⁷. Natural language processing (NLP) is a tool of medical informatics and uses computer based linguistics techniques, using artificial intelligence, to extract meaningful information from free text reports and has been in development since the 1950s^{18, 19}. Until recently the ability to use NLP within the medical field and gastroenterology was limited by linguistic variations and the terminology standards to accurately classify a detailed clinical event¹⁹.

NLP use within medical free-text is in the form of recognized clinical concepts matched towards standard terminologies (e.g. Systematized nomenclature of medicine - Clinical terms (SNOMED-CT)¹³ and Unified Medical Language System (UMLS)¹⁴). These concepts have been encoded into NLP systems and allow for unique concept identification (CUI) and secondary usage of this data²⁰ for information and knowledge processing. An example of this would be that under ICD-9 the most detailed level for a tubulovillous adenoma of the rectum would be code 211.3, “benign neoplasm of colon”. However, through SNOMED-CT we can capture through code 448428002 to get the defined concept of “tubulovillous adenoma of rectum”.

With the ability to capture detailed clinical concepts and process them on the large scale for quality tracking and research, we sought to determine the accuracy of natural language processing within colonoscopy records for defining the most advanced pathology, location of the most advanced pathology, size of the largest adenoma, and number of adenomas.

Materials and Methods

After IRB approval, data was extracted from the Richard L. Roudebush Veterans Administration Medical Center in Indianapolis, Indiana. The data for this study was obtained from an ongoing study of first-time colonoscopy. All data extracted were in electronic format. Inclusion and exclusion criteria were previously selected and the test characteristics were not explicitly outlined for this study and were applied using custom fully electronic software as part of the parent study design.

Inclusion criteria for the parent study cohort included all veterans aged 40 years and over who had an index outpatient colonoscopy between 2002 and 2009 for any indication.

Exclusion criteria for the cohort included: 1) previous VHA-based colonoscopy, 2) indication for colonoscopy of surveillance for neoplasia, 3) surgical resection of any part of the large intestine, 4) history of polyps or cancer of the colon or rectum, 5) history of Crohn’s disease, ulcerative colitis, or inflammatory bowel disease, 6) history of a hereditary polyposis syndrome, 7) history of hereditary non-polyposis colon cancer syndrome, or 8) having a colonoscopy earlier within the 8 year sampling frame (e.g. a person with two colonoscopies within the sampling frame would have the second excluded).

The extracted reports were linked as part of the parent study using study specific software to their corresponding pathology reports and de-identified for NLP analysis. There were 10,798 reports with 6,379 linked to pathology. 500 of the reports with linked pathology were randomly selected using MySQL random record selection for triplicate manual annotation. Figure 1 shows the flowchart for the study.

The 500 reports and their linked pathology were combined into a single text document. Each document had a unique document ID and four sections: findings, impression, specimen, and pathology.

To analyze the reports clinical text analysis and knowledge extraction system (cTAKES), an open-source freely available Java based NLP engine developed originally by the Mayo

Clinic²⁰, was used. It was written on top of the Unstructured Information Management Architecture (UIMA) Framework²¹. cTAKES was designed to process clinical notes and uses named entity recognition to map identified concepts to the UMLS Metathesaurus^{22, 23}, a large, multi-purpose, and multi-lingual thesaurus that contains millions of biomedical and health related concepts, their synonymous names, and their relationships²⁴. This entity recognition allows for capture of key clinical concepts to a standardized format for secondary usage (e.g. UMLS CUI). cTAKES also determines if the named entity is negated as well as the context of the named entity within the document to allow for more robustness than simple text mining strategies.

UMLS concepts were then matched against relevant terms such as ‘tubulovillous adenoma’ for concept unique identifier (CUI) C0334307. If the CUI was negated, it was excluded. These terms were searched and written to an NLP key. Sequential searching was used to identify the highest level of pathology from carcinoma to normal mucosa and included size identification in order to further classify the concept.

Definitions were agreed to a priori for the terms for each concept. The highest level of pathology was divided into five categories: 1) carcinoma, 2) advanced adenoma, 3) non-advanced adenoma, 4) hyperplastic polyp, and 5) normal mucosa / non-significant. Sessile serrated polyps (SSP) were included within the adenoma category. Advanced adenomas were defined as those with villous features, carcinoma in situ, high-grade dysplasia, or adenoma with size on colonoscopy report ≥ 10 mm with size determined by the endoscopist. Non-significant findings included: lipomas, benign colonic tissue, or no specimen for pathologic review.

The location of the most advanced lesion was defined as: proximal (cecum to and including splenic flexure), distal (descending colon to and including the rectum), and both proximal and distal equivalent for most advanced lesion.

The largest adenoma removed as well as the total number of adenomas removed were captured. Presence and absence of hemorrhoids and diverticulosis were captured.

Three of the authors (Imler/Imperiale/Kahi) independently manually reviewed all 500 randomized documents and their associated pathology reports to determine the: 1) highest level of pathology, 2) number of adenomas removed, 3) size of adenoma removed, 4) location of adenoma, and 5) presence of diverticulosis and hemorrhoids.

Recall, precision, accuracy, and f-measure were calculated for both the testing and training data sets. Recall was defined as: $[\text{true positives} / (\text{true positives} + \text{false negatives})]$ or (reports in agreement/positive reports by gold standard). Precision was defined as: $[\text{true positives} / (\text{true positives} + \text{false positives})]$ or (reports in agreement/positive reports by NLP). Accuracy was defined as $[(\text{true positives} + \text{true negatives}) / (\text{true positives} + \text{false positives} + \text{true negatives} + \text{false negatives})]$. The f-measure is defined as $[2 * (\text{precision} * \text{recall}) / (\text{precision} + \text{recall})]$ and is used for the measurement of information retrieval and measures the effectiveness of retrieval²⁵. The values for recall, precision, accuracy, and f-measure vary between 0-1 with 1 being the optimal.

150 of the annotated sample were used for NLP training with blinding to the other annotated documents. All 500 annotated gold standard documents were used for testing. Annotated linked colonoscopy and pathology reports were then run through the NLP engine to capture the highest level of pathology and previously defined concepts.

Results

Initial annotation revealed an initial triplicate annotator agreement of 62.4% on the entire document level and 86.8-95.8% on the individual concept level. Lowest agreement was in the most advanced lesion and the location of the most advanced lesion. Adjudication was done between the three annotators with agreement on 499 of the 500 (99.8%) documents and 2999 of the 3000 (99.97%) individual concepts. The one excluded concept was the location of the highest level of pathology. This instance had an advanced adenoma in the proximal colon and a pathology report stating both hyperplastic and adenomas in the rectum with colonoscopy report of multiple lesions removed with largest > 10 mm. It was determined that since the > 10 mm lesion in the rectum could either be hyperplastic or adenomatous, it was not possible to correctly classify this lesion.

NLP testing was done on the 499 documents with agreement in triplicate (gold standard). The training set was used for manipulation of the concept identification and details of these results are seen in Table 1.

The primary aim of accurately identifying the highest level of pathology was found to have an accuracy of 0.97 (97%) with an f-measure of 0.96. There was variation between the different pathology categories with all carcinomas being correctly identified. The lowest precision and recall was in the advanced adenoma category, with respective values of 0.91 and 0.92. Results for the test set are presented in Table 2.

The location of the most advanced lesion had an accuracy of 0.97 (97%) and an f-measure of 0.96. Precision ranged from 0.94-0.98 with recall from 0.89-0.99. Results for the test set are shown in Table 2.

The size of the largest adenoma was captured with an accuracy of 0.96 (96%) with an f-measure also of 0.96. Precision varied between 0.95-1 with recall from 0.92-1. Testing results are shown in Table 2.

The number of adenomas removed was captured with an accuracy of 0.84 with an f-measure of 0.62. Precision varied between 0 for > 10 adenomas removed to 1 for no adenomas removed. Recall varied between 0-0.95. Testing results from the NLP processing are presented in Table 2.

Unique concepts were also captured with concepts of diverticulosis and hemorrhoids having an accuracy of an accuracy and f-measure of 1.

The NLP system was run over the entire cohort (including non-annotated documents) to establish feasibility of the system outside of the test setting. There were 138 cases of adenocarcinoma (2.2%), 999 (15.7%) cases of advanced adenoma as the highest level of

pathology, 2,967 cases (46.5%) with the highest level of pathology of adenoma, 1,331 (20.9%) with the highest pathology of hyperplastic, and 944 (14.8%) with non-significant pathologic findings. Figure 2 shows the derived highest level of pathology and their frequencies. The non-annotated documents that were processed by the NLP system were manually reviewed by two of the authors and were found to have high accuracy - comparable to that of the testing set.

Discussion

Within gastroenterology the use of NLP has been very limited²⁶⁻²⁹. Several recent papers have looked at NLP use within the field of gastroenterology, focusing within colonoscopy quality metrics for concept extraction^{28, 30}.

Harkema published in 2011 looking at 21 variables for 19 quality measures from colonoscopy and pathology free-text reports. They evaluated these measures on 453 colonoscopy reports with 226 pathology reports linked. They were able to find an accuracy of the NLP system of 0.89 (0.62-1) for the metrics and an overall f-measure of 0.74 (0.49-0.89). The average agreement score was 0.62 (0.09-0.86) for the quality metrics between manual review and NLP derived outcomes²⁸.

Mehotra subsequently used the previously evaluated work and published in *Gastrointestinal Endoscopy* in 2012 looking at NLP-based quality measures in a cross-sectional study to generate provider performance on 7 of the quality measures. They were able to use 9 hospitals across their health system with 24,157 colonoscopy and pathology reports and outputted individual based quality measurements such as adenoma detection rate with a range of 14.9-33.9%³⁰.

Secondary use of clinical data, use of data for other than its initial purpose, for measures of severity or quality is an important concept that is likely to accelerate as clinical systems are introduced into electronic health records³¹⁻³³. Though the initial intent of these systems may not to provide a research platform, research will be required to decrease the cost of collecting and accurately aggregating. This is true within the field of gastroenterology, particularly for widespread use of free-text documents for endoscopic procedure documentation.

Our results show that we were able to accurately (84-97%) capture key clinical concepts regarding most advanced findings, polyp type, number and location. This type of information has many potential applications from tracking pathology specific detection rates, to clinical decision support systems for recommended follow-up intervals. Although our data were limited to first time colonoscopy within the VA system, however, there is no reason to believe that the endoscopic and pathologic findings would be labeled (or designated) differently than for subsequent colonoscopies.

These findings may be used to track and report adenoma detection rate by individual provider without the need for manual review of both the clinical document and the pathology report. It is also a potential way to track beyond simply adenoma detection into

more specific categories such as advanced adenoma, small adenoma (≤ 5 mm), or even proximal hyperplastic polyps for sessile serrated polyp (SSP) detection rates.

We were also able to show the deficiencies of manual review efforts for extraction of complex data. Among three gastroenterologists for 3000 concepts we had an initial agreement of 62.4% at the document level and 86.8-95.8% at the concept level. This may reflect the human error factor for large data extraction as the vast majority of these were errors of omission¹⁶ (e.g. did not see the word diverticula in a sentence). After adjudication and establishment of a gold standard for a concept, the NLP system was able to outperform expert reviewers, enhancing the advantage of using NLP systems for complex data extraction.

There are several limitations to our study. The first is that we used reports from a single institution (similar to Harkema/Mehotra) and that these were completed using a template driven software system (Wolters Kluwer®, Provation Medial®). This allowed for limited linguistic variation among the reports despite the numerous providers involved and thus may not be as generalizable. However, since the template system is widely used many sites would benefit from an unaltered extraction method and further refinement of the NLP when looked at across sites may produce similar results.

Secondly, we used a previously non-validated method for establishing a gold standard for NLP analysis. The use of three expert gastroenterologists as the annotators revealed an initial agreement of 62.4%, however the majority of these were errors of oversight of plainly visible information and likely reflect what a reviewing individual would accomplish given the amount of data points to be captured. We decided on the adjudication process and annotation of all the documents in triplicate in order to more robustly justify the findings of the NLP system. This further validates that an NLP system for extraction may be far superior to manual review given the lack of fatigability and missing of concepts.

Lastly, we used concepts that had high likelihood for capture and may not have required an advanced NLP system for processing. Since the majority of the documents were from template driven systems, it is feasible that the concepts could have been captured from the underlying software database. However, these documents are often edited in “free-text mode” after the concept has been generated within the software thus rendering the concept incorrect. NLP has the opportunity to look at the final written word and make a decision on the concept as the gold standard. Future studies will be looking at this template concept vs. NLP concept to assess the best way to extract the data.

Having an accurate method for extraction and aggregation of previously inaccessible data is believed to be part of the future of medicine, including gastroenterology. We have accurately collected validated concepts for linked pathology and clinical reports and prepared them for clinical research or for linked clinical care.

Conclusion

A freely available and easily implementable NLP system may be able to extract important concepts from colonoscopy and pathology reports and enable secondary use of data such as

highest level of pathology found. These findings may have clinical implications for tracking detection rates of not only adenomas, but also potentially advanced adenomas and proximal large hyperplastic polyps. Further research is needed on the performance of this NLP tool across multiple centers.

Acknowledgments

This work was performed at the Regenstrief Institute, Indianapolis, Indiana, and was supported in part by Grant Number T15LM007117 from the National Library of Medicine and by Veterans Administration Merit Review Grant IIR 08-062. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Library of Medicine or the National Institutes of Health.

Thank you to the Roudebush VAMC (Center of Excellence on Implementing Evidence-Based Practice, Health Services Research & Development (HSR&D)), Eric Sherer, Ph.D., Brian Brake, and Jason Larson for the data extraction and manipulation assistance.

Abbreviations

ICD-9	International Classification of Diseases, 9 th edition
CPT	Current Procedural Terminology
NLP	Natural language processing
cTAKES	Clinical text analysis and knowledge extraction system
SNOMED-CT	Systematized nomenclature of medicine - Clinical terms
UMLS	Unified medical language system
UIMA	Unstructured information management architecture
CUI	Concept unique identifier

References

1. Seeff LC, Richards TB, Shapiro JA, et al. How many endoscopies are performed for colorectal cancer screening? Results from CDC's survey of endoscopic capacity. *Gastroenterology*. 2004; 127:1670–7. [PubMed: 15578503]
2. Kahi CJ, Imperiale TF. Flexible sigmoidoscopy screening reduced colorectal cancer incidence and mortality in older adults. *Ann Intern Med*. 2012; 157:JC3–3. [PubMed: 22986395]
3. Schoen RE, Pinsky PF, Weissfeld JL, et al. Colorectal-cancer incidence and mortality with screening flexible sigmoidoscopy. *N Engl J Med*. 2012; 366:2345–57. [PubMed: 22612596]
4. Zauber AG, Winawer SJ, O'Brien MJ, et al. Colonoscopic polypectomy and long-term prevention of colorectal-cancer deaths. *N Engl J Med*. 2012; 366:687–96. [PubMed: 22356322]
5. Screening for colorectal cancer: recommendation and rationale. *Ann Intern Med*. 2002; 137:129–31. [PubMed: 12118971]
6. 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. [PubMed: 18384785]
7. Rex DK, Johnson DA, Anderson JC, et al. American College of Gastroenterology guidelines for colorectal cancer screening 2009 [corrected]. *Am J Gastroenterol*. 2009; 104:739–50. [PubMed: 19240699]

8. Qaseem A, Denberg TD, Hopkins RH Jr, et al. Screening for colorectal cancer: a guidance statement from the American College of Physicians. *Ann Intern Med.* 2012; 156:378–86. [PubMed: 22393133]
9. Schoenfeld P, Cash B, Flood A, et al. Colonoscopic screening of average-risk women for colorectal neoplasia. *N Engl J Med.* 2005; 352:2061–8. [PubMed: 15901859]
10. Lieberman DA, Weiss DG, Bond JH, et al. Use of colonoscopy to screen asymptomatic adults for colorectal cancer. Veterans Affairs Cooperative Study Group 380. *N Engl J Med.* 2000; 343:162–8. [PubMed: 10900274]
11. Ko CW, Dominitz JA, Green P, et al. Accuracy of Medicare claims for identifying findings and procedures performed during colonoscopy. *Gastrointest Endosc.* 2011; 73:447–453. e1. [PubMed: 20950800]
12. Lieberman D. Pitfalls of using administrative data for research. *Dig Dis Sci.* 2010; 55:1506–8. [PubMed: 20428943]
13. Medicine NLo. SNOMED CT. 2012; Volume 2012
14. Medicine, NLo. Unified Medical Language System. 2012; Volume 2012
15. Kennedy HG. Fatigue and fatigability. *Lancet.* 1987; 1:1145. [PubMed: 2883468]
16. Kennedy HG. Fatigue and fatigability. *Br J Psychiatry.* 1988; 153:1–5. [PubMed: 2906263]
17. AMIA. 2012.
18. Wikipedia.org. Natural language processing. 2012; Volume 2012
19. Ohno-Machado L. Realizing the full potential of electronic health records: the role of natural language processing. *J Am Med Inform Assoc.* 2011; 18:539. [PubMed: 21846784]
20. 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. [PubMed: 20819853]
21. Unstructured Information Management Applications (UIMA).
22. Lindberg C. The Unified Medical Language System (UMLS) of the National Library of Medicine. *J Am Med Rec Assoc.* 1990; 61:40–2. [PubMed: 10104531]
23. Walden V. Defining the UMLS Metathesaurus. *J AHIMA.* 1996; 68:38. [PubMed: 10163160]
24. Medicine USNLo. UMLS Metathesaurus Fact Sheet. 2012; Volume 2012
25. Van Rijsbergen, CJ. Information Retrieval: Butterworth. 1979.
26. 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. [PubMed: 21083794]
27. 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. [PubMed: 20595304]
28. 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. [PubMed: 21946240]
29. 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.* 2012; 32:188–97. [PubMed: 21393557]
30. 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; 75:1233–1239. e14. [PubMed: 22482913]
31. Rea S, Pathak J, Savova G, et al. Building a robust, scalable and standards-driven infrastructure for secondary use of EHR data: The SHARPn project. *J Biomed Inform.* 2012; 45:763–71. [PubMed: 22326800]
32. Holzer K, Gall W. Utilizing IHE-based Electronic Health Record systems for secondary use. *Methods Inf Med.* 2011; 50:319–25. [PubMed: 21431246]
33. Jain A, Hayden S. Using electronic health records to collect patient-specific performance measures and outcomes. *AMIA Annu Symp Proc.* 2008:1242–3. [PubMed: 18998832]

Figure 1. Flow chart of document usage for NLP testing and training

* Unable to categorize location of highest lesion due to pathology report mixed categorization of pathology at same location.

Figure 2.
Highest level of pathology for linked records.

Table 1

Precision, recall, accuracy, and F-measure for colonoscopy/pathology free text documents for the training set.

		Precision	Recall	Accuracy	F-measure
Most advanced lesion	None	1	1		
	Hyperplastic polyp	1	1		
	Tubular adenoma	0.99	1		
	Advanced adenoma	1	0.96		
	Carcinoma	n/a *	n/a *		
Overall				0.99	0.99
		Precision	Recall	Accuracy	F-measure
Location of most advanced lesion	None	0.94	1		
	Proximal	0.96	0.98		
	Distal	1	1		
	Proximal and distal equal	1	0.89		
Overall				0.98	0.97
		Precision	Recall	Accuracy	F-measure
Largest adenoma removed	None	0.93	0.98		
	<= 5 mm	0.98	0.92		
	6-9 mm	0.94	0.94		
	>= 10 mm	0.96	0.96		
Overall				0.95	0.95
		Precision	Recall	Accuracy	F-measure
Number of adenomas removed	0	1	0.98		
	1-2	0.84	0.79		
	3-10	0.55	0.63		
	>10	0.6	0.75		
Overall				0.83	0.77

* No adenocarcinomas within the training set data to test against.

Table 2

Precision, recall, accuracy, and F-measure for colonoscopy/pathology free text documents for the test set.

		Precision	Recall	Accuracy	F-measure
Most advanced lesion	None	0.99	0.98		
	Hyperplastic polyp	1	0.98		
	Tubular adenoma	0.98	0.98		
	Advanced adenoma	0.91	0.92		
	Carcinoma	0.9	1		
Overall				0.97	0.96
		Precision	Recall	Accuracy	F-measure
Location of most advanced lesion	None	0.94	0.98		
	Proximal	0.97	0.96		
	Distal	0.98	0.99		
	Proximal and distal equal	0.95	0.89		
Overall				0.97	0.96
		Precision	Recall	Accuracy	F-measure
Largest adenoma removed	None	0.95	1		
	<= 5 mm	0.97	0.96		
	6-9 mm	0.96	0.92		
	>= 10 mm	1	0.92		
Overall				0.96	0.96
		Precision	Recall	Accuracy	F-measure
Number of adenomas removed	0	1	0.95		
	1-2	0.86	0.84		
	3-10	0.64	0.66		
	>10	0	0		
Overall				0.84	0.62