ABSTRACT

Background and aims: Determining surveillance intervals for patients with colorectal polyps is critical but time-consuming and challenging to do reliably. We present the development and assessment of a pipeline that leverages natural language processing (NLP) techniques to automatically extract and analyze relevant polyp findings from free-text colonoscopy and pathology reports. Using this information, individual patients are categorized into six post-colonoscopy surveillance intervals defined by the U.S. Multi-Society Task Force on Colorectal Cancer.

Methods: Using a set of 546 randomly selected colonoscopy and pathology reports from 324 patients in a single health system, we used a combination of statistical classifiers and rule-based methods to extract polyp properties from each report type, associate properties to unique polyps, and classify a patient into one of six risk categories by integrating information from both report types. We then assessed the pipeline's performance by determining the positive predictive value (PPV), sensitivity, and F-score of the algorithm, compared to the determination of surveillance intervals by a gastroenterologist.

Results: The pipeline was developed using 346 (224 colonoscopy and 122 pathology) reports from 224 patients and evaluated on an independent test set of 200 (100 colonoscopy and 100 pathology) reports from 100 patients. We achieved an average PPV, sensitivity, and F-score of 0.92, 0.95, and 0.93, respectively, across targeted entities for colonoscopy. Pathology extraction achieved a PPV, sensitivity, and F-score of 0.95, 0.97, and 0.96. The system achieved an overall accuracy of 92% in assigning the recommended interval for surveillance colonoscopy.

Conclusions: The study demonstrates the feasibility of using machine learning to automatically extract findings and classify patients to appropriate risk categories and corresponding surveillance intervals. Incorporating this system can facilitate proactive and timely follow-up after screening colonoscopy and enable real-time quality assessment of prevention programs and providers.

Keywords: Natural language processing, quality improvement, high-risk polyps, surveillance.

INTRODUCTION

Colorectal cancer (CRC) is the second most common cause of cancer-related mortality in the United States (U.S.). However, 50% of deaths are preventable by screening[1, 2]. Screening for CRC reduces incidence and mortality by identifying cancers and precancerous colon polyps, including high-risk polyps, defined as 1 adenoma or sessile serrated polyp (SSP) ≥ 1cm, 1 adenoma with villous histology, 1 adenoma with high-grade dysplasia, 1 SSP with dysplasia, 1 traditional serrated adenoma (TSA), or 5-10 adenomas or SSP of any size[3]. High-risk polyps are associated with a two- to five-fold increased risk for subsequent CRC[4], and colonoscopic removal has been shown to decrease CRC incidence and mortality[5, 6]. Several professional societies have recommended that individuals with high-risk polyps undergo surveillance with repeat colonoscopy three years after diagnosis[3]; yet, over 50% of individuals with these polyp subtypes do not receive colonoscopy within the recommended surveillance interval[7-11]. Recent results from the Population-based Research Optimizing Screening through Personalized Regimens (PROSPR) consortium demonstrated surveillance rates between 18% and 53% in four separate integrated health delivery systems[10]. At our institution, only 26% of patients with high-risk polyps complete surveillance colonoscopy at three years, followed by 40% by 3.5 years and 54% by five years[12], which is due to many factors, including low rates of surveillance interval documentation, incorrect documentation of intervals, and lack of patient outreach around the time surveillance is due[13].

A primary barrier to the effective management of colorectal polyps is an inefficient and time-consuming manual process for identifying and tracking patients who require 3-year colonoscopy surveillance. Clinicians are tasked with synthesizing and interpreting information about polyp size, number, and histology from semi-structured colonoscopy and pathology reports, which are usually available at different times and in various locations within the electronic health record (EHR). Once these two independent reports have been located and reviewed, the clinicians need to combine information from both reports and reference detailed guidelines to

determine where their patient fits within one of six different surveillance intervals based on CRC risk[11, 14, 15] (**Figure 1**).

In this study, we describe the development and evaluation of a tool that uses machine learning (ML) and natural language processing (NLP) techniques in an automated pipeline to help physicians synthesize information from colonoscopy and pathology reports to assess CRC risk and recommend the appropriate surveillance interval. We demonstrate how the pipeline can reliably extract information from both pathology and colonoscopy reports as the basis for tracking patients to enable timely proactive surveillance of colorectal polyps. The work is one step towards our team's overall goal to improve adherence to CRC surveillance guidelines in our large health system.

MATERIALS AND METHODS

Data sources

The process for selecting the data is summarized in **Figure 2**. Following an institutional review board-approved protocol, we randomly selected a total of 546 colonoscopy and pathology reports from UCLA Health's electronic health record (Epic Systems, Verona, WI) from patients aged 18 or older who underwent colonoscopy procedures between March 2013 and April 2019. Excluded cases included patients who had 1) a personal or family history of CRC, 2) a personal or family history of colorectal polyps, 3) inflammatory bowel disease, 4) familial/hereditary polyposis syndromes. A subset of 346 colonoscopy and pathology reports (122 paired colonoscopy and pathology reports and 102 unpaired colonoscopy reports) representing 224 patients were initially annotated and used to train and test each step of our automated pipeline. During the annotation process, a set of 10 colonoscopy and pathology reports that were incorrectly paired (e.g., the pathology report was generated after the colonoscopy report but did not refer to the same colonoscopy procedure). These reports were not used to validate the surveillance interval recommendation algorithm but were included as part of the training set. Each

colonoscopy and pathology report was split into sections based on heading (e.g., Findings, Final Diagnosis) to maximize the number of training and testing examples, treating each section with a mention of polyps as an independent example. Consequently, we used 224 colonoscopy reports comprised of 417 sections and 122 pathology reports comprised of 241 sections.

A set of 200 reports (100 colonoscopy reports and 100 corresponding pathology reports) from 100 patients was retrieved from the the EHR to evaluate the complete system. We independently sampled cases from each surveillance interval group to ensure balanced representation across groups. **Table 1** summarizes the breakdown of reports for each pipeline development phase and evaluation.

Annotations

The annotation guidelines were developed by a gastroenterologist (FM) and two pathologists (BN, YK). They identified 13 polyp-related entities of interest found in colonoscopy and pathology report text (**Table 2**). The annotation guidelines then defined which entities to target for extraction from the semi-structured reports and served as a framework for representing structured polyp data.

Two physicians (AM, CS) and a trained analyst (OK) annotated the reports using the annotation guidelines and the brat rapid annotation tool[16]. An initial set of 50 reports (25 colonoscopy and 25 pathology) was annotated by the analyst and reviewed by a physician (CS or AM). These reviews occurred in batches, and discrepancies were discussed. Two types of annotations were used for this task: 1) entities, which assign pre-defined types to spans of text, and 2) relations, which label connections between entities. Entity annotations were used to train the statistical classifier to identify those entities containing polyp properties and extract their values for processing. The relation annotations were used to validate the process by which those properties, once identified, were grouped and associated with distinct polyps.

To develop and test the surveillance interval classification and to establish the independent test set, two physicians (FM, CS) independently reviewed and determined a

consensus recommended surveillance interval based on a review of both colonoscopy and pathology reports for the patient and according to guidelines[17] (**Figure 1**). Disagreements (15 total; 7%) were discussed, and a consensus was reached in all cases.

Pipeline overview

The overall pipeline (**Figure 2)** consists of three modules: 1) a *polyp property extraction* module identifies mentions of polyps and their relevant properties from colonoscopy and pathology reports; 2) a *polyp grouping* module relates properties associated with an individual polyp (e.g., morphology, location, size); and 3) a *surveillance interval classification* module integrates information between colonoscopy and pathology reports to determine the recommended interval for a repeat colonoscopy. The pipeline was developed using Python and the spaCy framework, an open-source library for NLP[18].

1. Polyp property extraction

The extraction module identifies mentions of polyps and related properties (e.g., size, location) that are documented in unstructured colonoscopy and pathology reports. We focused on the *Findings* and *Impressions* sections of colonoscopy reports and the *Final Diagnosis* and *Gross Description* sections of pathology reports. Within these sections, sentences are tokenized into individual words and fed into a statistical classifier followed by a rule-based pattern matcher. The statistical classifier was trained to identify properties using annotations within the development set. The classifier calculates the probability of a token (or set of tokens) labeled as one of 13 entities summarized in **Table 2** and chooses the most likely label. The rule-based pattern matcher serves two purposes: 1) to provide a secondary method for capturing key information (e.g., histology) that may influence surveillance interval; and 2) to examine each entity's context to identify common scenarios that result in the assignment of incorrect entities (e.g., erroneously assigning the size of a snare as the size of a polyp).

2. Polyp grouping

The grouping module aims to associate the extracted properties (e.g., number, size) with their unique polyp mentions. Colonoscopy and pathology reports frequently document more than one polyp finding. Accurately enumerating the total number of polyps is important in determining the appropriate surveillance interval. **Figure 3** illustrates how polyps and their properties are associated. In colonoscopy reports, individual sentences often refer to a single polyp (**Figure 3a**). Alternatively, a sentence may describe two or more unique polyp findings, in which case they might be distinguished by anatomical location (**Figure 3b**). In pathology reports, polyp mentions are often grouped based on location and the order that the samples were collected (**Figure 3c**). To handle these scenarios, the pipeline processes each property in sequence, creating a new polyp finding each time a unique location or histology is encountered.

3. Surveillance interval classification

The final module combines all polyp findings extracted from the colonoscopy and pathology reports to classify the patient into one of six risk-stratified categories (1, 2, 3, 4, 5, 6), each corresponding to a surveillance interval. Comparing the structured polyp findings to the conditions in **Figure 1**, patients are assigned to these intervals based on a rule-out strategy. **Figure 4** provides an illustrative example. Given the example polyps extracted from the colonoscopy and pathology reports, the module first considers the total number of polyps and the size of the largest polyp from the colonoscopy report. Based on the finding that a single, small (< 10mm) polyp was identified, risk categories 4 and 6 can be ruled out. Considering information from the pathology report, we can determine that the polyp is ‘sessile serrated’ without cytologic dysplasia. Therefore, categories 1 and 2 can be ruled out. Finally, the intersection of the two independently created sets of candidate categories is considered. Based on the combined summary of all extracted properties—a quantity of 1, no large polyp, no cytologic dysplasia—we rule out category 5. Thus, the patient falls into category 3, and a surveillance procedure in 5-10 years is recommended.

Evaluation

The primary study outcomes were performance measures for the final end-to-end NLP pipeline, as measured by positive predictive value, sensitivity, and F-Score based on a comparison of pipeline outputs to human annotations. Performance of the pipeline was assessed in four parts. First, the polyp extraction module was evaluated using a held-out set of 67 colonoscopy and 37 pathology reports. Second, the polyp grouping module was evaluated using a held-out test set of 180 colonoscopy and 96 pathology reports. Third, the surveillance interval classification module was evaluated using a held-out test set of 89 paired colonoscopy and pathology reports. The polyp grouping and surveillance interval classification models were given human annotations as inputs to characterize the upper bound of the module's performance. Finally, we performed a final end-to-end assessment of the entire pipeline using an independent test set of paired colonoscopy and pathology reports from 100 patients. For all assessments, we determined PPV, sensitivity, and F-score based on comparing the output of each module and the recommended surveillance interval as determined by human readers (AM, CS, FM).

RESULTS

NLP reliably extracts polyp descriptions from free-text reports

Our polyp extraction pipeline performed consistently high across both colonoscopy (overall F-score 0.93) and pathology reports (overall F-score 0.96) (**Table 3**). Descriptions such as morphology that have relatively consistent terminology achieved the highest PPV and sensitivity. Performance on pathology reports was higher overall, likely due to the structured pathology reports that utilize more standardized terminology than colonoscopy reports.

Polyp properties are accurately associated with unique polyp mentions

Our approach for associating extracted properties to the correct polyp mention achieved overall F-scores of 0.95 and 0.96 for colonoscopy and pathology reports, respectively (**Table 4**).

Surveillance intervals are assigned with high precision and recall

The distribution of cases used to test our recommendation approach across all six surveillance interval categories is summarized in **Table 1**. **Table 5** reports the performance across all risk categories. Performance ranged from a PPV of 0.91-1.00, a sensitivity of 0.85-1.00, and an F-score of 0.88-1.00 across all categories. The lowest-performing category, Category 4, was most often misclassified as Category 5, reflecting the highly overlapping nature of these risk categories. Category 5 is considered higher risk with a “3 year” surveillance interval compared to Category 4, which calls for surveillance in “3-5 years”.

Independent test set evaluation

Reports from 100 patients that were not previously used to develop or test pipeline modules were used to evaluate the entire pipeline. Surveillance intervals that were outputted by the system were compared to the surveillance recommendations generated by human readers (AM, CS, FM). Each of the module steps was executed in sequence without any manual correction. Overall, our system achieved F-scores ranging from 0.89 to 0.97 across all categories (**Table 6**). Category 5 achieved the lowest F-score (0.89), reflecting the limitations of our defined rules in disambiguating Category 5 from Category 6.

DISCUSSION

This study demonstrates that an algorithm using NLP and ML techniques can integrate information from colonoscopy and pathology reports to accurately categorize patients into groups by guideline-concordant surveillance intervals. On average, our system achieved a PPV of 0.92 and a sensitivity of 0.95 when extracting targeted polyp properties from colonoscopy reports. For pathology report extraction, the average PPV was 0.95, and sensitivity was 0.97. Using integrated data from colonoscopy and pathology reports, the complete system's overall accuracy was 92% for assigning the recommended interval for surveillance colonoscopy. Our methods and findings are consistent with other studies that have used NLP to determine surveillance intervals; however, our method is unique in implementing the recent post-polypectomy surveillance guidelines,

including sessile serrated lesions and large hyperplastic polyps, and improving generalizable of our algorithm by using a training set that reflects the high variability of language and structure of clinical reports written by multiple authors[19, 20].

The detection and removal of high-risk polyps is the cornerstone of CRC prevention and screening. Thus, it is essential to accurately determine surveillance intervals for patients who undergo screening and surveillance to minimize CRC and CRC-related mortality. The automated approach presented can help minimize error seen in the manual determination of surveillance intervals, automate surveillance interval determination, simplify clinical processes, and potentially increase adherence to surveillance guidelines. The system also has the potential to assist busy providers who must access data from several areas of the patient chart over multiple encounters to determine the appropriate CRC surveillance interval for each patient. Given the sheer number of screening colonoscopy performed daily, our system has great potential to improve CRC outcomes in health systems that face challenges to identifying and then achieving timely surveillance for patients with colorectal polyps[4].

A growing number of studies have shown the promise of applying artificial intelligence (AI) to gastrointestinal endoscopy, including polyp detection and characterization[21]. NLP applications have been explored in various contexts, primarily to calculate quality improvement metrics. For example, NLP has been used to extract mentions of adenomas and sessile serrated polyps to compute the adenoma detection rate, a common metric for colonoscopy quality improvement[22]. In studies like ours, NLP has been demonstrated to be an efficient alternative to manual synthesis of colonoscopy and pathology findings with high accuracy for polyp risk stratification (>84%) and guideline-concordant surveillance recommendations (>90%), performing better than clinicians in some cases[19, 23-28]. NLP has also been used to improve documentation of colonoscopy findings, assess colonoscopy quality, and generate colonoscopy result letters to patients and providers [24, 25, 29].

We envision several potential applications of our ML and NLP approach in health systems and gastroenterology practices. First, it can be used as a proactive and real-time approach to surveillance colonoscopy. By serving as a clinical decision support tool, this innovative and efficient approach can reduce the workload necessary to determine surveillance intervals and assist with the recall of patients with high-risk polyps (e.g., adenomas with high-risk features) for on-time surveillance. It can also be expanded to address other types of polyps (e.g., low-risk adenoma) that require longer surveillance intervals (e.g., 5 or 10 years). Second, the pipeline facilitates data-driven quality improvement programs that can help ensure timely follow-up. Metrics such as adenoma detection rate and adherence to recommended surveillance intervals can be easily calculated, allowing health systems to understand their screening programs' performance. Third, information extracted and structured from colonoscopy and pathology reports can be incorporated into risk models being developed to estimate the lifetime risk of CRC. Models that incorporate more detailed clinical information may achieve higher PPV than existing risk stratification methods.

This work is not without limitations. Because polyps are often not uniquely identified in colonoscopy reports at our institution, accurately estimating the total number of polyps is challenging. Polyp quantity and histology type have the largest impact on surveillance interval. Subtle differences in these two entities distinguish risk categories 2, 4, and 5; in some cases, the algorithm overestimates the number of polyps, which results in an incorrect classification. Moreover, finding a set of rules that will consistently disambiguate multiple polyps grouped together and correctly assign their individual properties is challenging. In particular, colonoscopy reports can have variable sentence structure depending on the physician. As such, our approach's performance may vary across providers and institutions. Electronic reporting systems that standardize individual polyps reporting would help in this regard. Another source of error is our preference to assign patients to the higher risk category (and more frequent surveillance) if uncertain. For example, Category 5 cases were often incorrectly assigned by our algorithm as

Category 6. While shorter surveillance intervals could potentially catch CRCs earlier when risk is uncertain, the incorrect classification may promote colonoscopy overuse in some cases. Lastly, the development and implementation of NLP algorithms like this one require resource allocation, both in constructing the pipeline, its validation, deployment in practice, and maintenance. Despite the high performance reported, a human-in-the-loop paradigm where a patient coordinator oversees and reviews assigned surveillance intervals is still recommended to ensure that patients receive timely and appropriate care.

The next steps in this multi-phase effort are to prospectively validate our NLP algorithm for patients who present for screening and to compare surveillance classifications to provider classifications. Once the system achieves consistent performance, we will incorporate the algorithm into our health system's preventive care EHR health maintenance module. This integration will allow us to automate surveillance reminders to patients (mailed and electronic messaging through the EHR-based patient portal) and providers (electronic messaging through the EHR) and improve physician-patient communication about screening colonoscopy findings in our health system. We plan to study the implementation, effectiveness, and acceptability of this multi-level intervention and to disseminate our findings and code to other health system partners to further study the algorithm in different patient populations and settings. This work will streamline provider workload and improve health outcomes in our health system and beyond for one of the most common and deadly malignancies in the U.S.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the financial support from the Melvin and Bren Simon Gastroenterology Quality Improvement Program, a UCLA Health Innovation Grant, and the National Science Foundation grant #1722516.

REFERENCES

1. Siegel RL, Miller KD, Goding Sauer A, et al. Colorectal cancer statistics, 2020. CA Cancer J Clin 2020;70:145-164.
2. Zauber AG. The impact of screening on colorectal cancer mortality and incidence: has it really made a difference? Dig Dis Sci 2015;60:681-91.
3. Gupta S, Lieberman D, Anderson JC, et al. Recommendations for Follow-Up After Colonoscopy and Polypectomy: A Consensus Update by the US Multi-Society Task Force on Colorectal Cancer. Am J Gastroenterol 2020;115:415-434.
4. Joseph DA, Meester RG, Zauber AG, et al. Colorectal cancer screening: Estimated future colonoscopy need and current volume and capacity. Cancer 2016;122:2479-86.
5. 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.
6. He X, Hang D, Wu K, et al. Long-term Risk of Colorectal Cancer After Removal of Conventional Adenomas and Serrated Polyps. Gastroenterology 2020;158:852-861 e4.
7. Laiyemo AO, Pinsky PF, Marcus PM, et al. Utilization and yield of surveillance colonoscopy in the continued follow-up study of the polyp prevention trial. Clin Gastroenterol Hepatol 2009;7:562-7; quiz 497.
8. Schoen RE, Pinsky PF, Weissfeld JL, et al. Utilization of surveillance colonoscopy in community practice. Gastroenterology 2010;138:73-81.
9. Braschi C, Pelto DJ, Hennelly MO, et al. Patient-, provider-, and system-level factors in low adherence to surveillance colonoscopy guidelines: implications for future interventions. J Gastrointest Cancer 2014;45:500-3.
10. Chubak J, McLerran D, Zheng Y, et al. Receipt of Colonoscopy Following Diagnosis of Advanced Adenomas: An Analysis within Integrated Healthcare Delivery Systems. Cancer Epidemiol Biomarkers Prev 2019;28:91-98.
11. Iskandar H, Yan Y, Elwing J, et al. Predictors of Poor Adherence of US Gastroenterologists with Colonoscopy Screening and Surveillance Guidelines. Dig Dis Sci 2015;60:971-8.
12. May F.P, Corona E., Yang L., Nguyen S., Lin C., Huang M., Shao P., Mwengela D., Didero M., Asokan I., Bui A., Hsu W., Maehara C., Naini B., Kang Y., Bastani R. Low Rates of Colonoscopic Surveillance Among Patients with High-Risk Adenomas. Abstract presentation at American College of Gastroenterology; San Antonio, TX; October 2019. Manuscript in preparation.

13. Myint A, Corona E, Yang L, et al. Gastroenterology visitation and reminders predict surveillance uptake for patients with adenomas with high-risk features. Sci Rep 2021;11:8764.
14. Patel N, Tong L, Ahn C, et al. Post-polypectomy Guideline Adherence: Importance of Belief in Guidelines, Not Guideline Knowledge or Fear of Missed Cancer. Dig Dis Sci 2015;60:2937-45.
15. Gupta S, Lieberman D, Anderson JC, et al. Recommendations for Follow-Up After Colonoscopy and Polypectomy: A Consensus Update by the US Multi-Society Task Force on Colorectal Cancer. Gastroenterology 2020;158:1131-1153 e5.
16. Stenetorp P, Pyysalo S, Topić G, et al. BRAT: a web-based tool for NLP-assisted text annotation, In Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics, 2012.
17. Gupta S, Lieberman D, Anderson JC, et al. Recommendations for Follow-Up After Colonoscopy and Polypectomy: A Consensus Update by the US Multi-Society Task Force on Colorectal Cancer. Gastroenterology 2020;158:1131-1153.e5.
18. Honnibal M, Montani I, Van Landeghem S, et al. spaCy: Industrial-strength Natural Language Processing in Python: Zenodo, 2020.
19. Imler TD, Morea J, Imperiale TF. Clinical decision support with natural language processing facilitates determination of colonoscopy surveillance intervals. Clin Gastroenterol Hepatol 2014;12:1130-6.
20. Karwa A, Patell R, Parthasarathy G, et al. Development of an Automated Algorithm to Generate Guideline-based Recommendations for Follow-up Colonoscopy. Clin Gastroenterol Hepatol 2020;18:2038-2045 e1.
21. Pannala R, Krishnan K, Melson J, et al. Emerging role of artificial intelligence in GI endoscopy. Gastrointestinal Endoscopy 2020;92:1151-1152.
22. Nayor J, Borges LF, Goryachev S, et al. Natural Language Processing Accurately Calculates Adenoma and Sessile Serrated Polyp Detection Rates. Dig Dis Sci 2018;63:1794-1800.
23. 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.
24. Raju GS, Lum PJ, Slack RS, et al. Natural language processing as an alternative to manual reporting of colonoscopy quality metrics. Gastrointest Endosc 2015;82:512-9.

25. Imler TD, Morea J, Kahi C, et al. Multi-center colonoscopy quality measurement utilizing natural language processing. Am J Gastroenterol 2015;110:543-52.
26. 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.
27. 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.
28. Abdul-Baki H, Schoen RE, Dean K, et al. Public reporting of colonoscopy quality is associated with an increase in endoscopist adenoma detection rate. Gastrointest Endosc 2015;82:676-82.
29. Skinner CS, Gupta S, Halm EA, et al. Development of the Parkland-UT Southwestern Colonoscopy Reporting System (CoRS) for evidence-based colon cancer surveillance recommendations. J Am Med Inform Assoc 2016;23:402-6.

FIGURE AND TABLES

FIGURES

Figure 1: Conditions for each polyp risk category and corresponding CRC colonoscopy surveillance interval recommendation; summarizes guidelines published in Gupta et al[17].

Figure 2: (a) Overview of the NLP pipeline. Inputs are free-text colonoscopy and pathology reports. Output is a colonoscopy surveillance interval recommendation (as categorized in Figure 1). **(b)** Summary of the case selection process. A "mismatched pair" refers to a colonoscopy and pathology report that did not reference the same procedure and was removed during annotation.

Figure 3: Illustration of the inputs/outputs of the polyp grouping module. **(a)** Example of a colonoscopy report with one polyp finding per sentence. **(b)** Example of a colonoscopy report with multiple polyp findings in one sentence. **(c)** Example of a pathology report with single and multiple polyp findings per sample that are grouped by anatomical location. (Note: properties highlighted in yellow are being shared across multiple polyp observations by the grouping algorithm)

Figure 4: Example of colonoscopy surveillance interval recommendation process. First, properties of polyps extracted from colonoscopy report (left) and pathology report (right) are independently summarized and used to narrow down possible risk categories. Then a combined summary is used to determine the final category and recommend the corresponding surveillance interval.

TABLES

Table 1: (A) Distribution of colonoscopy and pathology reports used to train and test each module of the NLP pipeline. (B) Surveillance interval distribution of patients used during pipeline development. (C) Surveillance interval distribution of patients included in the independent test set.

(A)

Module	Report Type	Training	Testing	Total
Polyp Extraction	Colonoscopy	157	67	224
	Pathology	85	37	122
Polyp Grouping	Colonoscopy	44	180	224
	Pathology	26	96	122
Surveillance Interval	Colonoscopy	23	89	112
	Pathology	23	89	112

(B)

Set	Total Patients	10 years	7-10 years	5-10 years	3-5 years	3 years	1 year
Training	23	4	4	4	4	4	3
Testing	89	16	10	10	13	31	9
Total	112	20	14	14	17	35	12

(C)

Total Patients	10 years	7-10 years	5-10 years	3-5 years	3 years	1 year
100	15	15	15	20	20	15

Table 2: Entities targeted from colonoscopy and pathology reports

Colonoscopy Entities		Pathology Entities	
Entity	**Description**	**Entity**	**Description**
Polyp finding	Mention of polyp, lesion, adenoma, etc	Polyp sample	Mention of polyp tissue in pathology sample
Location	Specific position in colon where polyp was found	Location	Specimen source as designated by sample label
Quantity	Number of distinct polyps in observation group	Quantity	Number of tissue fragments in sample
Measured size	Numeric size of polyp (mm or cm)	Measured size	Numeric size of polyp (mm or cm)
General size	General descriptive size of polyp (diminutive, large, etc)	Histology	Diagnosis of polyp histology (tubular adenoma, sessile serrated polyp, etc)
Morphology	Polyp characteristic (pedunculated, broad-based, etc)	High-grade dysplasia	Presence or absence of high-grade dysplasia in sample
		Cytologic dysplasia	Presence or absence of cytologic dysplasia in sample

Table 3: Performance of polyp entity extraction module on a test set of 67 colonoscopy and 37 pathology reports.

Report Type	Metric	Polyp	Morphology	Measured Size	General Size	Quantity	Location	Histology	High-Grade Dysplasia	Cytologic Dysplasia	Overall
Colonoscopy	PPV	0.96	0.99	0.97	0.87	0.81	0.90	-	-	-	0.92
	SENSITIVITY	0.90	0.98	0.96	0.97	0.91	0.97	-	-	-	0.95
	F-SCORE	0.93	0.98	0.96	0.92	0.86	0.94	-	-	-	0.93
Pathology	PPV	0.97	-	0.98	-	0.87	0.91	0.96	0.99	0.96	0.95
	SENSITIVITY	0.98	-	0.98	-	0.90	0.95	0.98	0.99	0.99	0.97
	F-SCORE	0.97	-	0.98	-	0.88	0.93	0.97	0.99	0.97	0.96

PPV: positive predictive value

Table 4: Performance of the polyp grouping module on a test set of 276 reports.

Report Type	Metric	Location	Measured Size	General Size	Quantity	Histology	High-Grade Dysplasia	Cytologic Dysplasia	Overall
Colonoscopy	PPV	0.97	0.98	0.98	0.99	-	-	-	0.98
	SENSITIVITY	0.88	0.94	0.98	0.99	-	-	-	0.93
	F-SCORE	0.92	0.96	0.98	0.99	-	-	-	0.95
Pathology	PPV	0.95	-	-	-	0.94	0.91	0.88	0.93
	SENSITIVITY	1.00	-	-	-	0.99	0.99	1.00	1.00
	F-SCORE	0.97	-	-	-	0.97	0.95	0.94	0.96

PPV: positive predictive value

Table 5: Performance of the surveillance interval classification module on a test set of 89 patients.

Metric	Category 1	Category 2	Category 3	Category 4	Category 5	Category 6	Overall
PPV	1.00	0.91	1.00	0.92	0.94	1.00	0.96
SENSITIVITY	1.00	1.00	0.90	0.85	0.97	1.00	0.95
F-SCORE	1.00	0.95	0.95	0.88	0.95	1.00	0.96

PPV: positive predictive value

Table 6: Performance of full pipeline (extraction, grouping, classification) on an independent test set of 100 patients.

Metric	Category 1	Category 2	Category 3	Category 4	Category 5	Category 6	Overall
PPV	0.93	1.00	0.94	0.90	0.94	0.83	0.92
SENSITIVITY	0.93	0.87	1.00	0.90	0.85	1.00	0.92
F-SCORE	0.93	0.93	0.97	0.90	0.89	0.91	0.92

PPV: positive predictive value