PERFORMANCE OF AN AUTOMATED NATURAL LANGUAGE PROCESSING TOOL TO IDENTIFY GUIDELINE-CONCORDANT SURVEILLANCE INTERVALS AFTER SCREENING COLONOSCOPY

Soroudi, Camille; De Silva, Sadie; Peterson, Emma Kate; Maehara, Cleo Kaiaki; Badiee, Jayraan; Myint, Anthony; Muthusamy, V. Raman; Esrailian, Eric; Hsu, William; May, P. Folasade

Characters: 2893/2900 (including spaces)
Abstract Category: AGA. GI Fellow-Directed QI Session
Due date: 12/1/2022 8:59pm

Introduction: Accurate determination and documentation of post-polypectomy surveillance intervals for screening colonoscopy is essential to reduce colorectal cancer (CRC) incidence and mortality. It is particularly important to ensure timely surveillance for patients with high-risk polyps that warrant 3-year follow-up. In order to improve guideline-concordant surveillance after screening colonoscopy, we sought to validate a previously developed natural language processing (NLP) algorithm that automates determination of post-polypectomy colonoscopy surveillance intervals.

Methods: The study setting is a large, academic healthcare system performing over 17,000 screening colonoscopies per year. We previously developed an automated NLP algorithm to identify, extract and analyze relevant data from free-text colonoscopy (number, size, location of polyps) and pathology reports (polyp histology) to determine post-polypectomy surveillance intervals based on 2020 USMSTF guidelines. We further refined the system through multiple rounds of performance assessment to improve accuracy of the tool. To validate the final algorithm, we used a random selection of screening colonoscopies performed from 2/1/2022-7/31/2022. Two board-certified physicians performed an independent chart review to determine the surveillance interval for each case; 100% inter-rater agreement was reached through discussion. We then determined the performance (sensitivity, specificity, PPV, F-score, accuracy) of the NLP algorithm to identify guideline-concordant surveillance intervals compared to chart review. For analyses, we excluded cases that warranted sooner than 1-year follow-up due to malignancy, inadequate bowel prep, incomplete colonoscopy, or incomplete polyp removal.

Results: Our validation cohort included 458 colonoscopies (unique individuals). Our cohort was 54.6% female with a mean age of 54.5 (s.d=8.2) (Table 1). The most common surveillance interval was 10 years (n=278). The NLP tool correctly classified surveillance intervals with an overall 89.5% sensitivity, 98.9% specificity, 94.5% PPV, 0.92 F-score, and 97.4% accuracy (Table 2). Test characteristics were variable across surveillance intervals (Table 2). Notably, the tool identified high-risk polyps (requiring 3-year surveillance) with very high sensitivity (96.9%) and specificity (99.1%).

Discussion: We developed an automated NLP tool that is highly sensitive and specific at classifying appropriate surveillance intervals after screening colonoscopy, particularly for procedures requiring 3-year follow-up. Next steps include integrating this tool into our electronic health record to flag these high-risk cases. This will assist with planned outreach to recall these patients at 3 years and help our health system prevent delays in appropriate surveillance for patients at higher risk of developing advanced adenomas and CRC.

Table 1. Demographics of patient population for the validation study cohort; N=458 patients.

Demographic Characteristic	**n (%) or mean (SD)**
Mean Age (SD)	54.5 (8.2)
Gender	
Male	208 (45.4)
Female	250 (54.6)
Race/Ethnicity	
Non-Hispanic White	209 (45.7)
Non-Hispanic Black	28 (6.1)
Hispanic	69 (15.1)
Non-Hispanic Asian	61 (13.3)
Non-Hispanic Other (American Indian or Alaskan Native, Native Hawaiian or Pacific Islander, null, and other race categories)	24 (5.2)
Unknown or Declined to Answer	67 (14.6)

Table 2. Performance metrics of NLP risk stratification tool to appropriately classify surveillance intervals after screening colonoscopy; N=458 procedures.

Performance Metric	**10 years (n=278)**	**7-10 years (n=105)**	**5-10 years (n=20)**	**3-5 years (n=23)**	**3 years (n=32)**	**Overall (all intervals)**
Sensitivity	92.4	89.5	70.0	60.9	96.9	89.5
Specificity	98.9	98.9	99.8	97.5	99.1	98.9
PPV	99.2	95.9	93.3	56.0	88.6	94.5
F-Score	0.957	0.926	0.800	0.583	0.926	0.919
Accuracy	95.0	96.7	98.4	95.6	98.9	97.4