Measuring the quality of colonoscopy: Where are we now and where are we going?

Timothy D. Imler, MD

Division of Gastroenterology and Hepatology, Indiana University School of Medicine, Indianapolis, Indiana
Department of Medicine, Indiana University School of Medicine, Indianapolis, Indiana
Division of Biomedical Informatics, Regenstrief Institute, LLC, Indianapolis, Indiana

Thomas F. Imperiale, MD

Division of Gastroenterology and Hepatology, Indiana University School of Medicine, Indianapolis, Indiana
Department of Medicine, Indiana University School of Medicine, Indianapolis, Indiana
Center of Innovation, Health Services Research and Development, Richard L. Roudebush VA Medical Center, Indianapolis, Indiana
Health Services Research, Regenstrief Institute, LLC, Indianapolis, Indiana

“Not everything that can be counted counts, and not everything that counts can be counted.” – *Albert Einstein*

Data to support adenoma detection rate (ADR) as a measure of colonoscopy quality and cancer protection continue to emerge,1 along with evidence that behavior changes favorably when ADR is tracked.2 However, ADR can be a challenging metric to obtain, given the disparity in systems for pairing colonoscopy and pathology reports. Although several surrogate measures have been suggested as appropriate for quantifying quality (eg, ADR, polypectomy rate), the potential for gaming the system makes them less attractive,3 implying that the easiest metrics to count may not be the most useful ones. What is needed are 1 or more robust quality measures and an automated, accurate method for obtaining them. The mean number of adenomas per screening colonoscopy has been proposed as less prone to gaming than the ADR, but no data have linked this measure to outcomes.4

This is the author's manuscript of the article published in final edited form as:

Imler, T. D., & Imperiale, T. F. (2015). Measuring the quality of colonoscopy: Where are we now and where are we going? Gastrointestinal Endoscopy, 82(3), 520–522. http://doi.org/10.1016/j.gie.2015.03.1961

Natural language processing (NLP) is a method of searching within text documents that has been used for obtaining multiple quality measures in endoscopy, especially the ADR (Table 1). Since 2011, several institutions have developed customized in-house automated systems for quantifying ADR, advanced ADR, and site-specific ADR (proximal/distal), precluding the need for manual review. As providers of an effective but expensive and invasive technology, we endoscopists will soon be required to quantify colonoscopy quality for payers.5 Although the potential for a system that uses artificial intelligence to track our services is no longer considered far-fetched, this vision must be tempered with consideration of what and how a metric is to be measured. There is a pressing need for a simple, yet accurate, method for moving toward an optimal system of tracking colonoscopy quality.

In this issue of Gastrointestinal Endoscopy, Raju and colleagues 6 report their experience with the computer application for ADR reporting (CAADRR). The CAADRR was developed to extract colonoscopy reports from existing endoscopy software and link them with associated pathology reports. These reports were then processed with text mining to determine colonoscopy indication (ie, screening) and the findings of adenomas and sessile serrated polyps (SSPs). The authors validated this system against a manually curated data set and showed accuracy rates of 91.3% for detecting screening examinations, 99.4% for adenomas, and 100% for SSPs. These high rates are consistent with previous reports of single-center accuracy with the use of NLP. 7, 8, 9 and 10 From text mining of the reports, Raju et al were able to measure individual endoscopist’s ADRs, which ranged from 22% to 62%, along with gender-stratified ADRs.

The authors are to be congratulated for creating a carefully constructed and sensible tracking system for several quality metrics. Although this work further establishes the ability of computer techniques to obtain a validated measurement of colonoscopy quality, it is not without limitations (as are previous efforts to advance NLP-based quality measurement of colonoscopies).6, 7, 8, 9 and 10 First, there appears to have been no differentiation between a training set and a test set, which increases both the potential for overtraining and the chance of less accurate performance when the method is applied to subsequent patient cohorts. Second, because traditional NLP was not used, the CAADRR system is unable to use context-specific term identification. As a result, the phrase “no … adenoma” would potentially count as an adenoma by the CAADRR system, whereas it would be recognized as a negated term by a complex NLP system; the misspelled word “adnoma” would be potentially missed with CAADRR but identified by NLP. Third, the CAADRR system has no method for finding adenomas 10 mm or larger, omitting the inclusion of this large subgroup of advanced adenomas from specific identification. The number of adenomas per colonoscopy, which may be a better quality metric than ADR, is also not possible to quantify in the CAADRR system. Furthermore, the accuracy of identifying a “screening” colonoscopy may be questioned because patients with rectal bleeding and a positive fecal blood test result were counted as “screening,” whereas those with “family history of colorectal cancer” were excluded without specifically being first-degree relatives. Last, this is a single-center tool that was created on templated endoscopy software with a single pathology reporting system. Although the tool may work well on its “home court,” the methods used to create it may not apply as well to other systems.

Despite advances made by several groups of investigators in automated measuring of selected quality determinations (Table 1), there are and will likely remain several other determinants of colonoscopy quality that are not reflected in procedure note documentation. These include adequacy of luminal distention and examining the proximal sides of flexures, folds, and valves—techniques associated with lower adenoma miss rates.11 Although video recording of colonoscopies and randomly reviewing withdrawal techniques is a possibility,12 the practicality of implementing these practices may be too costly and time-consuming.

In 2015, technology pervades our professional life, with each clinical encounter systematically logged into the electronic medical record. The opportunity to use secondary data to improve the quality of health services and patient outcomes is apparent. However, we must temper our enthusiasm with reality. We must accept that not all patients for a provider or panels of patients among providers are the same. For example, a quality metric such as ADR must be coupled with adjustment for factors that affect risk for colorectal neoplasia (eg, age, gender, certain family histories, results of previous screening tests). An adjusted ADR would appropriately reduce unwanted variations caused by these factors, resulting in a better quality metric. Furthermore, today's best quality metric may not be tomorrow's. Mean adenoma number per screening colonoscopy may be a more robust, more "tamperproof" metric,13 so any technology used for quality monitoring must be adaptable to changing measures of quality. With the challenge of detecting SSPs, an SSP detection rate may be most indicative of high-quality colonoscopy, although variations in pathologic interpretation must be addressed definitively before further study.14

The optimal system will allow an endoscopist to perform a colonoscopy, create a colonoscopy report in an electronic format (endoscopy software or dictation), associate that report to a subsequent pathology report, extract meaningful quality measures (eg, ADR, ADR per procedure, advanced ADR, proximal ADR), adjust these rates for important covariates, and report these to the provider, a quality monitoring system (eg, the GI Quality Improvement Consortium), and the payer, all without any human review. A seamless process using NLP would both allow full transparency of care and encourage practice improvement. We are not too far from such a system, but before we can count a quality measure, we should know and accept the optimal measure(s) to count.

Disclosure

All authors disclosed no financial relationships relevant to this publication.

References

1. Corley DA, Jensen CD, Marks AR, et al. Adenoma detection rate and riskof colorectal cancer and death. N Engl J Med 2014;370:1298-306.

2. Kahi CJ, Ballard D, Shah AS, et al. Impact of a quarterly report card on colonoscopy quality measures. Gastrointest Endosc 2013;77:925-31.

3. Wang HS, Pisegna J, Modi R, et al. Adenoma detection rate is necessary but insufficient for distinguishing high versus low endoscopist performance. Gastrointest Endosc 2013;77:71-8.

4. Kahi CJ, Vemulapalli KC, Johnson CS, et al. Improving measurement of the adenoma detection rate and adenoma per colonoscopy quality metric: the Indiana University experience. Gastrointest Endosc 2014;79:448-54.

5. Rex DK, Schoenfeld PS, Cohen J, et al. Quality indicators for colonoscopy. Gastrointest Endosc 2015;81:31-53.

6. Raju GS, Lum PJ, Slack R, et al. Natural language processing as an alternative to manual reporting of colonoscopy quality metrics. Gastrointest Endosc 2015;82:512-9.

7. 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.

8. 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.

9. 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.

10. Imler TD. Multi-center colonoscopy quality measurement utilizing natural language processing. Am J Gastroenterol. Epub 2015 Mar 10.

11. Rex DK. Colonoscopic withdrawal technique is associated with adenoma miss rates. Gastrointest Endosc 2000;51:33-6.

12. Rex DK, Hewett DG, Raghavendra M, et al. The impact of videorecording on the quality of colonoscopy performance: a pilot study. Am J Gastroenterol 2010;105:2312-7.

13. Denis B, Sauleau EA, Gendre I, et al. The mean number of adenomas per procedure should become the gold standard to measure the neoplasia yield of colonoscopy: a population-based cohort study. Dig Liver Dis 2014;46:176-81.

14. Payne SR, Church TR, Wandell M, et al. Endoscopic detection of proximal serrated lesions and pathologic identification of sessile serrated adenomas/polyps vary on the basis of center. Clin Gastroenterol Hepatol 2014;12:1119-26.

Table 1. Summary of selected studies using natural language processing–based extraction of adenoma detection rate

Study	Objectives	Methods	Results	ADR	What the study adds
Harkema et al (2011)7	Develop NLP application to measure colonoscopy quality	Newly created NLP engine tested against manual review of 453 free-text colonoscopy and 226 pathology reports for identifying data for 19 quality measures	Accuracy of 89% (range<comma> 62%-100%) for the 19 quality measures	Individual endoscopist ADRs range from 14.9% to 33.9%	NLP can identify with reasonable accuracy data elements for quality assessment of colonoscopy
Imler et al (2013)8	Create and test NLP program to identify the most advanced level of pathologic lesions found on colonoscopy	500 linked colonoscopy and pathology reports used to train and test an open-source NLP engine against paired annotation of reports by blinded endoscopists	NLP identified highest level of pathologic change with 98% accuracy. Accuracies for location<comma> size<comma> and number were 97%<comma> 96%<comma> and 84%<comma> respectively	Overall institutional ADR of 46.5% and advanced ADR of 15.7%	NLP can identify the most advanced level of pathologic lesion and its size and anatomic location
Gawron et al (2014)9	Use open-source NLP to assess total<comma> anatomic<comma> and advanced ADRs	Performance of an open-source NLP engine was tested against manual validation of 200 procedures with associated pathology reports	NLP reported screening indication and completed procedure with 98% accuracy<comma> and correct pathologic lesion and location with 94% accuracy	Overall ADR of 20.3%<comma> left-sided ADR of 10.1%<comma> right-sided ADR of 12.5%<comma>and advanced ADR of 4.4%	Established ability to measure advanced and anatomic ADRs
Imler (2015)10	Determine feasibility and performance of NLP in 13 VA	750 random colonoscopies from a sample of	NLP identified highest pathologic lesion with 94.6%-99.8% accuracy	ADR per center ranged from 19.3% to 38%; advanced	Demonstrates high accuracy for pathologic lesion

	centers for pathologic lesion and anatomic location	>40<comma>000 linked colonoscopy and pathology reports were used to train and test an open-source NLP engine against paired annotation of reports by blinded endoscopists	and location with 87%-99.8% accuracy	ADR was 7.7%; SSP detection rate was 0.6%; and proximal ADR was 11.4%	and location among multiple centers
Raju et al (2015)6	Develop NLP mechanism to identify screening colonoscopies and determine ADR	Multiple manually curated reports compared against NLP extraction with single gastroenterology review and mismatches assessed by a separate gastroenterologist	NLP identified 91.3% of screening examinations<comma> 99.4% of reports for adenomas<comma> and 100% of reports for SSPs	Overall ADR was 43%<comma> with individual endoscopist’s ADRs ranging from 22% to 62%	NLP can accurately identify screening examinations and report individual endoscopist’s ADR