Development of a Natural Language Processing System to Identify Timing and Status of Colonoscopy Testing in Electronic Medical Records

Joshua C. Denny, MD MS[1,2]; Josh F. Peterson, MD MPH[1,2,3]; Neesha N. Choma, MD[2,3]; Hua Xu, PhD[1]; Randolph A. Miller, MD[1]; Lisa Bastarache, MS[1]; Neeraja B. Peterson, MD MSc[2]
[1]Department of Biomedical Informatics, Vanderbilt University, Nashville, TN; [2]Division of General Internal Medicine and Public Health, Department of Medicine, Vanderbilt University Medical Center, Nashville, TN; [3]Veterans Administration, Tennessee Valley Geriatric Research Education Clinical Center (GRECC), Nashville, TN

Abstract

Colorectal cancer (CRC) screening rates are low despite proven benefits. We developed natural language processing (NLP) algorithms to identify temporal expressions and status indicators, such as "patient refused" or "test scheduled." The authors incorporated the algorithms into the KnowledgeMap Concept Identifier system in order to detect references to completed colonoscopies within electronic text. The modified NLP system was evaluated using 200 randomly selected electronic medical records (EMRs) from a primary care population aged ≥50 years. The system detected completed colonoscopies with recall and precision of 0.93 and 0.92. The system was superior to a query of colonoscopy billing codes to determine screening status.

Introduction

Screening for CRC is recommended for average-risk individuals age 50 years and older, but is underutilized.[1] Current methods for determining CRC screening status (patient self-report, physician report, medical claims data, and manual chart abstraction) are either time-consuming and expensive or inaccurate. We investigated the use of an NLP system to detect the timing and receipt of colonoscopies.

Methods

The primary outcome measure was recall and precision for the NLP algorithm's identification of completed colonoscopies as compared to a gold standard review of all available data by a physician. We extended an existing NLP system, the KnowledgeMap concept identifier[2], with novel NLP algorithms to 1) identify and interpret time descriptors (e.g., "6/2003" or "5 years ago") and associate them with clinical events; and 2) assign values for concept certainty and status (e.g., "*never had* colonoscopy" or "*needs* a colonoscopy") to each identified concept.

Results

Manual review identified 159 unique completed colonoscopies in the test set. The NLP system identified 1,208 sentences with references to colonoscopies, of which 518 contained a timing reference, and 514 contained a status indicator. The recall and precision were 0.94 and 0.95, respectively, for identifying and assigning timing information to the colonoscopies. The recall and precision of the algorithm to detect status indicators were 0.82 and 0.95, respectively. Overall, the system detected completed colonoscopies with a recall and precision of 0.93 and 0.92, respectively. A query of colonoscopy billing codes identified 106 (67%) of the colonoscopies detected by NLP, and identified one colonoscopy not previously detected.

Conclusion

Using NLP algorithms to detect timing and status on EMR records can identify patients who received colonoscopies with high recall and precision which was superior to billing records queries. These data suggest that a robust system to identify receipt of CRC testing should incorporate NLP methods.

References

1. Winawer S, Fletcher R, Rex D, et al. Colorectal cancer screening and surveillance: clinical guidelines and rationale-Update based on new evidence. *Gastroenterology.* Feb 2003;124(2):544-560.
2. Denny JC, Smithers JD, Miller RA, Spickard A, 3rd. "Understanding" medical school curriculum content using KnowledgeMap. *J Am Med Inform Assoc.* Jul-Aug 2003;10(4):351-362.