Facilitating post-surgical complication detection through sublanguage analysis

Hongfang Liu, PhD[1], Sunghwan Sohn, PhD[1], Sean Murphy[1], Jenna Lovely, PharmD R.Ph[2], Matthew Burton, MD[1,2], James Naessens, ScD[1], David W Larson, MD[2]
[1]Department of Health Sciences Research [2] Department of Surgery
Mayo Clinic College of Medicine, Rochester, MN 55905

Abstract

Identification of postsurgical complications is the first step towards improving patient safety and health care quality as well as reducing heath care cost. Existing NLP-based approaches for retrieving postsurgical complications are based on search strategies. Here, we conduct a sublanguage analysis study using free text reports available for a cohort of patients with postsurgical complications identified manually to compare the keywords identified by subject matter experts with words/phrases automatically identified by sublanguage analysis. The results suggest that search-based approaches may miss some cases and the sublanguage analysis results can be used as a base to develop an information extraction system or support search-based NLP approaches by augmenting search queries.

Introduction

Identification of postsurgical complications is the first step towards improving patient safety and health care quality as well as reducing health care cost (1). Multiple methods currently are being used to identify patients with postsurgical complications. The Agency for Healthcare Research and Quality (AHRQ) Patient Safety Indicators (PSI) are based on administrative hospital ICD-9 discharge diagnoses enhanced with present on admission (POA) modifiers (2). The American College of Surgeons National Surgical Quality Improvement Program (NSQIP) bases its assessment on clinical registry data abstracted directly from a sample of medical records (3). Consumer Reports used POA-enhanced hospital claims data to identify cases with prolonged (i.e., longer than expected) risk-adjusted post-operative lengths of stay (prLOS), which served as a surrogate indicator for serious non-fatal inpatient adverse outcomes (4).

With the rapid adoption of electronic medical records (EMRs) and the accelerated advance of health information technology (HIT), detection of postsurgical complications based on EMRs via natural language processing (NLP) offers a potential powerful alternative to either administrative data or labor-intensive manual chart reviews (5, 6). For example, Murff et al. (6) showed NLP-based postsurgical complication detection achieves higher sensitivity and lower specificity compared to PSI when using a randomly selected sample of Veterans Affairs Surgical Quality Improvement Program (VASQIP)–reviewed surgical inpatient admissions. The NLP approach applied in their study is a generic SNOMED coding system followed by an inclusion-exclusion search. Alsara et al (7) applied a similar search strategy to identify pertinent risk factors for postsurgical acute lung injury.

Meanwhile, in the field of NLP, sublanguage-oriented information extraction (IE) has shown to be promising in extracting a variety of structured data from clinical reports (8, 9). In this paper, we conducted sublanguage analysis related to postsurgical complications utilizing clinical narratives related to a cohort of colorectal surgical patients to compare words/phrases identified by sublanguage analysis with a list of keywords identified by subject matter experts and to explore the potential of developing IE applications for postsurgical complication detection.

Background

Colorectal postsurgical complications

Postsurgical complications may be general or specific to the type of surgery undertaken. In this paper we studied complications that may happen after colorectal surgery. Table 1 shows the complications and their definition considered in this paper adapted from: American Society of Colon & Rectal Surgeons (http://www.fascrs.org/physicians/education/core_subjects/2011/Complications/) and Society of international radiology (http://www.sirweb.org/).

Table 1. Colorectal postsurgical complication definitions

Postsurgical Complication	Abbreviation	Definition
Deep vein thrombosis (DVT)/pulmonary embolism (PE)	DVTPE	DVT is the formation of a blood clot, known as thrombus that usually occurs in the leg although it can happen in other parts of the body. Part of the clot can break off and travel to the lung, where it blocks the oxygen supply, casing heart failure, known as PE.
Bleeding	BLEED	Anastomotic bleeding is common and varies in severity. More serious bleeding can be managed with epinephrine and saline retention enemas. If this fails, surgical intervention can be performed.
Wound infections	INFECTION	Wound infections occur in 5-15% of patients following colorectal surgery and typically present around the fifth postoperative day and treated by opening of the overlying skin incision.
Myocardial infraction	MI	Acute myocardial infarction occurring during surgery or within 30 days after surgery.
Ileus	ILEUS	Ileus is simply defined as bowel obstruction. CT scan of the abdomen and pelvis has the sensitivity of 90-100% for diagnosis and evaluation of small bowel obstruction.
Abscess/Leak	ABSCESS	Extravasation of contrast material limited to the perianastomotic space often results in the development of an abscess, a pocket of infected fluid and pus. This is usually managed by insertion of a radiologically-guided percutaneous drainage catheter. Anastomotic leak varies depending on the level of anastomosis. Small bowel and ileocolic anastomoses have the lowest rates and coloanal anastomoses have the highest rates.

Sublanguage analysis

The hypothesis behind an IE system is the property of inequalities of likelihood in the sublanguage. We focus on two kinds of information: domain taxonomy and semantic lexicon where domain taxonomy here refers to report types while semantic lexicon refers to a collection of words/phrases. Specifically, we argue that clinical text for patients with surgical complications can have different distribution regarding report types. In addition, words and phrases for patients with surgical complications can also have different distributions compared to those with no complications. Identifying words/phrases with high inequality of likelihood can be used to assist the knowledge engineering process of developing an IE system or formulating the queries to identify positives (here, postsurgical complications).

Materials and Experimental Methods

Colorectal surgical cohort

The cohort considered here contains 1,856 colorectal surgical cases for 1,416 patients between 2005 and 2013 enrolled at Mayo Clinic Rochester. For this cohort, a quality improvement project has documented the postsurgical complications defined in Table 1. The cohort has been used as a retrospective data set for developing an IE system for identifying postsurgical complications. A collection of keyword patterns relevant to specific postsurgical complications has been assembled in the period of six months by subject matter experts. Table 2 lists the current keyword patterns. Since there can be multiple surgical cases per patient, to reduce patient-specific sublanguage, we limit to one surgical case per patient in this study. For each patient, we chose the case with the latest surgical date in case of multiple surgical cases for that patient.

MedTagger

MedTagger is a concept mention detection and normalization tool released open source through open health natural language processing (10, 11). It consists of three components: dictionary lookup allowing flexible mapping, machine learning-based concept mention detection, and pattern-based information extraction. In this study, we used MedTagger to identify phrases present in MedLex, a general semantic lexicon created for the clinical domain (12).

Table 2. Keyword patterns (as regular expressions) identified by subject matter experts. “\W” means punctuations, “\w+” means one or more letters, “\s+” means one or more blank spaces, and P? means the pattern P occurs zero or once.

Postsurgical Complication	**Keyword patterns**
Deep vein thrombosis (DVT)/pulmonary embolism (PE)	dvt; vein(\W|\s+)?thrombosis; venous(\W|\s+)?thrombosis; venous(\W|\s+)?thromboembolism; vte; vena(\W|\s+)?cava thrombosis; pe; pulmonary(\W|\s+)?embol(\w+)?; pulmonary embol(\w+)?; pulmonary(\W|\s+)?thromboembol(\w+)?
Bleeding	bleed(\w+)?; hemorrhage; acute blood loss; acute(\W|\s+)?anemia; acute blood loss anemia; post.?op(\w+)? anemia
Wound infections	wound(\W|\s+)?infection; cellulitis; contamination within the abdomen
Myocardial infraction	ami; attacks? coronary; attacks? heart; cardiac infarctions?; coronary attacks?; heart attacked; heart attacks?; heart infarctions?; infarctions? myocardial; infarctions? of heart; infarctions? heart; infarcts? myocardial; myocardial(\W|\s+)?infarcts?; myocardial(\W|\s+)?infarctions?; myocardial necrosis
Ileus	ileus; enteric tube; naso(\W|\s+)?gastric; naso(\W|\s+)?enteric; ng; ngt; ng(\W|\s+)?tube; small(\W|\s+)?bowel obstruction; sbo; partial small(\W|\s+)?bowel obstruction; psbo; poi
Abscess/Leak	Abscess; intra(\W|\s+)?abdominal infection; intra(\W|\s+)?abd infection; abdominal infection; leak; anastomotic leak; fistul(\w+)?

Data processing and analysis

We extracted various free text reports within 30 days from the surgical dates for the cohort. We lowercased all words and acquired words mentioned in the reports. We then applied MedTagger to obtain clinical concept phrases mentioned in the reports. We considered words and phrases as candidate concept keywords.

For each complication, we applied the following equations adapted from point-wise mutual information (http://en.wikipedia.org/wiki/Mutual_information) to assess the inequality of likelihood of report types and words/phrases:

$$Inequality(rpt, com) = log2((N(rpt, com) + 0.01)/N(com)) - log2(N(rpt)/N) \quad \text{(Eq 1)}$$

$$Inequality(con, com) = log2\left(N(con, com)\right) * (log2\left(\frac{N(con,com)+0.01}{N(com)}\right) - log2\left(\frac{N(con)}{N}\right)) \quad \text{(Eq 2)}$$

where N is the number of surgical cases, $N\ (rpt)$ is the number of cases having report type rpt, $N\ (con)$ is the number of cases having concept con, $N(rpt, com)$ is the number of surgical cases with complication com and report type rpt, $N(con, com)$ is the number of surgical cases with complication com and concept con, and $N(com)$ be the number of surgical cases with complication com. Note that in Eq 2, we penalized those concepts with low co-occurrence with the complications.

Results and discussion

We retrieved a total of 23,558 reports with an average of 16.6 reports per patient. After excluding report types with lower than 100 occurrences, we computed the inequality measures of the remaining report types. Table 3 shows the statistics of reports and postsurgical complications. Figure 1 shows the inequality measures of report types where close to 0 indicates no difference of the distribution of the cases with or without the specific complication (computed using Eq 1). It indicates that surgical cases with postsurgical complication generally yield more reports and for certain report types such as radiology and ECG, we observe over two fold increase.

For words/phrases, we filtered out those occurred in less than three patients and obtained a total of 21,910 unique words/phrases. Table 4 lists the top words/phrases ranked according to inequality measures computed using Eq 2). When comparing Table 2 and Table 4, some of the patterns identified by subject matter experts are also ranked top by sublanguage analysis. We underlined words/phrases captured by keyword patterns and also crafted by subject matter experts. Majority of the patterns identified in Table 2 do not appear in Table 4 and verse versa. One extreme case is INFECTION where none of the top ranked words/phrases can be found through keyword patterns. Another example is, *abscess*, a keyword for ABSCESS which appears in reports of 336 cases and only 20 are cases (with a rank of 345 for ABSCESS). Meanwhile, sublanguage analysis identified some keywords which can be a strong signal of complications. For example, *low hemoglobin* (32 out of 93 patients containing *low hemoglobin* in reports are cases) which ranked the third for BLEED most likely indicates bleeding. At the same time, the term, *bleeding*, appears in the reports of 408 patients but only 57 of them are bleeding cases (ranked 156 according to inequality). However, some of the keywords identified by sublanguage analysis may not have obvious semantic relationships with the postsurgical complications. For example, *knee* appears in five out of the seven cases of DVTPE with a total occurrence of 61. It ranks 13 for DVTPE regarding its inequality. But there is no obvious semantic relationship between *knee* and *DVTPE*. Further investigation is needed to find out hidden semantic relationships.

Table 3. Statistics of reports and postsurgical complications.

Report Type	**Definition**	**#Patients**	**ABSCESS**	**MI**	**BLEED**	**AFIB**	**ILEUS**	**DVTPE**	**INFECTION**
PRG	Progress notes	877	19	1	54	11	99	5	28
MIS	Misc. notes	887	20	3	56	11	87	7	38
SUM	Discharge summary	1404	23	5	78	17	136	7	46
OPN	Operation notes	1411	23	5	80	17	137	7	47
THP	Therapy	737	18	3	49	12	85	6	37
CON	Consultant notes	556	19	4	50	13	66	7	38
ADM	Admission	1003	21	4	59	15	110	5	32
AM	-	931	18	1	55	11	101	4	28
PP	Post procedure	801	16	1	51	10	92	3	25
SV	Subsequent visits	433	8	4	24	5	41	3	18
ECG	ECG reports	291	16	5	44	17	48	6	24
PAA	-	567	11	1	32	6	58	1	10
LIN	-	510	15	0	43	11	71	5	29
RB	-	418	8	1	17	4	46	0	6
RAD	Radiology reports	248	19	1	24	7	58	6	24
LE	Limited evaluation	140	4	2	20	5	16	1	10
SUP	Supplements	116	4	1	14	4	10	1	5
Total Cases		**1416**	**23**	**5**	**80**	**17**	**137**	**7**	**47**

One limitation of our sublanguage analysis is that we used an existing cohort with postsurgical complications identified. We have noticed the annotation of the cohort is quite noisy and some of the false postsurgical complications are actually true cases. The inequality metrics obtained here may not reflect the true inequality of the likelihood. However, we can use the words/phrases identified with high inequality of likelihood to bootstrap a better annotated data set for developing advanced informatics tools for postsurgical complication detection.

Conclusion

In this study, we have investigated the use of sublanguage analysis to facilitate NLP-enabled postsurgical complication detection. The study indicates that search-based approaches may miss some cases. The sublanguage analysis results can be used as a base to develop an information extraction system or support search-based NLP approaches by augmenting search queries. There are multiple future studies planned. One is to work with subject matter experts to improve the annotation quality of the cohort through bootstrapping. The other is to explore the use of machine learning approaches as well as sublanguage-supported search-based techniques for postsurgical complication detection.

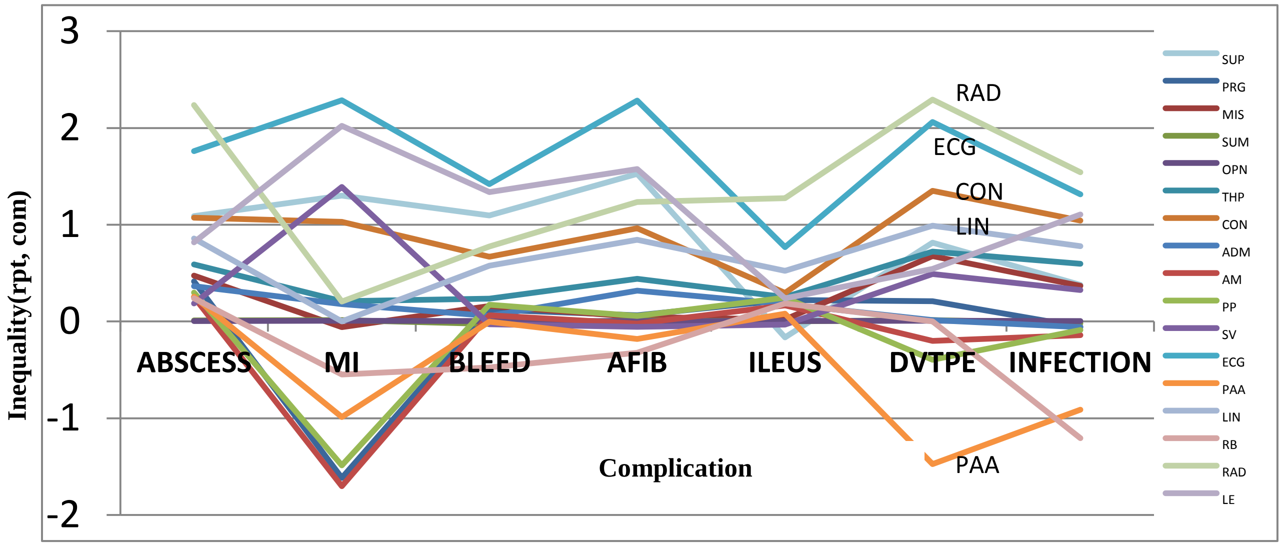

Figure 1. Inequality of report types with respect to complications

Table 4. Top 20 words/phrases identified by sublanguage analysis ranked by inequality.

Postsurgical Complication	**Words/Phrases**
Deep vein thrombosis (DVT)/pulmonary embolism (PE)	thrombosis; extremity edema; edema secondary; extremities bilateral; peripheral edema; wrap; negative fluid balance; left lower extremity; pe protocol; bowel ischemia; knee; popliteal vein; common femoral vein; fungal infection; sedative thromboembolism; left upper extremity; skin wound; pulmonary embolism; triglycerides
Bleeding	plasma fresh frozen; blood loss anemia; low hemoglobin; red blood cells; transfusion; transferred to icu; s hemoglobin; coagulation; gi bleed; intensive care; systolic blood pressure; hypotensive; spontaneous bacterial peritonitis; clot; hemodynamic; bilateral prophylactic mastectomy; extremity edema
Wound infections	open wound; dressing changes; wound packing; wet; vac; vacs; vacuum assisted closure; right internal jugular; bun creatinine; granulation tissue; wound care enteral nutrition; cva; keflex; wound status; granulation; fluid overload; on ventilator; wound edges; chronic pyelonephritis
Myocardial infraction	cardiac enzymes; cardiac catheterization; dominance; coronary angiography; ekg changes; cardiac monitor; st depression; ecg sinus rhythm; color flow Doppler; transthoracic echocardiogram; ischemic heart disease; coronary artery; troponin; acute myocardial infarction; coronary angiogram; metabolic acidosis; bilateral prophylactic mastectomy
Ileus	ileus; decompression; nasogastric tube; total parenteral nutrition; ng tube clamped; abdominal distention; npo; parenteral nutrition; reglan; small bowel dilatation; feels bloated; abdominal x ray; adynamic ileus; vomiting; emesis
Abscess/Leak	pressors; septic shock; resuscitated; resuscitation; sepsis; contamination; anasarca; arterial blood gas; hemodynamic instability; abdominal sepsis; mechanical ventilation; chronic pyelonephritis; hypotensive; fluid resuscitation; neo synephrine; norepinephrine; sonogram; vasopressors; vancomycin; positive pressure ventilation; pressure support

Acknowledgement

This work was made possible by joint funding from National Institute of Health grants R01LM009959A1, R01GM102282A1, and National Science Foundation grant ABI:0845523.

References

1. Leaper D, Whitaker I. Post-operative Complications: Oxford University Press; 2010.
2. Romano PS, Mull HJ, Rivard PE, et al. Validity of selected AHRQ patient safety indicators based on VA National Surgical Quality Improvement Program data. Health services research. 2009;44(1):182-204.
3. Birkmeyer JD, Shahian DM, Dimick JB, et al. Blueprint for a new American College of Surgeons: national surgical quality improvement program. Journal of the American College of Surgeons. 2008;207(5):777-82.
4. Fry DE, Pine M, Jones BL, Meimban RJ. Adverse outcomes in surgery: redefinition of postoperative complications. The American Journal of Surgery. 2009;197(4):479-84.
5. FitzHenry F, Murff HJ, Matheny ME, et al. Exploring the Frontier of Electronic Health Record Surveillance: The Case of Postoperative Complications. Medical care. 2013;51(6):509-16.
6. Murff HJ, FitzHenry F, Matheny ME, et al. Automated identification of postoperative complications within an electronic medical record using natural language processing. JAMA: the journal of the American Medical Association. 2011;306(8):848-55.
7. Alsara A, Warner DO, Li G, Herasevich V, Gajic O, Kor DJ. Derivation and validation of automated electronic search strategies to identify pertinent risk factors for postoperative acute lung injury. Mayo Clinic Proceedings; 2011: Elsevier; 2011. p. 382-8.
8. Sager N, Friedman C, Lyman MS. Medical language processing: computer management of narrative data. 1987.
9. Friedman C. A broad-coverage natural language processing system. Proceedings / AMIA Annual Symposium AMIA Symposium. 2000:270-4.
10. Torii M, Wagholikar K, Liu H. Using machine learning for concept extraction on clinical documents from multiple data sources. J Am Med Inform Assoc. 2011 Sep-Oct;18(5):580-7.
11. Liu H, Bielinski SJ, Sohn S, et al. An Information Extraction Framework for Cohort Identification Using Electronic Health Records. AMIA Summits Transl Sci Proc. 2013 Mar 18;2013:149-53.
12. Liu H WS, Li D, Jonnalagadda S, Sohn S, Wagholikar K, Haug PJ, Huff SM, Chute CG Towards a semantic lexicon for clinical natural language processing Annual Symposium of American Medical Informatics Association; 2012; Chicago; 2012.