Detecting Evidence of Intra-abdominal Surgical Site Infections from Radiology Reports Using Natural Language Processing

Alec B. Chapman, BM[1], Danielle L. Mowery, PhD[1,2], Douglas S. Swords, MD[3], Wendy. W. Chapman, PhD[1,2], Brian T. Bucher, MD[1,3]

[1]Biomedical Informatics, University of Utah School of Medicine, Salt Lake City, UT;

[2]IDEAS Center, George E. Wahlen Veterans Affairs Medical Center, Salt Lake City, UT;

[3]Pediatric Surgery, University of Utah School of Medicine, Salt Lake City, UT

Abstract

Free-text reports in electronic health records (EHRs) contain medically significant information - signs, symptoms, findings, diagnoses - recorded by clinicians during patient encounters. These reports contain rich clinical information which can be leveraged for surveillance of disease and occurrence of adverse events. In order to gain meaningful knowledge from these text reports to support surveillance efforts, information must first be converted into a structured, computable format. Traditional methods rely on manual review of charts, which can be costly and inefficient. Natural language processing (NLP) methods offer an efficient, alternative approach to extracting the information and can achieve a similar level of accuracy. We developed an NLP system to automatically identify mentions of surgical site infections in radiology reports and classify reports containing evidence of surgical site infections leveraging these mentions. We evaluated our system using a reference standard of reports annotated by domain experts, administrative data generated for each patient encounter, and a machine learning-based approach.

Introduction

Health care-associated infections (HAIs), such as surgical site infections (SSIs), affect one in every twenty hospitalized patients and account for $10 billion dollars in potentially preventable health care expenditures annually[1]. Identification and reporting of HAI after selected procedures is a required hospital quality measure by several federal agencies. In addition, detection of HAIs in a timely manner may alter treatment courses, reducing hospital costs and improving patient care.

Information indicating an HAI may be recorded in a variety of locations in the electronic health record (EHR): physician notes, radiology reports, and microbiology reports. One common method of documenting HAIs is using administrative data such as International Classification of Diseases (ICD). Due to the lack of specificity in administrative billing codes and timing of their assignment, these codes cannot reliably be used to identify HAIs in an accurate and timely manner. Furthermore, administrative data are often insufficient and incomplete due to the underutilization of structured data fields, lack of standardization, and unknown quality of clinical data[2]. Relevant information may instead be documented in free-text reports and text fields of the EHR. However, these reports are unstructured and the information contained within them must be extracted into a structured, computable format in order to utilize them. The traditional method of extraction is manual chart review, such as in the National Surgical Quality Improvement Program (SQIP). Specifically, these programs leverage trained surgical case reviewers to manually extract data from EHRs to detect and to report the development of HAIs[3,4]. However, manual abstraction is expensive, labor-intensive, and time-consuming. More efficient methods must be considered to extract this information.

Natural language processing (NLP) systems can be developed to automatically extract HAI information from textual data as an efficient and effective method of HAI detection. Melton et al. trained the NLP system MedLEE to identify adverse events (AEs) from discharge summaries[5]. MedLEE uses grammatical rules to identify concepts and their semantic context, which are then mapped to controlled vocabularies[6]. MedLee outperformed manual review for AEs, but their study included only one postoperative HAI (wound infection), which the system achieved a low precision (0.34) for detecting wound infections. Their overall recall (0.28) and precision (0.45) were fairly low, while their specificity was quite high (0.98). Penz et al. combined MedLEE and phrase-matching techniques that utilize regular

expressions to identify AEs related to the placement of central venous catheters[7]. By combining these two methods, they achieved good recall (0.72), specificity (0.80), and precision (0.64). False positives occurred due to indications of risk, in which the doctors discussed the hypothetical risks of a catheter placement. Some mentions were also missed due to misspellings in the text. The authors compared the NLP performance to that of ICD-9 and CPT codes. They concluded that NLP techniques are far more sensitive than using administrative data, which captured less than 11.5% of all central line placements.

Several studies have specifically focused on the effectiveness of NLP at detecting surgical AEs. FitzHenry et al. applied NLP methods to surveil for the presence multiple surgical complications, including pneumonia, wound infection, and urinary tract infections[8]. They achieved moderate to high performance for detecting each type of complication measured: pneumonia (recall: 0.90 and specificity: 0.80); wound infection (recall: 0.63 and specificity: 0.77); and urinary tract infections (recall: 0.80 and specificity: 0.95). Similarly, Murff et al. compared the performance of NLP techniques in the detection of surgical AEs to the performance of using patient safety indicators[9]. The NLP system performance ranged from moderate to high for detecting pneumonia (recall: 0.64 and specificity: 0.95) and sepsis (recall: 0.89 and specificity: 0.94). Their system had a higher recall and lower specificity than administrative data. Both of these studies were conducted in Veteran Affairs (VA) centers and used a gold standard of manual review performed by a VA-SQIP nurse.

The main limitation of these studies is the lack of additional clinically relevant information (such as anatomic location) in development of these detection systems. The goal of this study was to evaluate the performance of an NLP algorithm in the extraction of evidence of intra-abdominal surgical site infections that specifies the anatomic locations of these SSIs. The overall hypothesis is that *NLP methods will achieve a level of accuracy equal to or greater than that of manual chart review and greater than that of administrative data*. Our objective was (1) to develop a knowledge base that supports deep semantic extraction of evidence of intra-abdominal surgical site infection mentions and their anatomic locations and (2) to test a knowledge base-powered NLP algorithm's ability to make report-level predictions of whether evidence of an intra-abdominal surgical site infection is present. We evaluated our NLP algorithm's performance using a reference standard of expert-annotated text and compared its performance against other baseline approaches such as administrative data and machine learning.

Methods

The overall approach for this study included four steps: (1) identifying surgical patients and collecting their associated EHR data; (2) annotating text data and developing a knowledge base; (3) detecting mentions of SSIs and classifying reports for the presence of a SSI; and (4) evaluating automated methods against the reference standard annotations for SSIs. Institutional review board approval (IRB) was obtained.

Patient Identification and Dataset Extraction

We identified a cohort of patients undergoing gastrointestinal surgery from MIMIC-III Critical Care Database and collected their de-identified EHR data. Specifically, we collected administrative information (admission date, procedure date, ICD-9 diagnosis codes, and Current Procedural Terminology (CPT) codes). We limited the study to patients undergoing gastrointestinal surgery covering topics of esophageal, gastric, small intestine, large intestine, liver, pancreas, abdominal wall, ovarian, uterus, kidney procedures based on 28 CPT codes (e.g., CPT 49000=Under Incision Procedures on the Abdomen, Peritoneum, and Omentum). We limited our review and process to only include computed tomography (CT) reports due to their standardized structure and reliable reporting of findings. From this cohort, all the CT reports within 30 days after the surgical procedure were collected for our dataset.

Annotation Study

We conducted an annotation study using our development dataset to generate a reference standard of mention-level and report-level classifications using the full dataset. Two domain experts, a surgeon (author BB) and a surgical resident (author DS), annotated these reports using the annotation tool, the extensible Human Oracle Suite of Tools (eHOST)[10]. In these reports, evidence for a SSI is commonly described by phrases such as "fluid collection" or "abscess"[11]. To confirm the mentions (subsequently referred to as "fluid collections") were referencing findings related to the gastrointestinal tract surgery, we limited our search to detecting mentions co-occurring in the same sentence with descriptions of a gastrointestinal anatomic location such as "liver" or "abdomen". The annotators annotated sentences containing mentions of fluid collection as one of the following classes: *positive evidence of fluid collection*; *negated evidence of fluid collection*; *indication for exam*; and annotated mentions of *anatomic location* within these sentences marked as *positive evidence of fluid collection* (**Table 1**). After annotation was complete, we then randomly split our dataset into development set (n=565 documents for 409 patients) and test set (n=100 documents for 96 patients).

Table 1. Mention-level annotation schema; **Bolded**=terms indicating fluid collection; *italics*=contextual terms indicating anatomic location, negation, uncertainty, historicity, and indication.

Class Name	**Definition**	**Examples**
Positive Evidence of Fluid Collection	A positive mention of fluid collection mentioned as potentially occurring at some point in time in the same sentence as a relevant anatomic location.	"**Fluid collection** is seen in the *abdomen*."
Negated Evidence of Fluid Collection	A clear statement ruling out the possibility of fluid collection.	"There is *no evidence* of **fluid collection** in the *abdomen*."
Indication for Exam	A phrase indicating that the purpose of the exam is to check for fluid collection.	"*PURPOSE OF EXAM: Rule out* **abscess.**"
Anatomy	A part of the body in the GI tract that occurs in the same sentence as a mention of fluid collection.	"**Hematomas** are seen around in the *right lower quadrant*."

Mention Detection

To automatically identify and classify evidence of surgical site infection from each report according to our annotation schema, we developed an NLP system called SSI-Detect. We created a module called Fluid Collection Finder (fcFinder) that targets and classifies fluid collections*. fcFinder leverages the pyConText library, an adaptation of the ConText algorithm. ConText utilizes trigger terms and termination points to extract and associate contextual features such as negation, temporality and experiencer to targeted terms[12]. Specifically, pyConText extends the ConText algorithm by leveraging NetworkX digraphs to relate targets with modifiers and supports report-level classification[13]. pyConText has been adapted to support several use cases: to identify pulmonary embolisms, to classify cancer history, and to flag reports with significant carotid stenosis findings[13-15]. Using the development set, we created a knowledge base lexicon of 46 fluid collection targets and 128 anatomic location modifiers*. fcFinder utilizes this knowledge base to implement the pyConText algorithm and identify relationships between contextual and anatomical modifiers and fluid collection targets. After creating these relationships, fcFinder applies rules to classify each sentence containing a mention of a fluid collection finding as being *positive evidence of fluid collection*, *negated evidence of fluid collection*, or *indication for exam*. **Figure 1** depicts an example pyConText markup sentence with a target and its associated modifiers of negation and anatomic location.

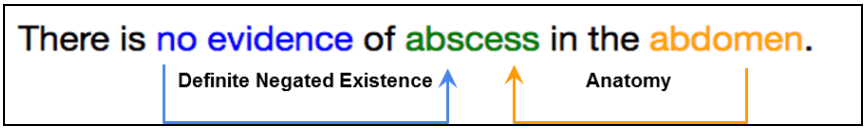

Figure 1. An example of *negated evidence of fluid collection*. The target, "abscess", is modified by "no evidence", a forward-direction negation modifier, and "abdomen", a bidirectional anatomic modifier.

Evaluation

We computed the performance of fcFinder by comparing the mention-level findings of positive evidence, negative evidence, and indication with the annotated reference standard. Definitive, historical, and probable evidence were all considered positive evidence mentions and could be matched with one another. Two findings were considered a match if their spans overlapped and they had the same class. We defined a true positive (TP) as a correctly identified mention, a false positive (FP) as a spuriously flagged mention, and a false negative (FN) as a missed mention. We evaluated fcFinder's performance overall and within each class using F1-score (**Eq. 1**), recall (**Eq. 2**), and precision (**Eq. 3**).

(**Eq. 1**) $F1 = \frac{2TP}{2TP + FP + FN}$ (**Eq. 2**) $Precision = \frac{TP}{TP + FP}$ (**Eq. 3**) $Recall = \frac{TP}{TP + FN}$

Report Classification

Using the mention-level annotations created by fcFinder, we classified a report as either *fluid collection-present* or *fluid collection-not present* using a python module called fcClassifier. If a report contained at least one positive mention-level annotation, the report was labeled as *fluid collection-present*. If it contained only negative or indication findings, or no findings at all, then it was labeled as *fluid collection-not present.* Additionally, we assessed fcClassifier's performance using three configurations: 1) without context; 2) without anatomy; and 3) with all modifiers. First, we implemented fcClassifier without context using only the targets without any linguistic (negation, historicity, indication) or semantic (anatomy) modifiers. This could essentially be considered a keyword search. We then tested fcClassifier using linguistic modifiers, but no semantic modifiers. This allowed us to evaluate the importance of excluding negated mentions of fluid collection and of requiring the use of anatomic location in order to accurately classify reports. Finally, we tested fcClassifier using both linguistic and semantic modifiers.

Baselines

We generated two baseline approaches for classifying each report comparison against fcClassifier: 1) ICD-9 codes and 2) n-grams. First, we developed a simple ICD-9 classifier that predicts *fluid collection-present* for each report from a patient encounter encoded with ICD-9 code 998.53. For any report from a patient without this ICD-9 code, it predicts *fluid collection-not present*. We then experimented with several machine learning models utilizing n-grams (1-4 window word features) from the development set, excluding words occurring in an English stopword list and word features occurring less than twice. To reduce the likelihood of overfitting, we applied the following feature selection strategy. We determined an upper bound threshold of n features to consider approximated using the square root of total training reports as 25 features. We ranked the informativeness of these 25 features by p-value using Chi-square with 5-fold cross-validation. We evaluated each classifier's performance at intervals of 5 features (5, 10, 15, 20, 25 features) and selected the highest performing classifier according to F1-score with the fewest number of features. We experimented using several machine learning classification models, including a naive bayes, random forest classifier, and linear support vector machine (SVM). For this study, we applied the most predictive classifier,

the linear SVM, leveraging the 15 highest-ranked word features to the test set. In **Table 2,** we compared these n-gram features to fcFinder's knowledge base and determined some terms are shared with our rule-based approach, giving us confidence this could be reasonable baseline approach.

Table 2. 15 top-ranked features for n-gram classifier. **Bolded=**shared terms with fcFinder's knowledge base.

Collection, fluid, **fluid collection**, **hematoma**, cm, drain, **collections**, drainage, **fluid collections**, **subcapsular**, pigtail, catheter, rim, noted, extravasation

Evaluation

We compared the performance of each baseline approach as well as fcClassifier and its various configurations against the testing set reference standard using accuracy (**Eq. 4**), F1-score, recall, and precision. We defined true positives (TP) as correctly classified reports as *fluid collection-present* and true negatives (TN) as correctly classified reports as *fluid collection-not present*. Specifically, we aimed to determine how well each approach could predict a report as *fluid collection-present* or *fluid collection-not present* in the test set. We assessed whether the differences in performance between the fcClassifier using all modifiers and all other approaches were statistically different using McNemar's test.

(**Eq. 4**) $Accuracy = \frac{TP + TN}{All\ reports}$

Results

Annotation Study

We calculated the inter-annotator agreement between annotators over 4 batches. There were a total of 249 mention-level annotations in the test set (**Table 3**). The most prevalent mention-level annotations were *positive evidence of fluid collection* (57%; 142 of 249 annotations) followed by *indication for exam* (32%; 80 of 240 annotations). The most prevalent report-level classifications were *fluid collection-not present* (69%). F1-scores were high and consistent for mention-level classes ranging from 0.88 to 0.89 and near excellent for report-level classification (0.96).

Table 3. Inter-annotator Agreement in Test Set using F1-score (surrogate for Cohen's Kappa).

Mention-level	**Number of Annotations (%)**	**F1-scores**
Overall	249 (100%)	**0.91**
Positive evidence of fluid collection	142 (57%)	0.89
Negated evidence of fluid collection	27 (11%)	0.88
Indication for exam	80 (32%)	0.88
Report-level		
Fluid collection status	100 (100%)	**0.96**

Mention Detection

We evaluated fcFinder's ability to identify and classify mentions of fluid collection from a reference standard using F1-score, recall, and precision (**Table 4**). Results were comparable between the development and test datasets (not shown). The algorithm performed with high F1-scores ranging from 0.87 to 0.90 for each category. Recall and precision was comparable (+/-2 points) for *overall* and for *evidence of fluid collection.* Recall was notably lower (-18 points) than precision for *indication for exam.* In contrast, recall was higher (+9 points) than precision for *negated evidence of fluid collection*.

Table 4. Mention-level comparison of fcFinder against the reference standard in the test test. **Bolded**=highest score for each metric.

	F1-score	Recall	Precision
Overall	0.90	0.89	0.91
Positive evidence of fluid collection	**0.91**	0.92	0.90
Negated evidence of fluid collection	**0.91**	**0.96**	0.87
Indication for exam	0.87	0.79	**0.97**

Report Classification

The prevalence of fluid collections in the development set (*fluid collection-present:* 42% and *not present:* 58%) was slightly skewed to positive cases compared to the test set (*fluid collection-present:* 31% and *not present:* 69%). We evaluated fcClassifier's ability to classify reports containing positive mentions of fluid collection compared to two baseline approaches: ICD-9 codes and n-grams (Table 4). ICD-9 codes achieved moderate performance (F1: 0.48; recall: 0.45). The use of targets without context for report classification resulted in improvement in classification (F1: 0.66; recall: 1.0). However, this improvement in recall results in a reduction of precision (0.49). fcClassifier without anatomic references improves precision (0.64) without sacrificing recall (1.0). fcClassifier with anatomic references performs with the highest F1 (0.88) and precision (0.82) with high recall (0.93). A comparison of each baseline to fcClassifier using McNemar's test suggests this improvement is not by chance alone.

Table 5. Report-level classification of fcClassifier against the test set reference standard. **Bolded=**best performance for each metric.

Classification Approach	Accuracy	F1-score	Recall	Precision	P-value vs. fcClassifier
ICD-9 codes	0.72	0.48	0.45	0.52	<0.0001
SVM n-grams	0.86	0.76	0.71	0.81	<0.0001
fcClassfier w/o context	0.70	0.66	**1.0**	0.49	<0.0001
fcClassifer w/o anatomy	0.88	0.79	**1.0**	0.64	0.0002
fcClassifier	**0.95**	**0.88**	0.93	**0.82**	--

Discussion

We developed an NLP algorithm called fcFinder to automatically detect fluid collections from radiology reports. Our system utilized lexical and semantic features to determine the context of mentions of fluid collections. Based on the mention-level findings, we then classified reports as either *fluid collection-present* or *not present*. Our system performed well at both the mention and report levels, outperforming all compared methods.

Error Analysis

We examined instances where incorrect mention level-annotations caused misclassification of reports. Most instances of misclassification were caused by terms that were missing in our lexicon. The errors made by the system can be grouped into the following categories:

- **False positive target:** Our lexicon included the literal "collectin", which was meant to match misspellings of "collection", but also matched "fluid collecting", which is not equivalent to a fluid collection. This resulted in an incorrect *positive evidence of fluid collection* annotation, which caused the document to be incorrectly classified as *fluid collection-present*.

- **Missing anatomic location modifier:** We did not have the literal "postoperative site" in our lexicon, resulting in a false classification of *fluid collection-not present*. We also deliberately excluded the literal "subcutaneous" as an anatomic modifier, due to its lack of specificity, but it was once used by the annotators, leading to another false negative classification.

- **Missing pseudoliteral:** Our lexicon included a number of "pseudoliteral" terms that would explicitly exclude phrases that could lead to false positives, such as "collection **of gas**" or "subpleural collection". One report was incorrectly classified because "collection of **intraluminal** gas" was not included in our lexicon of pseudoliterals. A similar error was caused by the phrase "pockets of gas".

- **Abbreviation:** A number of mistakes were made due to clinicians' use of a question mark, "?", as an abbreviation for indication for exam. This was not included in our lexicon and several phrases such as "?abscess" were not annotated as *indication for exam*. This is an example of one of the biggest challenges facing clinical NLP, which is the informal and inconsistent nature of clinical text.

- **Reference standard problem:** A small number of mismatches were due to mistakes made by the annotators in the reference standard.

Key Findings

Contextual Features

Overall, fcFinder performed well with each mention-level class: *positive evidence* (F1: 0.91), *negated evidence* (F1: 0.91), and *indication for exam* (F1: 0.87). This allowed fcClassifier to achieve high results with report classification (accuracy: 0.95; F1: 0.88). fcClassifier's performance demonstrates the importance of linguistic and semantic context. ICD-9 codes, which provide information at a patient level and do not consider any information that is contained within a report, performed much worse at report classification (F1: 0.48). Searching for target concepts without taking into account any context resulted in a perfect recall (1.0), but a very low precision (0.49). When utilizing linguistic and semantic context, fcClassifier maintained a high recall (0.93) while achieving a much higher precision (0.82). This high precision could enable a more accurate selection of reports for review, greatly reducing the amount of work needed to evaluate and detect surgical site infections. This demonstrates the value of including contextual features when extracting information from clinical reports.

Anatomic Location Modifiers

Our intention was to identify only reports that contained fluid collections in the site of recent gastrointestinal surgeries. To do this, we developed a knowledge base of 128 anatomic location modifiers. The use of these modifiers enabled us to exclude reports that contained mentions of other collections, such as pleural or sternal collections, but no gastrointestinal fluid collections. The importance of anatomical designation is shown by the decrease in overall classification performance (-9 points) when anatomic modifiers were excluded from the criteria for mention-level findings. The significant increase in both recall and precision using anatomic modifiers is a valuable contribution towards the effective and efficient identification of surgical site infections. This knowledge base should be expanded to other anatomic regions if utilized in other settings.

N-gram Classifier

The best-performing baseline model was the n-gram SVM classifier (F1: 0.76). Despite not using any contextual features, this model achieved fairly high accuracy (0.86) and precision (0.81), and reasonable F1 (0.76) and recall (0.71). While the SVM's precision was comparable to fcClassifier's (0.81 vs. 0.82), its recall was much lower (0.71 vs. 0.93). Given the top 15 word features used by the n-gram model, we can conclude that the model does not take into account any linguistic (negation, historicity) or semantic (anatomic locations) context. However, the presence of procedural terms (“drainage”, “catheter”, “drain”) suggests the model includes situational context from the narrative, namely the treatment of fluid collection, which might be equally useful for asserting the presence of fluid collection. Future work could combine situational with linguistic and semantic context. Additionally, a system that utilizes structured data, such as ICD-9 codes, in combination with NLP could achieve even better results than those we report here.

Value of Rule-based Methods

This study demonstrates the potential of rule-based NLP to offer a semantically rich and detailed narrative of patient care. Techniques such as administrative data and machine learning offer very little detail beyond binary classification of a report. Rule-based NLP methods, such as those utilized by fcFinder, can extract much more detailed information. Here we showed the value of leveraging both linguistic and anatomic location modifiers to restrict SSI findings to a particular anatomic region. Similar methods can also model the administration and effectiveness of treatment in conjunction with the change of severity and course of a disease over time, without requiring much extra effort in developing new models to address each new question. This can offer a much more detailed and specific review of patient care, and can also enable a generalized approach to evaluating quality of treatment.

Comparison to Other Works

By conventional standards, our system achieved acceptable results. No other study has specifically aimed to extract mentions of fluid collection as evidence of SSI, but previously studies have successfully extracted other evidence of SSIs. Melton et al.'s algorithm, which utilized MedLEE, achieved a detected postoperative wound infection with a precision of 0.34[5]. The algorithm implemented by FitzHenry et al. to detect wound infection achieved a recall of 0.63 and specificity of 0.77[8]. Murff et al. report a recall of 0.89 and specificity of 0.94 in the detection of sepsis[9]. We achieved a recall of 0.92 and precision of 0.90 in the detection of positive mentions of fluid collection. Our algorithm introduces the novel technique of connecting targets with anatomic location modifiers. This technique contributed to our overall performance and was a unique method of using NLP to detect a surgical AE. While some studies have evaluated the extraction of anatomic site of a condition, to our knowledge, no studies have done so in connection to surgical site infections or other surgical adverse events[16-17].

Limitations and Future Work

There are several limitations to our study. We used reports from the MIMIC-III Critical Care Database. These reports were written for surgical patients in intensive care. Because of the severity the patient’s condition, the prevalence of surgical site infections is higher than what we would expect to see in the general population. Reports

were then more likely to contain mentions of our targets, fluid collection. This may not directly generalize to the general population, where we would expect the prevalence of surgical site infection to be ~8-10% of patients. In this testing sample, 31% of notes were found to have *fluid collections-present*. However, the MIMIC-III dataset is openly available and the work that we have done with it can be reproduced and compared against. Furthermore, in the future, we will apply these methods to a more general population which could more broadly validate the results of this study.

Another limitation is the focused scope of this project. We looked only at radiology reports, specifically including only CT reports because of their relatively consistent structure and content. We focused on one specific anatomic region, and our targets included a relatively small lexicon of a single clinical concept, fluid collection, as evidence of surgical site infection. The CDC lists three additional criteria for defining an organ/space surgical site infection: purulent drainage from a drain, organisms found in a culture, and diagnosis of a surgical site infection by a surgeon or attending physician[18].

Despite the relatively narrow focus of this project, it has proven the utility of a number of system features that can be applied in future studies and future implementations of SSI-Detect. In particular, this study has shown how a knowledge base of anatomic modifiers combined with NLP contextual tools results in a much higher precision and much richer semantic detail. This study is the first step in a larger framework of applying NLP methods to identify surgical adverse events. Future work should focus on expanding the knowledge base which we developed for this study, improving the performance of our system on non-enriched text, conducting a data analysis across various report types to test the generalizability of our methods, and expanding these methods to address other anatomic regions and clinical problems.

Conclusions

The automated detection of adverse surgical event has the potential to revolutionize how providers and hospitals detect and report quality measures. Currently, these efforts rely on manual chart review which limit the scalability and generalizability of quality measurement activities. In addition, leveraging automated detection into clinical decision support services may lead to better surveillance and the potential to improve patient care. We executed automated detection of evidence of intra-abdominal surgical site infections using a rule-based NLP system that achieved accuracy similar to manual chart review. In addition, this system outperformed other approaches using administrative data or SVM machine learning techniques. Future work in this field will seek to broaden our NLP approach and leverage addition types of structured and unstructured healthcare data to the detection of postoperative adverse events.

Acknowledgements

We thank our anonymous reviewers for their invaluable feedback on this manuscript.

* The source code and lexicon for fcFinder is openly available at https://github.com/abchapman93/fcFinder.

References

1. Zimlichman E, Henderson D, Tamir O, et al. Health care-associated infections: a meta-analysis of costs and financial impact on the US health care system. JAMA Intern Med. 2013 Dec 9-23;173(22):2039-2046. doi: 10.1001/jamainternmed.2013.9763.
2. Etzioni DA, Lessow CL, Lucas HD, et al. Infectious Surgical Complications are Not Dichotomous: Characterizing Discordance Between Administrative Data and Registry Data. Ann Surg. 2016 Oct 17. doi: 10.1097/SLA.0000000000002041.
3. Ko CY, Hall BL, Hart AJ, Cohen ME, Hoyt DB. The American College of Surgeons National Surgical Quality Improvement Program: achieving better and safer surgery. Jt Comm J Qual Patient Saf. 2015;41(5):199-204. PubMed PMID: 25977246.
4. Price CS, Savitz LA. Improving the measurement of surgical site infection risk stratification/outcome detection. Final report (Prepared by Denver Health and its partners under Contract No. 290-2006-00-20). AHRQ Publication No. 12-0046-EF. Rockville, MD: Agency for Healthcare Research and Quality; 2012.

5. Melton GB, Hripcsak G. Automated detection of adverse events using natural language processing of discharge summaries. Journal of the American Medical Informatics Association : JAMIA. 12(4):448–57.
6. Friedman C, Alderson PO, Austin JH, Cimino JJ, Johnson SB. A general natural-language text processor for clinical radiology. J Am Med Inform Assoc. 1994 Apr;1(2):161–74.
7. Penz JFE, Wilcox AB, Hurdle JF. Automated identification of adverse events related to central venous catheters. Journal of Biomedical Informatics. 2007;40(2):174–182.
8. FitzHenry F, Murff HJ, Matheny ME, Gentry N, Fielstein EM, Brown SH, et al. Exploring the frontier of electronic health record surveillance: the case of postoperative complications. Medical care. 2013;51(6):509–16.
9. Murff HJ, FitzHenry F, Matheny ME, Gentry N, Kotter KL, Crimin K, et al. Automated identification of postoperative complications within an electronic medical record using natural language processing. JAMA : the Journal of the American Medical Association.
10. South BR, Shen S, Leng J, Forbush TB, DuVall SL, Chapman WW. 2012. A prototype tool set to support machine-assisted annotation. In Proceedings of the 2012 Workshop on Biomedical Natural Language
11. Bucher B, Mowery DL, Castine M, Chapman WW. Towards Automatic detection of surgical site infections. 2017 Joint Summits on Translational Science. San Francisco, CA.
12. Chapman WW, Chu D, Dowling JN. ConText: An algorithm for identifying contextual features from clinical text. In: Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing. Stroudsburg, PA, USA: Association for Computational Linguistics; 2007. p. 81–88. (BioNLP '07).
13. Chapman BE, Lee S, Kang HP, Chapman WW. Document-level classification of CT pulmonary angiography reports based on an extension of the ConText algorithm. Journal of Biomedical Informatics. 2011;44(5):728–737.
14. Wilson RA, Chapman WW, DeFries SJ, Becich MJ, Chapman BE. Automated ancillary cancer history classification for mesothelioma patients from free-text clinical reports. Journal of Pathology Informatics. 2010 Jan 1;1(1):24.
15. Mowery DL, Chapman BE, Conway M, South BR, Madden E, Keyhani S, et al. Extracting a stroke phenotype risk factor from Veteran Health Administration clinical reports: an information content analysis. J Biomed Semantics. 2016 May 10;7.
16. Dligach D, Bethard S, Becker L, Miller T, Savova GK. Discovering body site and severity modifiers in clinical texts. J Am Med Inform Assoc. 2014 May 1;21(3):448–54.
17. Pham A-D, Névéol A, Lavergne T, Yasunaga D, Clément O, Meyer G, et al. Natural language processing of radiology reports for the detection of thromboembolic diseases and clinically relevant incidental findings. BMC Bioinformatics. 2014;15:266.
18. Mangram AJ, Horan TC, Pearson ML, Silver LC, Jarvis WR. Guideline for prevention of surgical site infection, 1999. Hospital Infection Control Practices Advisory Committee. Infect Control Hosp Epidemiol. 1999 Apr;20(4):250-278-280.