Published in final edited form as:

Med Care. 2013 June ; 51(6): 509–516. doi:10.1097/MLR.0b013e31828d1210.

Exploring the Frontier of Electronic Health Record Surveillance: The Case of Post-Operative Complications

Fern FitzHenry^{1,3}, Harvey J. Murff^{1,2,5}, Michael E. Matheny^{1,2,3,4,5}, Nancy Gentry¹, Elliot M. Fielstein^{1,3}, Steven H Brown^{1,2,3}, Ruth M Reeves^{1,3}, Dominik Aronsky^{1,3}, Peter L. Elkin⁶, Vincent P. Messina¹, and Theodore Speroff^{1,2,4,5}

¹Tennessee Valley Healthcare System, Veterans Affairs Medical Center, Nashville, TN

²Division of General Internal Medicine and Public Health, Vanderbilt University School of Medicine, Nashville, TN

³Department of Biomedical Informatics; Vanderbilt University, Nashville, TN

⁴Department of Biostatistics, Vanderbilt University, Nashville, TN

⁵Vanderbilt Institute for Medicine and Public Health, Nashville, TN

⁶Mount Sinai Medical Center, New York, NY

Abstract

Background—The aim of this study was to build electronic algorithms using a combination of structured data and natural language processing (NLP) of text notes for potential safety surveillance of nine post-operative complications.

Methods—Post-operative complications from six medical centers in the Southeastern United States were obtained from the Veterans Affairs Surgical Quality Improvement Program (VASQIP) registry. Development and test datasets were constructed using stratification by facility and date of procedure for patients with and without complication. Algorithms were developed from VASQIP outcome definitions using NLP coded concepts, regular expressions, and structured data. The VASQIP nurse reviewer served as the reference standard for evaluating sensitivity and specificity. The algorithms were designed in the development and evaluated in the test dataset.

Results—Sensitivity and specificity in the test set were 85% and 92% for acute renal failure, 80% and 93% for sepsis, 56% and 94% for deep vein thrombosis, 80% and 97% for pulmonary embolism, 88% and 89% for acute myocardial infarction, 88% and 92% for cardiac arrest, 80%

Corresponding author. Theodore Speroff, Ph.D., Professor, Center for Health Services Research, GRECC/HSR 4th Floor, VA Tennessee Valley Healthcare System, 1310 24th Avenue South, Nashville, TN 37212, Phone (615) 873-6057 (VA office), Cell (615) 636-9476, Fax (615) 873-7981, ted.speroff@va.gov.

Conflict of Interest Disclosures: All authors will complete and submit the ICMJE Form for Disclosure of Potential Conflicts of Interests. There are no conflicts of interest during the past three years to report for any of the contributing authors except Peter Elkin MD. Dr. Elkin declares financial interest in the natural language processing software with the prospect of potential future royalties.

Disclaimer: The views expressed in this article are those of the authors and do not necessarily reflect the position or policy of the Department of Veterans Affairs or the United States government.

Additional Contributions: The authors would like to acknowledge The VA Surgical Quality Data Use Group (SQDUG) for its role as scientific advisors and for the critical review of data use and analysis presented in this manuscript. We would like to thank the collaborative investigators for their work in this study: Debra Jo Barrett MSN, Lexington VA Medical Center, Lexington, KY; William G. Cheadle MD, Louisville VA Medical Center, Louisville, KY; Brad L. Roper PhD, Memphis VA Medical Center, Memphis, TN; Teresa England RN, James H. Quillen VA Medical Center, Mountain Home, TN; Sandra E. Shaw RN, Huntington VA Medical Center, Huntington, WV. Robert Alan Greevy, Jr, PhD, Vanderbilt University, Zhao Zou, MS, Vanderbilt University, Robert Dittus MD MPH, VA Tennessee Valley Healthcare System, William Nylander MD, VA Tennessee Valley Healthcare System (TVHS), Nashville, TN and the members of the POEM Team, TVHS.

and 90% for pneumonia, 95% and 80% for urinary tract infection, and 80% and 93% for wound infection, respectively. A third of the complications occurred outside of the hospital setting.

Conclusions—Computer algorithms on data extracted from the electronic health record produced respectable sensitivity and specificity across a large sample of patients seen in six different medical centers. This study demonstrates the utility of combining natural language processing with structured data for mining the information contained within the electronic health record.

Keywords

surgical complications; adverse events; patient safety; electronic surveillance; natural language processing; text mining

Introduction

Improving healthcare quality and patient safety has been a major priority for the past fifteen years since the publication of the Institute of Medicine's report on iatrogenic injury.¹ Surgical complications comprise more than half of the healthcare adverse events and most are preventable injuries that occur to patients as a result of their medical management and not their underlying disease process.^{1, 2} Variation in quality of surgical care among multiple facilities prompted the Veterans Affairs (VA) to initiate the first national surveillance program for measuring complications and instituting a structure for systematic improvement in quality of surgical care.³ This commitment of a health system to deliver the safest medical care possible for its patients requires the ability to detect and record complications.^{4, 5} Hence, a key challenge for improving patient safety is an accurate and widely implementable means of identifying the occurrence of complications.

The monitoring of surgical complications for enhancing quality and patient safety starts with data collection, but provider self-report grossly underreports safety issues,^{6, 7} manual chart review is expensive and requires an enormous commitment of healthcare resources,^{8–11} and administrative data using ICD-9-CM codes¹² designed for billing are limited.¹³ An innovative and promising alternative for monitoring patient safety is the application of more sophisticated search strategies that incorporate natural language processing to “mine” clinical documents to detect complications.¹⁴

Prior studies extracting information from text have made only limited use of the full electronic health record with most focusing on only a single type of document such as discharge summary, radiology report, imaging report, or diagnostic test.^{15–22} Few studies exist that have processed large samples of documents across multiple institutions,^{8, 10} and none have included the complete repository of clinical notes across settings. If effective, computerized extraction of information from the total health record, including both structured data (such as clinical laboratory values and vital signs) and textual notes, would make an attractive means for routine surveillance of complications. Because the VA has an integrated electronic health record system, requiring all outpatient and inpatient documentation to be electronically stored, the VA is a suitable environment for demonstrating the potential strength of automated data extraction from the electronic health record (EHR).

In this study, we created algorithms based on both structured data and text codified by natural language processing software to define surgical complications reported in the VA Surgical Quality Improvement Program (VASQIP), a prospective surgical outcome surveillance program for comparing centers on risk adjusted outcomes and quality improvement.^{3, 23} The goal of the study was to develop a Post-Operative Event Monitor

(called POEM) to detect surgical complications using the full EHR across six hospitals. We compare the performance of POEM with VASQIP's review process consisting of manual nurse chart review as the reference standard for defining and coding occurrence of surgical complications.

Methods

Sample

The study cohort included 45,159 surgical procedures on 33,565 patients from 1999 to 2006 at six VA medical centers in the Southeastern United States which comprise our **Veterans Integrated System Network (VISN)**. The VASQIP protocol limited chart review to major non-cardiac surgical procedures.³ For each VASQIP surgical procedure, trained nurse reviewers use a uniform data collection process to abstract the health records of patients to detect and record 30-day post-operative complications. An operation is defined as outpatient if the length of stay in the facility is less than 24 hours. Cases with multiple surgical procedures within a 30-day period are pooled together to attribute patient post-surgical adverse events to the index surgery. Of the 21 potential post-operative complications abstracted by VASQIP, we targeted nine adverse events, including those with higher incidence of occurrence and those used by the Agency for Healthcare Research and Quality as patient safety indicators^{23–25} that electronically detected complications using ICD-9 codes. The patient safety indicator complications included sepsis, deep vein thrombosis, pulmonary embolism, myocardial infarction, and pneumonia and the additional complications included acute renal failure, cardiac arrest, urinary tract infection, and wound infection. VASQIP nurse inter-rater agreement on post-operative occurrences averages 95% and the kappa coefficient has been estimated at 0.67 for superficial incisional surgical site infections (SSI), 0.73 for deep incisional SSI, 0.57 for wound disruption, 0.82 for cardiac arrest, 0.46 for sepsis, 0.70 for urinary tract infection, 0.73 for acute renal failure, 0.65 for pneumonia, 0.60 for myocardial infarction, 0.81 for deep vein thrombosis and 0.89 for pulmonary embolism.^{3, 26} The local Institutional Review Board in each of the six participating VA facilities approved the study.

In this study, we retained all of the 4,093 surgical procedures identified by VASQIP with one or more of the nine complications. We randomly selected ten percent of the remaining VASQIP surgical procedures in the cohort with no complications, stratified by medical center and date of procedure. The total study sample included 8,186 surgical procedures in 7,743 patients. The study sample was then divided at random stratified by type of complication into a development set ($n = 4,098$ surgical procedures) and a test set ($n = 4,088$ surgical procedures) which over weighted the occurrence of surgical complications relative to the incidence in the VA population.

Structured data included the surgical complication outcomes obtained from the VASQIP program and demographics, vital signs, clinical laboratory, and pharmacy documentation in free text notes occurring 15 days preceding date of surgery to rule out pre-existing conditions and 30 days following the date of the surgical procedure to detect post-operative complications. Types of notes were classified by the "note title" of the electronic template. The 291,665 notes processed for inpatient and same day surgery patients included 17,960 imaging, 6,560 discharge summary, 6,867 operative, and 229,885 clinical progress notes (some operative notes were documented using a progress note template). The number of notes processed averaged 107 per case. These narrative clinical notes were parsed electronically for medical terminology and mapped to SNOMED CT concepts using the Multi-threaded Clinical Vocabulary Server Natural Language Processing program.²⁷ SNOMED CT is considered the medical terminology with the greatest coverage of medical concepts.²⁸ Concepts in SNOMED CT are hierarchically organized, accounting for

synonymy such that the terms heart attack and myocardial infarction are coded to the same concept. The full study database was comprised of the structured data, codified SNOMED CT concepts, and regular expressions (word phrases) for each surgical procedure.

Development of Algorithms

Clinical algorithms to identify post-operative complications were based on VASQIP defined criteria.^{29–31} Clinical teams comprising general internal medicine and emergency medicine physicians, a surgeon, nurses, VASQIP nurse, and informaticians operationalized the VASQIP definitions into rules that were programmed using SQL database queries (see supplemental Appendix for details on definitions and algorithms). Rule components searching for indexed concepts from the SNOMED-CT vocabulary uniformly included the “children” of a hierarchical parent concept after the “children” concepts were reviewed to eliminate ones that were pediatric, perinatal, or non-postoperative manifestations of the conditions (e.g. tropical infections, black lung, etc.). For example, the rule for acute renal failure included the SNOMED-CT parent concept of “dialysis”, whose descendent concepts, “dependency on renal dialysis”, “finding of measures of dialysis”, “finding of adequacy of dialysis”, and “adequacy of hemodialysis” were also captured by the rule.^{3, 14} For wound infection, the algorithm included the SNOMED-CT parent concept “wound” but excluded SNOMED-CT children concepts such as “bite” wound, “sting” wound, and “insertion” wound because they do not characterize surgical wounds. These rule-based algorithms were then applied to the development database to detect the instances of the nine complications (wound infections were combined into a single complication). The complications identified by POEM were compared to the reference standard to calculate sensitivity and specificity. Developing the algorithms for each of the nine adverse events was an iterative process conducted on the development set attempting to optimize both sensitivity and specificity. Sensitivity was defined as the proportion of VASQIP identified events that were identified by POEM. Specificity was defined as the proportion of procedures without a VASQIP identified event that were not flagged by the corresponding POEM rules. The final rule maximized sensitivity without a loss in specificity, procedures on use of ROC curves and rule combinations are detailed in our prior publication.²⁵

Analyses

The unit of analysis was the surgical procedure. POEM was run on extracted data from each surgical procedure in the test set to evaluate the performance of the final algorithm. For each complication, patients designated by VASQIP having the adverse event were cases and all patients in the dataset without the specific complication were controls. Statistics included the sensitivity, specificity, and positive predictive value for each of the nine surgical complications and the 95% confidence interval using the Wilson Score method in R.³² Descriptive statistics were performed using SAS version 9.2 (Cary, NC).

RESULTS

The development and test set patients were similar in demographics and clinical characteristics (Table 1). Patients had a mean age of 63 years and 95% were male, which was typical for the VA population (Table 1). Seventy-three percent of the surgical procedures were inpatient. Seventeen percent had an American Society of Anesthesiologist (ASA) pre-operative score indicating severe systemic disease.³³ The sample was distributed across the six VA facilities in the same proportion as the surgeries in the VASQIP database, reflecting the size and surgical volume of the medical centers.

The performance of POEM in the development and test sets is shown in Table 2. The performance in the test set was similar to the sensitivity and specificity in the development

set, indicating that the rule development was not over fitted. The sensitivity for the final rule across the 9 complications ranged from a low of 0.56 for DVT to a high of 0.95 for urinary tract infection (see Table 2). Specificity ranged from a low of 0.63 for wound infection to a high of 0.97 for pulmonary embolism. Considering sensitivity and specificity concurrently, POEM performed best in detecting cardiac arrest at a sensitivity of 0.88 (95% CI = 0.83, 0.93) and specificity of 0.92 (95% CI = 0.92, 0.93) and worse for wound infection, sensitivity of 0.77 and specificity of 0.63.

The positive predictive value for each complication ranged from 0.15 for DVT to 0.58 for pneumonia (Table 3). Because positive predictive value is affected by prevalence, we also report the positive and negative likelihood ratios to express the ability of an algorithm to discriminate between cases and controls.³⁴ A likelihood ratio of 1.0 indicates an equivocal or indeterminate test that conveys no meaningful information. Positive likelihood ratio values greater than one characterize the strength of the algorithm to detect the complication and negative likelihood ratio values less than one indicates that the result is associated with the absence of the complication. For example, a likelihood ratio of 10 for acute renal failure (Table 3) indicates that patients with the surgical complication are 10 times as likely to test positive by POEM as those without the complication. The positive likelihood ratio ranged from 2.1 for wound infection to 27 for pulmonary embolism. The negative likelihood ratio ranged from 0.46 for DVT to 0.07 for urinary tract infection.

One advantage of using the full EHR is the inclusion of documents for outpatient visits. The rate of surgical complications recorded during hospitalization, post-discharge and after same-day surgery is shown in Table 4. The sensitivity performance of POEM remained fairly consistent across settings, except for the detection of deep vein thrombosis. In general, Table 4 shows that the algorithms identified post-operative events that occurred both after hospital discharge in the case of inpatient surgeries and after outpatient surgical procedures.

Discussion

This study demonstrates the feasibility of automated detection and surveillance applications that combine structured data and text-mined EHR data. In all nine classes of complications, the likelihood ratio showed considerable gain in the value of information for detecting complications compared to related prevalence rates. There was, however, substantial variation among the different post-operative complications as the 0.56 sensitivity for DVT was 20 points lower than the 0.77 for wound infection and 39 points lower than the 0.95 sensitivity for urinary tract infection. A review of POEM errors revealed that VASQIP nurses mistakenly classified some arterial thrombosis, generally occurring after vascular surgical procedures such as bypass, and surgical site hematomas as DVT. Therefore, some true positive cases as judged by the VASQIP review in our analysis were likely negative and adjusting for reference standard misclassification might improve the POEM measures of sensitivity and specificity. These results and the kappa coefficients of inter-rater agreement reinforce that, despite annual certification and experience, the nurses were indeed a reference and not a gold standard for our evaluation. The consequence is that the POEM findings reported in our paper are conservative. Interestingly, POEM detected a substantial number of complications during the post-discharge period. These findings reveal the ability of systems like POEM to use records that extend across settings to capture information on the full 30-day post-discharge episode of care, a period when many complications occur.

Hanauer and colleagues³⁵ used a customized keyword search of only dictated clinical notes to obtain a sensitivity of 1.0 and specificity of 0.93 for post-surgical myocardial infarction and a sensitivity of 0.93 and specificity of 0.96 for post-surgical pulmonary embolism. The sensitivities in the Hanauer study were higher than the 0.88 for myocardial infarction and

0.80 for pulmonary embolism using POEM. String searches as applied in the Hanauer study required exact matching and a high degree of specification within the particular dataset to obtain such performance characteristics. The Hanauer keyword approach hand tailored over 4,000 strings of keywords, including 1,741 to exclude negative terms leading to false positives. This approach required on average 1.5 customized negative terms for every false positive alert eliminated. The keyword approach appears to provide high accuracy but it is highly customized and does not represent a generalizable approach that reaches across the vast domain of medical concepts and across wide regional variations in style of documentation. Unlike the keyword approach, most NLP tools have an advantage of sharing a unified, large, and sophisticated medical vocabulary for knowledge representation of extracted medical concepts. NLP rules leverage the hierarchical relations of higher-order “parent” concepts and underlying “child” concepts contained in the standardized SNOMED CT medical terminology and common acronyms and abbreviations have been incorporated into the NLP vocabulary. In addition, NLP identifies positive, negative, and uncertain assertions in text, breaking up text into sentences and phrases that can be linked as components of rules (e.g. two concepts must appear in the same sentence) or provide context for interpretation.

In a similar vein, Elkin and colleagues³⁶ were able to achieve a sensitivity of 1.0 and a specificity of 0.98 in detecting pneumonia in imaging reports. NLP applications conducted on a single type of note from a single institution for a single purpose generally perform quite well.^{36–38} In our study, only 33% of the pneumonia cases had a text-based imaging report and, in VASQIP, pneumonia may be diagnosed on clinical or culture findings without confirmatory imaging. This suggests that generalizable, real world applications of NLP may require the processing of an entire clinical record in order to provide inclusive population surveillance.

The sheer volume and diversity of documents both across providers within a medical center and between medical centers poses a substantial challenge to information exchange.³⁹ In their report on a computer-assisted algorithm to identify patients with surgical site infection, Price and Savitz¹¹ found a drop in sensitivity from an initial value of 0.93 to 0.38 when the algorithm which relied on structured data only was spread and findings pooled across additional medical centers. They concluded that further improvements may be difficult without information extraction from free-text notes and there could be value in exploring natural language programming. In our study, we applied NLP to thousands of patient records documenting over eight thousand surgical encounters. Even within an integrated healthcare system such as the VA, we encountered wide heterogeneity in notes varying according to type of note, clinical provider, discipline, and practice patterns in style of documentation and individual preferences for formatting and structuring notes, local norms by medical centers in approval of forms, and inpatient and outpatient documentation. A key outcome of our study was demonstrating generalizability and economy of scale using general all-purpose NLP tools across this heterogeneity of documentation.

Another important finding of our study was that almost a third of targeted post-operative complications occurred either after hospital discharge or after an outpatient surgical procedure. In 400 consecutively discharged patients from a general medical service, Forster et al. found that 19% of the patients experienced a post-discharge complication.⁴⁰ Hence, patient safety programs using the patient safety indicators that depend on administrative hospital coding of ICD-9 discharge diagnoses would consequently underestimate the overall number of adverse events due to the inability to capture outpatient and post-discharge complications.⁴¹ As such the patient harm and cost to the healthcare system related to complications could be substantially higher than previously estimated.

Post-surgical adverse events result in higher morbidity, mortality, and costs.^{23, 31, 42, 43} The implementation of VASQIP in VA hospitals has resulted in a 45% reduction in morbidity and 27% reduction in mortality.⁴⁴ However, the program requires manual chart abstraction costing several million dollars a year.³ A trained VASQIP nurse abstractor can cover approximately 1,600 charts per year, but labor consumes 85% of the cost of the program and fiscal constraint limits the coverage to just a proportion of surgeries. If automated tools can be built to interface with the electronic health record and support the full continuum of an episode of care, then the burden for data collection can be reduced and the coverage for surveillance and quality improvement can be extended to the total patient population.^{14, 45} A first step toward such electronic surveillance is building automated algorithm tools like POEM that use structured data and natural language processing to screen the population of records.

Our study was conducted within the integrated healthcare system of the VA. A challenge for NLP is the documentation practices of clinicians and the forms and templates used for documentation not suitable for NLP processing. Notes often contain short-hand abbreviations like “r” for the word “are” and context specific acronyms like “MI” for the state of Michigan versus myocardial infarction³⁵ and “CPRS” for the computer interface to the VA electronic health record versus cardio-pulmonary resuscitation. Templated notes often contain leading questions for the clinician to answer or check which are difficult to separate from the clinician’s entry of data and consequently confound the correct interpretation using NLP. Some reports, such as electrocardiograms, are often not in suitable format for text processing⁴⁶ and many are images scanned into the record that cannot be directly text processed. A multitude of such factors contributing to problems in quality of documentation in addition to events not captured or recorded in documentation also contribute to errors in NLP. These factors will require modification to either documentation practices or to enhancing NLP to recognize and interpret more contexts correctly in general purpose approaches such as the one applied in this study.

The institutionalization of the VA electronic health record provides one of the more extensive computerized environments matched by only a few non-VA settings. Although the VA setting limits the generalizability of this study, it does not diminish the impact on the potential meaningful use of the electronic record. The use of all electronic data in this study, both structured and narrative text, is a harbinger for the future development of informatics in healthcare.

Despite these limitations, our study demonstrated a robust clinical use for the application of NLP technology. Advances in applied informatics and further refinements to the POEM electronic chart review may enhance the tool’s reliability and reproducibility for detecting pre-specified outcomes; eliminate costly and time-consuming chart reviewer training; and demonstrate scalability and portability to a broad range of clinical conditions. Further enhancements to our NLP tool development include a machine learning tool for automated chart abstraction and annotation,⁴⁷ NLP extraction of event timing capable of distinguishing between pre-operative conditions and post-operative complications,⁴⁸ and near-real time electronic extraction systems that provide clinical information for human-adjudicated¹¹ follow-up and decision making.

Conclusion

Using informatics to facilitate quality improvement is emerging as one of the meaningful end results of federal reimbursement incentives intended to facilitate adoption of the EHR.⁴⁹ Hospital systems with established electronic health records, such as the VA, have a viable platform for automated tools that conduct routine assessment of patient safety, potentially reducing or replacing manual review efforts. The breadth and scope of this study

demonstrates that the informatics field is providing easier access to patient clinical data that will substantially enable safety monitoring and quality improvement.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Funding/Support: This study was supported through the Department of Veterans Affairs grant SAF-03-223. Dr. Matheny is supported by the Veterans Health Administration HSR&D Career Development Award CDA-08-020. Drs. Matheny and Speroff are supported by the Veterans Health Consortium for Health Informatics Research (CHIR) awards HIR 09-001 and HIR 09-003. This material is the result of work supported with resources and the use of facilities at the VA Tennessee Valley Healthcare System

Role of the Sponsor: The sponsors were not involved in the design and conduct of the study, collection, management, analysis, or interpretation of the data; and preparation, review, or approval of the manuscript.

References

1. Kohn, LT.; Corrigan, J.; Donaldson, MS. To err is human : building a safer health system. Washington, D.C: National Academy Press; 2000.
2. Gawande AA, Thomas EJ, Zinner MJ, et al. The incidence and nature of surgical adverse events in Colorado and Utah in 1992. *Surgery*. 1999; 126:66–75. [PubMed: 10418594]
3. Khuri SF, Daley J, Henderson W, et al. The Department of Veterans Affairs' NSQIP: The first national, validated, outcome-based, risk-adjusted, and peer-controlled program for the measurement and enhancement of the quality of surgical care. *Annals of Surgery*. 1998; 228:491–507. [PubMed: 9790339]
4. Bruce J, Russell EM, Mollison J, et al. The measurement and monitoring of surgical adverse events. *Health Technology Assessment (Winchester, England)*. 2001; 5:1–194.
5. Fuchshuber PR, Greif W, Tidwell CR, et al. The power of the national surgical quality improvement program-achieving a zero pneumonia rate in general surgery patients. *Permanente Journal*. 2012; 16:39–45. [PubMed: 22529758]
6. Cullen DJ, Bates DW, Small SD, et al. The incident reporting system does not detect adverse drug events: a problem for quality improvement. *Joint Commission Journal on Quality Improvement*. 1995; 21:541–548. [PubMed: 8556111]
7. Desikan, R.; Krauss, MJ.; Dunagan, WC., et al. Reporting of adverse drug events: examination of a hospital incident reporting system. Henriksen, K.; Battles, JB.; Marks, ES., et al., editors. Rockville MD: Agency for Healthcare Research and Quality; 2005. p. 145-160.
8. Gandhi TK, Seger DL, Bates DW. Identifying drug safety issues: from research to practice. *International Journal for Quality in Health Care*. 2000; 12:69–76. [PubMed: 10733086]
9. Jha AK, Kuperman GJ, Teich JM, et al. Identifying adverse drug events: development of a computer-based monitor and comparison with chart review and stimulated voluntary report. *Journal of the American Medical Informatics Association*. 1998; 5:305–314. [PubMed: 9609500]
10. Murff HJ, Patel VL, Hripcsak G, et al. Detecting adverse events for patient safety research: a review of current methodologies. *Journal of Biomedical Informatics*. 2003; 36:131–143. [PubMed: 14552854]
11. Price, CS.; Savitz, LA. 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; Mar. 2012 Improving the measurement of surgical site infection risk stratification/outcome detection.
12. Iezzoni LI, Daley J. A description and clinical assessment of the computerized severity index. *Quality Review Bulletin*. 1992; 18:44–52. [PubMed: 1574320]

13. Rosen AK, Rivard P, Zhao S, et al. Evaluating the patient safety indicators: how well do they perform on Veterans Health Administration data? *Medical Care*. 2005; 43:873–884. [PubMed: 16116352]
14. Hammermeister K. The National Surgical Quality Improvement Program: learning from the past and moving to the future. *American Journal of Surgery*. 2009; 198:S69–73. [PubMed: 19874938]
15. Melton GB, Parsons S, Morrison FP, et al. Inter-patient distance metrics using SNOMED CT defining relationships. *Journal of Biomedical Informatics*. 2006; 39:697–705. [PubMed: 16554186]
16. Elkin PL, Brown SH, Balas A, et al. Biosurveillance evaluation of SNOMED CT's terminology (BEST Trial): coverage of chief complaints. *Studies in Health Technology and Informatics*. 2008; 136:797–802. [PubMed: 18487829]
17. Suzuki T, Yokoi H, Fujita S, et al. Automatic DPC code selection from electronic medical records: text mining trial of discharge summary. *Methods of Information in Medicine*. 2008; 47:541–548. [PubMed: 19023491]
18. Womack JA, Scotch M, Gibert C, et al. A comparison of two approaches to text processing: facilitating chart reviews of radiology reports in electronic medical records. *Perspectives in Health Information Management*. 2010; 7:1a.
19. Fiszman M, Chapman WW, Aronsky D, et al. Automatic detection of acute bacterial pneumonia from chest X-ray reports. *Journal of the American Medical Informatics Association*. 2000; 7:593–604. [PubMed: 11062233]
20. Pakhomov SV, Buntrock JD, Chute CG. Automating the assignment of diagnosis codes to patient encounters using example-based and machine learning techniques. *Journal of the American Medical Informatics Association*. 2006; 13:516–525. [PubMed: 16799125]
21. Chapman WW, Christensen LM, Wagner MM, et al. Classifying free-text triage chief complaints into syndromic categories with natural language processing. *Artificial Intelligence in Medicine*. 2005; 33:31–40. [PubMed: 15617980]
22. Chu, D.; Dowling, JN.; Chapman, WW. Evaluating the effectiveness of four contextual features in classifying annotated clinical conditions in emergency department reports. *AMIA; Annual Symposium Proceedings/AMIA Symposium*; 2006. p. 141-145.
23. Khuri SF, Henderson WG, Daley J, et al. The patient safety in surgery study: background, study design, and patient populations. *Journal of the American College of Surgeons*. 2007; 204:1089–1102. [PubMed: 17544068]
24. AHRQ. Guide to patient safety indicators version 3.1 (Revised March 2007). Rockville, MD: 2007.
25. Murff HJ, FitzHenry F, Matheny ME, et al. Automated identification of postoperative complications within an electronic medical record using natural language processing. *Journal of the American Medical Association*. 2011; 306:848–855. [PubMed: 21862746]
26. Davis CL, Pierce JR, Henderson W, et al. Assessment of the reliability of data collected for the Department of Veterans Affairs national surgical quality improvement program. *Journal of the American College of Surgeons*. 2007; 204:550–560. [PubMed: 17382213]
27. Brown SH, Speroff T, Fielstein EM, et al. eQuality: Electronic quality assessment from narrative clinical reports. *Mayo Clinic Proceedings*. 2006; 81:1472–1481. [PubMed: 17120403]
28. Chute CG, Cohn SP, Campbell KE, et al. The content coverage of clinical classifications. *Journal of the American Medical Informatics Association*. 1996; 3:224–233. [PubMed: 8723613]
29. Itani KM. Fifteen years of the national surgical quality improvement program in review. *American Journal of Surgery*. 2009; 198:S9–S18. [PubMed: 19874939]
30. Rowell KS, Turrentine FE, Hutter MM, et al. Use of national surgical quality improvement program data as a catalyst for quality improvement. *Journal of the American College of Surgeons*. 2007; 204:1293–1300. [PubMed: 17544087]
31. Khuri SF, Henderson WG, Daley J, et al. Successful implementation of the Department of Veterans Affairs' national surgical quality improvement program in the private sector: the patient safety in surgery study. *Annals of Surgery*. 2008; 248:329–336. [PubMed: 18650645]
32. Newcombe RG. Interval estimation for the difference between independent proportions: comparison of eleven methods. *Statistics in Medicine*. 1998; 17:873–890. [PubMed: 9595617]

33. American Society of Anesthesiologists. [Accessed 4–26–2012] ASA physical status classification system. 2012. Available at: <http://www.asahq.org/Home/For-Members/Clinical-Information/ASA-Physical-Status-Classification-System>
34. Hunink, M.; Glasziou, P.; Siegel, J., et al. Decision making in health and medicine: integrating evidence and values. New York, NY: Cambridge University Press; 2001.
35. Hanauer DA, Englesbe MJ, Cowan JA Jr, et al. Informatics and the American College of Surgeons national surgical quality improvement program: automated processes could replace manual record review. *Journal of the American College of Surgeons*. 2008; 208:37–41. [PubMed: 19228500]
36. Elkin, PL.; Froehling, D.; Wahner-Roedler, D., et al. NLP-based identification of pneumonia cases from free-text radiological reports. *AMIA 2008 Symposium Proceedings*; 2008. p. 172-176.
37. Stanfill MH, Williams M, Fenton SH, et al. A systematic literature review of automated clinical coding and classification systems. *Journal of the American Medical Informatics Association*. 2010; 17:646–651. [PubMed: 20962126]
38. Evans, RS.; Lloyd, JF.; Aston, VT., et al. Computer surveillance of patients at high risk for and with venous thromboembolism. *AMIA 2010 Symposium Proceedings*; 2010. p. 217-221.
39. Fontaine P, Ross SE, Zink T, et al. Systematic review of health information exchange in primary care practices. *Journal of the American Board of Family Medicine*. 2010; 23:655–670. [PubMed: 20823361]
40. Forster AJ, Murff HJ, Peterson JF, et al. The incidence and severity of adverse events affecting patients after discharge from the hospital. *Annals of Internal Medicine*. 2003; 138:161–167. [PubMed: 12558354]
41. Zhan C, Miller MR. Excess length of stay, charges, and mortality attributable to medical injuries during hospitalization. *Journal of the American Medical Association*. 2003; 290:1868–1874. [PubMed: 14532315]
42. Ingraham AM, Richards KE, Hall BL, et al. Quality improvement in surgery: the American College of Surgeons national surgical quality improvement program approach. *Advances in Surgery*. 2010; 44:251–267. [PubMed: 20919525]
43. Guillamondegui OD, Gunter OL, Hines L, et al. Using the national surgical quality improvement program and the Tennessee surgical quality collaborative to improve surgical outcomes. *Journal of the American College of Surgeons*. 2012; 214:709–714. [PubMed: 22265639]
44. Khuri SF, Daley J, Henderson WG. The comparative assessment and improvement of quality of surgical care in the Department of Veterans Affairs. *Archives of Surgery*. 2002; 137:20–27. [PubMed: 11772210]
45. American College of Surgeons. [Accessed 8–14–2011] American College of Surgeons national quality improvement program. 2006. Available at: http://acsnsqip.org/main/resources/resources_faqs.jsp
46. Warner HR Jr. Can natural language processing aid outpatient coders? *Journal of American Health Information Management Association*. 2000; 71:78–81.
47. Gobbel, GT.; Reeves, R.; Jayaramaraja, S., et al. Evaluation of the RapTAT automated annotation system for identifying post-traumatic stress disorder-related concepts in clinical note. *AMIA Conference*; Chicago, IL. 2012.
48. Reeves RM, Ong FR, Matheny ME, et al. Detecting temporal expressions in medical narratives. *International Journal of Medical Informatics*. 2012 May 15.
49. Centers for Medicare and Medicaid Services. [Accessed July 20, 2010] Medicare quality monitoring System. 2010. Available at: http://146.123.40.205/QualityInitiativesGenInfo/15_MQMS.asp

Table 1

Description of the Study Sample

	Development Set	Test Set
Surgical Procedures (n)	4098	4088
Age, mean (S.D.)	62.8 (13.0)	63.2 (12.8)
Gender, % male	94.7	95.2
Race, %		
White	78.8	79.3
African-American	13.4	13.6
Other or Unknown	7.8	7.1
Current Smoker, %	41.4	39.8
Inpatient Procedure, %	72.8	73.9
Surgical Specialty Type, %		
General Surgery	38.7	39.7
Orthopedic	22.1	21.2
Vascular	11.2	11.7
Urology	10.0	10.9
Cardiothoracic	6.4	6.0
Neurosurgery	4.9	4.3
Ear Nose Throat	1.6	1.5
Other	4.1	3.4
Emergent Cases, %	12.1	12.0
ASACLAS ¹ , %		
Normal	1.9	1.7
Mild systemic disease	24.0	24.8
Severe systemic disease	56.2	55.2
Threat to Life	16.7	17.2
Moribund	1.3	1.1
Medical Center, %		
A	10.1	10.1
B	18.5	18.7
C	15.2	14.9
D	17.6	18.2
E	16.5	16.3
F	22.1	21.8
Surgical Adverse Event ² , %		
Acute Renal Failure	2.3	2.4

	Development Set	Test Set
Cardiac Arrest	4.2	4.2
Deep Vein Thrombosis	1.6	1.9
Myocardial Infarction	2.2	2.3
Pneumonia	14.7	14.8
Sepsis	7.0	6.9
Urinary Tract Infection	13.5	13.5
Wound Infection	19.6	19.6
Any of the above	49.3	50.0

¹ASACLAS= American Society of Anesthesiology Classification

²Representation of complications in the study were overly weighed in the sample and do not indicate VA population-based rates of complication which are much lower.

Table 2

POEM Rule Performance in the Development and Test Data Sets

Adverse Event	Development Set n = 4098 procedures		Test Set Sample n = 4088 procedures		Difference between Test and Development Sets	
	Sensitivity (95% CI) [/]	Specificity (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)
Acute Renal Failure	0.81 (0.73, 0.88)	0.93 (0.92, 0.93)	0.85 (0.76, 0.90)	0.92 (0.91, 0.92)	0.04 (−0.08, 0.15)	−0.01 (−0.02, 0.0)
Cardiac Arrest	0.81 (0.75, 0.86)	0.95 (0.94, 0.95)	0.88 (0.83, 0.93)	0.92 (0.92, 0.93)	0.07 (−0.02, 0.14)	−0.02 (−0.03, −0.01)
Deep Vein Thrombosis	0.54 (0.42, 0.65)	0.94 (0.93, 0.95)	0.56 (0.45, 0.67)	0.94 (0.93, 0.95)	0.02 (−0.14, 0.19)	0.0 (−0.01, 0.01)
Myocardial Infarction	0.90 (0.83, 0.95)	0.90 (0.89, 0.91)	0.88 (0.80, 0.93)	0.89 (0.88, 0.90)	−0.02 (−0.12, 0.08)	−0.01 (−0.03, 0.0)
Pneumonia	0.77 (0.74, 0.80)	0.91 (0.90, 0.92)	0.80 (0.76, 0.83)	0.90 (0.89, 0.91)	0.03 (−0.02, 0.07)	−0.01 (−0.03, 0.0)
Pulmonary Embolism	0.67 (0.52, 0.79)	0.98 (0.98, 0.98)	0.80 (0.66, 0.89)	0.97 (0.96, 0.98)	0.13 (−0.05, 0.31)	−0.01 (−0.02, 0.0)
Sepsis	0.87 (0.82, 0.90)	0.93 (0.92, 0.93)	0.88 (0.84, 0.91)	0.92 (0.91, 0.93)	0.01 (−0.04, 0.07)	−0.01 (−0.02, 0.0)
Urinary Tract Infection	0.92 (0.90, 0.94)	0.80 (0.79, 0.81)	0.95 (0.93, 0.96)	0.80 (0.79, 0.81)	0.03 (−0.01, 0.05)	0.0 (−0.01, 0.03)
Wound Infection	0.79 (0.76, 0.81)	0.66 (0.64, 0.67)	0.77 (0.74, 0.80)	0.63 (0.61, 0.64)	−0.02 (−0.06, 0.02)	−0.03 (−0.05, 0.01)

[/] CI = Confidence Interval

Table 3

Value of Information: Positive Predictive Value and Likelihood Ratio Positive of POEM Identification of Adverse Events

Adverse Event	Prevalence of adverse event in the Study Sample ¹	Positive Predictive Value (95% CI) ² in the test set	Likelihood Ratio Positive ³ (95% CI)	Likelihood Ratio Negative ³ (95% CI)
Acute Renal Failure	0.02	0.20 (0.16, 0.24)	10.0 (8.8, 11)	0.17 (0.10, 0.27)
Cardiac Arrest	0.04	0.34 (0.30, 0.39)	12.0 (10, 13)	0.13 (0.09, 0.20)
Deep Vein Thrombosis	0.02	0.15 (0.11, 0.20)	9.2 (7.3, 12.0)	0.46 (0.36, 0.60)
Myocardial Infarction	0.02	0.16 (0.13, 0.19)	7.8 (7.0, 9.8)	0.13 (0.07, 0.23)
Pneumonia	0.15	0.58 (0.54, 0.61)	7.8 (7.0, 8.7)	0.23 (0.19, 0.27)
Pulmonary Embolism	0.01	0.23 (0.17, 0.30)	27.0 (21, 34)	0.21 (0.11, 0.37)
Sepsis	0.07	0.44 (0.40, 0.48)	10.0 (9.4, 12)	0.13 (0.10, 0.18)
Urinary Tract Infection	0.14	0.43 (0.40, 0.46)	4.8 (4.5, 5.1)	0.07 (0.05, 0.09)
Wound Infection	0.19	0.33 (0.31, 0.35)	2.1 (1.9, 2.2)	0.37 (0.33, 0.42)

¹ Study Sample is the combined development and test set, n= 8,186 procedures. The test set included n = 4,088 procedures.

² CI = Confidence Interval.

³ Positive Likelihood Ratio is the sensitivity divided by one minus the specificity and Negative Likelihood Ratio is one minus the sensitivity divided by specificity.

Table 4

POEM Performance Sensitivity on the Total Study Sample[/] for Adverse Events by Location within Hospital, Post-Discharge, and Outpatient

Adverse Event	Inpatient Surgical Procedures N = 6001		Outpatient Surgical Procedures N = 2185
	Sensitivity on Hospital Events	Sensitivity on Post-Discharge Events	Sensitivity
Acute Renal Failure	135/159 (0.85)	19/24 (0.79)	8/12 (0.67)
Cardiac Arrest	267/315 (0.85)	21/25 (0.84)	7/9 (0.78)
Deep Vein Thrombosis	47/95 (0.49)	30/40 (0.75)	5/14 (0.36)
Myocardial Infarction	144/157 (0.92)	15/21 (0.71)	9/10 (0.90)
Pneumonia	802/1017 (0.79)	107/138 (0.78)	34/49 (0.69)
Pulmonary Embolism	46/64 (0.72)	18/23 (0.78)	2/3 (0.67)
Sepsis	423/485 (0.87)	59/69 (0.86)	16/17 (0.94)
Urinary Tract Infection	694/729 (0.95)	243/261 (0.93)	95/112 (0.85)
Wound Infection	560/670 (0.84)	512/657 (0.78)	165/266 (0.62)

[/]The total study sample (n=8,186 procedures) is the development and test set of VASQIP procedures combined together.