New and Increasing Rates of Adverse Events Can be Found in Unstructured Text in Electronic Health Records using the Shakespeare Method

Roselie A. Bright, ScD[1]; Katherine Dowdy[2]; Summer K. Rankin, PhD[2]; Sergey V. Blok, PhD[2]; Lee Anne Palmer, VMD, MPH[3]; Susan J. Bright-Ponte, DVM, MPH[3]

[1] Office of the Commissioner, Food and Drug Administration (FDA), Silver Spring, MD, USA. Corresponding author: Roselie.Bright@fda.hhs.gov.
[2] Booz Allen Hamilton, McLean, VA, USA
[3] Center for Veterinary Medicine, Food and Drug Administration, Rockville, MD, USA

ABSTRACT

Background: Text in electronic health records (EHRs) and big data tools offer the opportunity for surveillance of adverse events (patient harm associated with medical care) (AEs) in the unstructured notes. Writers may explicitly state an apparent association between treatment and adverse outcome ("attributed") or state the simple treatment and outcome without an association ("unattributed"). We chose to study EHRs from 2006-2008 because of known heparin contamination during this timeframe. We hypothesized that the prevalence of adulterated heparin may have been widespread enough to manifest in EHRs through symptoms related to heparin adverse events, independent of clinicians' documentation of attributed AEs.

Objective: Use the Shakespeare Method, a new unsupervised set of tools, to identify attributed and unattributed potential AEs using the unstructured text of EHRs.

Methods: We studied 21,287 adult critical care admissions divided into three time periods. Comparisons of period 3 (7/2007 to 6/2008) to period 2 (7/2006 to 6/2007) were used to find admissions notes to review for new or increased clinical events by generating Latent Dirichlet Allocation topics among words in period 3 that were distinct from period 2. These results were further explored with frequency analyses of periods 1 (7/2001 to 6/2006) through 3.

Results: Topics represented unattributed heparin AEs, other medical AEs, rare medical diagnoses, and other clinical events; all were verified with EHRs notes review and frequency analysis. The heparin AEs were not attributed in the notes, diagnosis codes, or procedure codes. Somewhat different from our hypothesis, heparin AEs increased in prevalence from 2001 through 2007, and decreased starting in 2008 (when heparin AEs were being published).

Conclusions: The Shakespeare Method could be a useful supplement to AE reporting and surveillance of structured EHRs data. Future improvements should include automation of the manual review process.

KEYWORDS

Electronic health records; big data; patient harm; patient safety; public health; product surveillance, postmarketing; natural language processing; proof of concept study; critical care; heparin; drug contamination; humans

INTRODUCTION

Avoidable patient harm continues to be a significant problem [1]. To learn of patient harm known as adverse events (AEs) related to products it regulates, FDA relies on spontaneous reports from manufacturers, healthcare providers, and the general public. Published deficiencies of these reports [2-10] include well known biases in reporting. Now that electronic healthcare records (EHRs) are very common [11] and seen as more informative than billing codes from payment claims [6, 12, 13] we have an opportunity to leverage them for automated surveillance of AEs [6, 14, 15].

Many methods for finding AEs in text [6, 7, 9, 16-38] rely on predefining the possible AEs. Writers may explicitly state an apparent association between treatment and adverse outcome ("attributed") or state the simple treatment and outcome without an association ("unattributed"). More critically, attributed and unattributed potential AEs (PAEs) may not necessarily be captured in structured data (e.g., diagnosis and procedure codes) [14, 23, 39].

Many medical care AEs occur at higher frequency in hospital critical care settings, related to complex illnesses, invasive procedures, and relatively long lists of treatments [40, 41]. In previous work, we performed a comparison of transfused to non-transfused admissions to critical care at a major teaching hospital [42] that successfully found potential blood transfusion adverse events, while addressing many published challenges (such as synonyms, overlapping meanings, and nonstandard terms) with using unstructured EHRs text [5, 11, 14, 19, 23].

We hoped the Shakespeare Method [42] would overcome the challenges of EHRs text to detect not only clinical and administrative changes but also trending potential AEs (PAEs), including heparin contamination PAEs which were first reported early in 2008 [43].

METHODS

The Shakespeare strategy is to find unusual, significant words that were new or increased in the most recent time period, use topic analysis to find words that tended to occur together, examine admissions that were prominent for topics of interest, and then evaluate how well the topics performed [42].

Study Population

We used EHRs for critical care admissions within an adult hospital, Beth Israel Deaconess Medical Center, Boston, MA the Medical Information Mart for Intensive Care III (MIMIC-III) [35, 44], which used one medical record system in 2001-2008 and another afterwards. We received the real dates, within several weeks, for the earlier data. MIMIC III is publicly available to those meeting human subjects research requirements. The research was designated not human subjects research by the FDA Institutional Review Board under the Code of Federal Regulations [45].

We wanted to simulate real-time analysis to find new or increasing events in the most recent time period. MIMIC-III data were collected with two sequential EHRs, so we selected the longer, earlier

period of exclusive use of the earlier EHRs (7/2001-6/2008). We restricted the admissions to patients > 16 years old because this was a hospital for adults.

Preprocessing

We concatenated in chronological order all text notes for a hospital admission into a document. We removed the personally identifying information mask string and lowercased the text, and retained punctuation, numerals and stop words (because they convey clinical information and are sometimes components of abbreviations).

Since our methods would be based on the frequencies of words, we eliminated duplicate sentences because they do not represent additional information and give weight to variable personal duplication practices. We removed widespread duplicated sentences and lists within the notes, using Bloatectomy [46].

Word Extraction

We utilized sci-kit learn's CountVectorizer [47, 48] to convert each document into a bag of words vector where each dimension is represented by the frequency of each n-gram present in the document (see Figure 1a and 1b). Details are in Table 1.

We then divided the study population into three cohorts: (Period 1) admissions starting between 7/1/2001 and 6/30/2006 (14,410 documents); (Period 2) 7/1/2006-6/30/2007 (3,581 documents), and (Period 3) 7/1/2007-6/30/2008 (3,296 documents).

To focus on new or increasing AEs, we reduced the number of words to analyze by filtering by whether they were unusual and increasing (or new) in period 3 compared to period 2 (see Figures 1c, 1d and 2a). We adopted two parallel approaches, shown in Figure 2: through binary classification of the notes, and analysis of term frequency between periods 3 and 2.

For the binary classification, we fit two classification models: logistic regression (LR) with L2 / ridge regularization [49] and multinomial naïve Bayes (NB) [50, 51]. Model evaluation found LR outperformed NB (with a weighted average F1 score of 0.76 compared to NB's weighted average F1 of 0.69), but that NB more effectively identified completely new terms in the target time period.

After evaluating the models, we re-fit both models without a train-test split on the entire 24-month dataset and combined the top 5,000 features from LR (those with the highest positive coefficient, associated with the positive target class) and the top 5,000 features from NB (those with the lowest log probability ratio). Combining lists resulted in a set of 9,896 terms.

We used frequency analysis to find emerging rare clinical events. We identified two groups of terms: those which appeared in fewer than 10% of documents in period 2 and saw a 30% increase in raw frequency in period 3, and any terms that never appeared in period 2 and did appear in period 3. For those new terms appearing in period 3, we filtered out digit-only terms (a large number of terms in this group).

For the final feature set, we took the intersection of terms identified from the binary classification and frequency analysis processes. This resulted in 6,122 significant terms identified from the initial 117,049

unique terms in documents from period 3 (5.2% of terms). We re-vectorized (Figure 1e) the 12-month corpus from period 3 using the combined feature list as our vocabulary (which has the effect of filtering the notes to only include terms in the vocabulary).

Topic Modeling and Interpretation

The co-occurrence of words in documents in the last time period was analyzed with Latent Dirichlet Allocation (LDA) topic analysis [52]. We chose the final number of topics (20) based on a balance of large and small topics and at least one topic with no substantive words. We used the words with the highest scores of their relationship to topics (Figure 1f), as well as the topic document scores that indicate the probability of the topic fit for a document (Figure 1g), to explore topic meanings. We manually read the three top-scoring documents for each topic (Figure 1h).

Statistical Analysis of Words and Codes Suggested by Manual Review of the Topics

Documents from selected individual admissions, as well as summary data from 7/2001 to 6/2008 were used to evaluate whether any topics formed around AEs. Most topics inspired time plots of selected words, diagnosis codes, or procedure codes through periods 1, 2, and 3. Slopes were analyzed for changes [53, 54].

For this report, out of concern for patient privacy we substituted generic words (such as “condition01”, “condition02”, etc.) for rare conditions, drugs, events, and languages because the year of admission is being presented. Related substitute words (e.g., “condition09a”, “condition09b”) were used for synonyms.

RESULTS

Table 2 shows the statistics for each topic. The strength of the maximum word score in a topic roughly corresponded with the number of admissions that had strong matches with the topic. The words in many of the topics seem to readily suggest interpretations, for example: long complex stay (topic 18), heart problem (3), trauma (19), cardiac catheterization (7), brain (1), cardiac catheterization (17), abdomen (12), uterus (16), and a foreign language (2). The other topics seemed broad.

Common topics

For the most common topics, the admissions with the top three topic match scores are summarized in Table 3. For the topics with words that suggested an interpretation, the records supported the interpretations. For the other topics, the records suggested interpretations that were consistent with the top words. Each of the three top scoring admissions within a topic were quite similar to each other (an indication that the topics were coherent and the model was working correctly; with the exception of the third admission in topic 3).

The top three scoring documents for topic 18 described long complex stays, which included large numbers of notes. The general words in the topic (“for”, “hr”, “plan”, “cont”, “today”, “skin”, and “are”) are nearly ubiquitous in periods 2 and 3. The words indicating mechanical ventilation (“vent”, “intubated”, and “trach”) were present in between 51% and 58% of the admissions per quarter in

periods 2 and 3, with a slight, not clinically significant increase for period 3. The lengths of stay and numbers of notes also did not vary between periods 2 and 3.

We noticed that among the five records in Table 3 that mentioned cardiac catheterization, all mentioned explicit or implied dosing with heparin followed the same day with hypotension that required treatment (heparin is generally involved with cardiovascular procedures) [55].

Topics 3 and 7 both have cardiac catheterization for heart problems in common; for five out of six instances, the procedure or heparin administration was followed by hypotension (four instances) that needed to be treated or heart rhythm deterioration (one instance). To investigate whether these potential heparin AEs were increasing 7/2001-6/2008, we plotted two measures of exposure (invasive cardiac procedure code and "heparin") and a measure of AE ("hypotension"). The proportion of admissions that had invasive cardiovascular procedure codes (see Figures 3a and b) declined overall (see Figure 3a), but had a local increase in period 3, compared to period 2. In contrast to the procedures, the words "heparin" and "hypotension" showed an overall rough increase over the entire timeframe. We also noticed that the proportion of admissions with invasive cardiology codes that had the word "hypotension" increased gradually over time (Figures 3a and b), followed by a drop in the last quarter; the pattern was similar and weaker for the proportion of admissions with "heparin" that also had "hypotension". There was a decrease in "hypotension" in the last quarter, both as a proportion of all admissions, and as a proportion of either indicator of having been exposed to heparin.

Other common topics

Topic 19 (and 13) corresponded with trauma. Figure 4 showed that trauma diagnosis and procedure codes increased steadily over time through periods 1-3.

The brain topic (1 and 17, combined) is centered around admissions for brain injury: either bleeding, ischemia, or trauma. Figures 5a, 5b, and 5c show that there were local increases in codes for bleeding and ischemia for period 3 compared to period 2. There were slight increases in the codes for all three types of brain injuries overall. The text words indicating these conditions showed similar trends.

Topic 4 describes prolonged drainage after abdominal surgery. The index surgeries were performed before admission for two instances and during hospitalization for the third. Figure 6 shows that codes for wounds were quite infrequent. However, long patient stays with words for leaky surgical wound or catheter were more common, rose gradually over time, and had a local increase in period 3, compared to period 2.

Condition01 was the subject of the admissions with the top match scores for topic 12. The codes and words were generally rare for the 3 periods and showed a local increase between periods 2 and 3.

Less common topics

Summaries of admissions with topic matching scores for the less common topics are shown in Table 4. We examined the top scoring admissions matched to topic 11 and all admissions matched to the others. All admissions in this Table had topic match scores for the index topic of <0.15 (column 2). Despite each admission in Table 4 having at least one strong topic match score for at least one of the strong topics in Table 3, the topics in Table 4 are distinct from those in Table 3. Some of the topics have admissions that have common aspects (topics 11, 10, 2, 9).

Fourteen PAEs evident in the notes were distributed among the less common topics: 13 related to medical therapy (6 medications, 3 medical devices, 2 procedures, and 2 combinations) and 2 non-medical. Five drug and all of the medical device PAEs are published in the product labels and/or medical literature. Nine of the PAEs occurred outside the hospital and were related to the reason for admission. The diagnosis and procedure codes generally did not give enough information to understand the specific cause and associated potential AE. Figure 7 shows that while the proportions over the 7 years of admissions with allergy and anaphylaxis words steadily decreased, the diagnosis codes for drug AEs and for surgical or procedure AEs increased slightly over time.

The other rare and infrequent terms, related diagnosis or procedure codes, and foreign language sentences were rare throughout all three time periods and increased during period 3.

DISCUSSION

We succeeded in our expectation of finding increases in clinical events and our hope of finding increases in AEs, especially AEs that would not have been reported because they were not attributed. We found increases in hypotension following heparin or presumed-heparin exposure. Hypotension occurring in the cardiac catheterization lab could be a vasovagal reaction [56]. However, vasovagal reaction generally does not respond to fluids and drugs for raising blood pressure, and all our observed patients' hypotension did respond to treatment. Hypotension can occur as anaphylaxis begins and, alone, may reflect mild anaphylaxis. We note that the nurses and physicians that described the sequence of events did not link sudden hypotension to heparin and the diagnosis codes did not reflect any awareness of a link. The warnings from FDA and the Centers for Disease Control and Prevention about heparin in the winter of 2007-2008 were for anaphylaxis due to adulterated heparin [57, 58]. Knowledge of the extent of the distribution of adulterated heparin products was not specific, so it may have been in the hospital's stock at the time. We had expected to see increases starting in 2006 because a few articles indicate heparin may have been adulterated before 2007 [59-61], but were surprised that the increases had started before 2006. The reduction in the last quarter coincided with recalls of contaminated heparin products and lend credibility to the idea that contaminated heparin was in slowly increasing use at this hospital for many years. We are struck that such a high proportion of the invasive cardiac catheter patients in the last two years experienced hypotension following heparin exposure (either as explicitly documented administration or implicitly in the catheter coating).

The types of clinical event changes we detected from period 2 to period 3 were: increases in patients with common conditions (heart disease, brain injuries, trauma, and complex conditions associated with long hospital stays), increases in rare conditions, change in administration (foreign language portion), and adverse events of concern.

The increases in common conditions may have reflected hospital marketing [62].

The increases in rare conditions could have reflected chance, or marketing as a referral center.

Nine of the adverse events happened outside the hospital and illustrate the utility of hospital records for monitoring severe reactions that occur in other health facilities or outside the healthcare system. Our

method was useful for detecting words that are rare in hospital records, partly reflecting events that normally occur outside the hospital.

The topic with the highest document score exhibited typical behavior of a topic containing words that are common to most documents. The filter that was removing words comprised of only digits also removed digits from some words. This resulted in some high frequency words getting into the vocabulary. When topic modeling, this resulted in high scores for these common words in the topics where they were correlated (as expected this happens in several topics) and created a common word topic (topic 18). This topic is a *noise* topic; the LDA model will put words that are low scoring and not correlated with other topics into their own noise topic in order to deal with noise and frequent words. Because this topic included words that were frequent in almost all documents, as expected the document topic scores for this topic were high [63]. This was dealt with by looking at the other more coherent topics that were assigned to each document (essentially ignoring this common-noise topic, capturing what most documents have in common. The top scoring words in this topic that were general survived the ensemble filtering method as an artifact of the digit-removal step. For future work, we recommend removing this step from the filtering process and relying on the classification terms to filter out irrelevant variations of terms.

Our method worked despite:

- the known challenges posed by clinical text notes
- restriction to one major hospital
- lack of all surgical and non-CCU nursing notes, and variable lack of physician, nursing, or discharge summary notes, probably reflecting hospital policy of gradually converting types of notes to EHRs [64]
- errors up to several weeks in dates.

Different, and hopefully improved, results may be derived from EHRs databases that are more complete and have actual dates.

Much of our manual work to evaluate topics could be reduced with a combination of natural language processing and dictionaries of clinical terms. Dictionaries should include standard acronyms and common abbreviations, and should try to account for context when the meaning of term could be ambiguous. The ability to decipher ongoing care notes will be important for noticing unrecognized signals of AEs.

CONCLUSIONS

We suggest that heparin contamination may have occurred earlier than previously recognized in the winter of 2007-2008, at a lower rate.

Our method successfully aided in the detection of a variety of medical product AEs that were not attributed in clinicians' notes, suggesting that this method could be a useful supplement to existing post-marketing surveillance programs at local as well as national levels. The method also found other changes in clinical care experiences. The method is easy to execute and understand and could be adopted by subnational public and private entities. It finds potential adverse events that are candidates for causality assessment with epidemiology or other clinical studies.

Our method enabled manual review of key EHRs by narrowing interest from the original large volume of words used in notes. Future improvements could include automation of the manual review process.

ACKNOWLEDGEMENTS

We thank enthusiastic support by our FDA and Booz Allen Hamilton supervisors, Department of Health and Human Services innovation programs (Ignite Accelerator and Data Science CoLab), and Alistair Johnson, DPhil, of the MIMIC-III program, Massachusetts Institute of Technology. George Plopper, PhD, of Booz Allen Hamilton, provided project and consultation support. Many FDA colleagues offered ideas and feedback regarding the selection of the case and the final paper. All authors had access to the data. All authors are responsible for the study topic, design, and interpretation. Dr. Bright, Ms. Dowdy, Dr. Rankin, and Dr. Blok are responsible for data processing and analysis.

Conflict of interest and disclaimer: The research was done with FDA support and under contract HHSF223201510027B between FDA and Booz Allen Hamilton Inc. None of the authors have other relevant financial interests. The opinions are those of the authors and do not represent official policy of either the FDA or Booz Allen Hamilton.

ABBREVIATIONS USED MORE THAN ONCE

AE	Adverse events
AF	Atrial fibrillation
BIDMC	Beth Israel Deaconess Medical Center
CABG	Coronary artery bypass graft
CCU	Critical (or Intensive) Care Unit
CPR	Cardiopulmonary resuscitation
DMII	Diabetes mellitus, type 2
DVT	Deep vein thrombosis
EHRs	Electronic healthcare records
FDA	Food and Drug Administration
HD	Hospital day
HIT	Heparin induced thrombocytopenia
IABP	Intra-aortic balloon pump
IPH	Intraparenchymal hemorrhage
IV	Intravenous
LDA	Latent Dirichlet Allocation algorithm for topic modeling
LR	Logistic regression supervised learning algorithm
MCA	Middle cerebral artery
MIMIC-III	Medical Information Mart for Intensive Care III
MRI	Magnetic resonance image
MVA	Motor vehicle accident
MVC	Motor vehicle collision
NB	Naïve Bayes supervised learning algorithm
NLP	Natural language processing
O2	Oxygen

OR	Operating room
PAE	Potential adverse event
PICC	Peripherally inserted central catheter
POD	Post-operative day
tPA	Tissue plasminogen activator
UTI	Urinary tract infection

REFERENCES

1. Brewer T, Colditz GA. Postmarketing surveillance and adverse drug reactions: current perspectives and future needs. *JAMA*. 1999 Mar 3;281(9):824-9. PMID:10071004. DOI: 10.1001/jama.281.9.824.

2. Scott HD, Thacher-Renshaw A, Rosenbaum SE, et al. Physician reporting of adverse drug reactions: Results of the Rhode Island Adverse Drug Reaction Reporting Project. *JAMA.* 1990;263:1785-1788. PMID:2313850. doi:10.1001/jama.1990.03440130073028.

3. Bright RA, Nelson RC. Automated support for pharmacovigilance: a proposed system. *Pharmacoepidemiol Drug Saf*. 2002; 11(2):121-125. PMID:11998536. DOI:10.1002/pds.684.

4. Samore MH, Evans RS, Lassen A, et al. Surveillance of medical device-related hazards and adverse events in hospitalized patients. *JAMA.* 2004; 291:325-34. PMID:14734595 DOI:10.1001/jama.291.3.325.

5. Bright RA. Strategy for surveillance of adverse drug events. *Food Drug Law J.* 2007; 62(3):605-615. PMID:17915403.

6. Hoang T, Liu J, Pratt N, Zheng VW, et al. Authenticity and credibility aware detection of adverse drug events from social media. *Int J Med Inform*. 2018 Dec;120:101-115. PMID:30409335. doi:10.1016/j.ijmedinf.2018.09.002.

7. Classen D, Li M, Miller S, Ladner D. An electronic health record-based real-time analytics program for patient safety surveillance and improvement. *Health Aff (Millwood)*. 2018 Nov;37(11):1805-1812. PMID:30395491. DOI:10.1377/hlthaff.2018.0728.

8. Wang L, Rastegar-Mojarad M, Ji Z, et al. Detecting pharmacovigilance signals combining electronic medical records with spontaneous reports: A case study of conventional disease-modifying antirheumatic drugs for rheumatoid arthritis. *Front Pharmacol*. 2018 Aug 7;9:875. PMID:30131701. DOI:10.3389/fphar.2018.00875.

9. Alghamdi AA, Keers RN, Sutherland A, Ashcroft DM. Prevalence and nature of medication errors and preventable adverse drug events in paediatric and neonatal intensive care settings: A systematic review. *Drug Saf*. 2019 Dec;42(12):1423-1436. PMID:31410745. DOI:10.1007/s40264-019-00856-9.

10. Molina FJ, Rivera PT, Cardona A, et al. Adverse events in critical care: Search and active detection through the Trigger Tool. *World J Crit Care Med*. 2018 Feb 4;7(1):9-15. PMID:29430403. DOI:10.5492/wjccm.v7.i1.9.

11. Report to Congress: Update on the adoption of health information technology and related efforts to facilitate the electronic use and exchange of health information. Office of the National Coordinator for Health Information Technology, US Department of Health and Human Services. 2016 Feb. https://www.healthit.gov/sites/default/files/Attachment_1_-_2-26-16_RTC_Health_IT_Progress.pdf.

12. Taggart M, Chapman WW, Steinberg BA, et al. Comparison of 2 natural language processing methods for identification of bleeding among critically ill patients. *JAMA Netw Open*. 2018 Oct 5;1(6):e183451. PMID:30646240. DOI:10.1001/jamanetworkopen.2018.3451.

13. Jin Y, Li F, Vimalananda VG, Yu H. Automatic detection of hypoglycemic events from the electronic health record notes of diabetes patients: Empirical study. *JMIR Med Inform*. 2019 Nov 8;7(4):e14340. PMID:31702562. DOI:10.2196/14340.

14. Melton GB, Hripcsak G. Automated detection of adverse events using natural language processing of discharge summaries. *J Am Med Inform Assoc*. 2005; 12:448-457. PMID:15802475. DOI:10.1197/jamia.M1794.

15. Patadia VK, Schuemie MJ, Coloma PM, Herings R, van der Lei J, Sturkenboom M, Trifiro G. Can electronic health records databases complement spontaneous reporting system databases? A historical-reconstruction of the association of rofecoxib and acute myocardial infarction. *Front Pharmacol*. 2018 Jun 6;9:594. PMID:29928230. DOI:10.3389/fphar.2018.00594.

16. Young IJB, Luz S, Lone N. A systematic review of natural language processing for classification tasks in the field of incident reporting and adverse event analysis. *I J Med Inf.* 2019; 132: 103971. PMID:31630063. DOI:10.1016/j.ijmedinf.2019.103971.

17. Fortenberry M, Odinet J, Shah P, et al. Development of an electronic trigger tool at a children's hospital within an academic medical center. *Am J Health Syst Pharm*. 2019 Nov 13;76(Suppl_4):S107-S113. PMID:31724037. DOI:10.1093/ajhp/zxz222.

18. Zhou L, Siddiqui T, Seliger SL, et al. Text preprocessing for improving hypoglycemia detection from clinical notes - A case study of patients with diabetes. *Int J Med Inform*. 2019 Sep;129:374-380. PMID:31445280. DOI:10.1016/j.ijmedinf.2019.06.020.

19. Mesfin YM, Cheng A, Lawrie J, Buttery J. Use of routinely collected electronic healthcare data for postlicensure vaccine safety signal detection: A systematic review. *BMJ Glob Health*. 2019 Jul 8;4(4):e001065. PMID:31354969. DOI:10.1136/bmjgh-2018-001065.

20. Morel M, Bacry E, Gaiffas S, Guilloux A, Leroy F. ConvSCCS: convolutional self-controlled case series model for lagged adverse event detection. *Biostatistics*. 2019 Mar 8. pii: kxz003. PMID:30851046. DOI:10.1093/biostatistics/kxz003.

21. Dandala B, Joopudi V, Devarakonda M. Adverse drug events detection in clinical notes by jointly modeling entities and relations using neural networks. *Drug Saf*. 2019 Jan;42(1):135-146. PMID:30649738. DOI:10.1007/s40264-018-0764-x.

22. Wunnava S, Qin X, Kakar T, Sen C, Rundensteiner EA, Kong X. Adverse drug event detection from electronic health records using hierarchical recurrent neural networks with dual-level embedding. *Drug Saf*. 2019 Jan;42(1):113-122. PMID:30649736. DOI:10.1007/s40264-018-0765-9.

23. Bagattini F, Karlsson I, Rebane J, Papapetrou P. A classification framework for exploiting sparse multi-variate temporal features with application to adverse drug event detection in medical records. *BMC Med Inform Decis Mak*. 2019 Jan 10;19(1):7. PMID:30630486. DOI:10.1186/s12911-018-0717-4.

24. Rafter N, Finn R, Burns K, et al. Identifying hospital-acquired infections using retrospective record review from the Irish National Adverse Events Study (INAES) and European point prevalence survey case definitions. *J Hosp Infect*. 2019 Mar;101(3):313-319. PMID:30590090. DOI:10.1016/j.jhin.2018.12.011.

25. Li F, Liu W, Yu H. Extraction of information related to adverse drug events from electronic health record notes: Design of an end-to-end model based on deep learning. *JMIR Med Inform*. 2018 Nov 26;6(4):e12159. PMID:30478023. DOI:10.2196/12159.

26. Jeong E, Park N, Choi Y, Park RW, Yoon D. Machine learning model combining features from algorithms with different analytical methodologies to detect laboratory-event-related adverse drug reaction signals. *PLoS One*. 2018 Nov 21;13(11):e0207749. PMID:30462745. DOI:10.1371/journal.pone.0207749.

27. Santiso S, Perez A, Casillas A. Exploring joint AB-LSTM with embedded lemmas for adverse drug reaction discovery. *IEEE J Biomed Health Inform*. 2019 Sep;23(5):2148-2155. PMID:30403644. DOI:10.1109/JBHI.2018.2879744.

28. Chu J, Dong W, He K, Duan H, Huang Z. Using neural attention networks to detect adverse medical events from electronic health records. *J Biomed Inform*. 2018 Nov;87:118-130. PMID:30336262. DOI:10.1016/j.jbi.2018.10.002.

29. Wang SV, Maro JC, Baro E, et al. Data mining for adverse drug events with a propensity score-matched tree-based scan statistic. *Epidemiology*. 2018 Nov;29(6):895-903. doi: 10.1097/EDE.0000000000000907.

30. Martins RR, Silva LT, Bessa GG, Lopes FM. Trigger tools are as effective as non-targeted chart review for adverse drug event detection in intensive care units. *Saudi Pharm J*. 2018 Dec;26(8):1155-1161. doi: 10.1016/j.jsps.2018.07.003.

31. Whalen E, Hauben M, Bate A. Time series disturbance detection for hypothesis-free signal detection in longitudinal observational databases. *Drug Saf*. 2018 Jun;41(6):565-577. PMID:30074538. DOI:10.1007/s40264-018-0640-8.

32. Zhou X, Douglas IJ, Shen R, Bate A. Signal detection for recently approved products: Adapting and evaluating self-controlled case series method using a US claims and UK electronic medical records database. *Drug Saf*. 2018 May;41(5):523-536. PMID:29327136. DOI:10.1007/s40264-017-0626-y.

33. Nydert P, Unbeck M, Härenstam KP, et al. Drug Use and Type of adverse drug events-identified by a trigger tool in different units in a Swedish pediatric hospital. *Drug Healthc Patient Saf*. 2020 Jan 31;12:31-40. PMID:32099481. DOI:10.2147/DHPS.S232604. eCollection 2020.

34. Chen L, Gu Y, Ji X, Sun Z, Li H, Gao Y, Huang Y. Extracting medications and associated adverse drug events using a natural language processing system combining knowledge base and deep learning. *J Am Med Inform Assoc*. 2020 Jan 1;27(1):56-64. PMID:31591641. DOI:10.1093/jamia/ocz141.

35. Ju M, Nguyen NTH, Miwa M, Ananiadou S. An ensemble of neural models for nested adverse drug events and medication extraction with subwords. *J Am Med Inform Assoc*. 2020 Jan 1;27(1):22-30. PMID:31197355. DOI:10.1093/jamia/ocz075.

36. Griffey RT, Schneider RM, Todorov AA. Adverse events present on arrival to the emergency department: The ED as a dual safety net. *Jt Comm J Qual Patient Saf*. 2020 Apr;46(4):192-198. PMID:32007399. DOI:10.1016/j.jcjq.2019.12.003.

37. Pandya AD, Patel K, Rana D, et al. Global Trigger Tool: Proficient adverse drug reaction autodetection method in critical care patient units. *Indian J Crit Care Med*. 2020 Mar;24(3):172-178. PMID:32435095. DOI:10.5005/jp-journals-10071-23367.

38. McIsaac DI, Hamilton GM, Abdulla K, et al. Validation of new ICD-10-based patient safety indicators for identification of in-hospital complications in surgical patients: A study of diagnostic accuracy. *BMJ*

Qual Saf. 2020 Mar;29(3):209-216. PMID:31439760. DOI:10.1136/bmjqs-2018-008852. Epub 2019 Aug 22.

39. de Vos MS, Hamming JF, Chua-Hendriks JJC, Marang-van de Mheen PJ. Connecting perspectives on quality and safety: patient-level linkage of incident, adverse event and complaint data. *BMJ Qual Saf*. 2019 Mar;28(3):180-189. PMID:30032125. 10.1136/bmjqs-2017-007457.

40. Bates DW, Cullen DJ, Laird N, et al. Incidence of adverse drug events and potential adverse drug events implications for prevention. *JAMA*.1995; 274: 29-34. PMID:7791255. DOI:10.1001/jama.1995.03530010043033.

41. Kane-Gill SL, Kirisci L, Verrico MM, Rothschild JM. Analysis of risk factors for adverse drug events in critically ill patients. *Crit Care Med*. 2012; 40(3): 823–828. PMID:22036859. 10.1097/CCM.0b013e318236f473.

42. Bright RA, Rankin SK, Dowdy K, et al. Potential Blood Transfusion Adverse Events Can be Found in Unstructured Text in Electronic Health Records using the "Shakespeare Method". *MedRxiv* 2021;2021.01.05.21249239. DOI:10.1101/2021.01.05.21249239.

43. Baxter issues urgent nationwide voluntary recall of heparin 1,000 units/ml 10 and 30ml multi-dose vials *NDC NUMBERS 0641-2440-45, 0641-2440-41, 0641-2450-45 and 0641-2450-41; LOTS: 107054, 117085, 047056, 097081, 107024, 107064, 107066, 107074, 107111.* Food and Drug Administration. 2008 January 25. http://wayback.archive-it.org/7993/20170111131710/http://www.fda.gov/Safety/Recalls/ArchiveRecalls/2008/default.htm?Page=5.

44. Johnson AEW, Pollard TJ, Shen L, et al. MIMIC-III, a freely accessible critical care database. *Sci Data*. 2016; 3:160035. https://doi.org/10.1038/sdata.2016.35.

45. Code of Federal Regulations Title 45 Part 46 Protection of Human Subjects, Subpart A—Basic HHS Policy for Protection of Human Research Subjects, §46.101 (b) (4). 2000 Oct 1. https://www.govinfo.gov/content/pkg/CFR-2000-title45-vol1/pdf/CFR-2000-title45-vol1-part46.pdf.

46. Rankin SK, Bright R, Dowdy K. Bloatectomy (Version v0.0.12). *Zenodo*. 2020, June 26. http://doi.org/10.5281/zenodo.3909030.

47. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine Learning in Python. *J Mach Learn Res.* 2011; 12: 2825-2830. https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf.

48. Sklearn.feature_extraction.text.CountVectorizer. *Scikit-learn Machine Learning in Python*. Scikit-learn developers. 2020. http://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html.

49. Marafino BJ, Boscardin WJ, Dudley RA. Efficient and sparse feature selection for biomedical text classification via the elastic net: Application to ICU risk stratification from nursing notes. *J Biomed Inform*. 2015; 54: 114-120. PMID: 25700665. DOI:10.1016/j.jbi.2015.02.003.

50. Witten IH, Frank E, Hall MA, Pal CJ. *Data Mining: Practical machine learning tools and techniques*, 4th ed. Elsevier. 2016. Paperback ISBN: 9780128042915. eBook ISBN: 9780128043578.

51. Tang B, Kay S and He H. Toward optimal feature selection in naive Bayes for text categorization.*arXiv.* 2016; 1602: 02850. DOI:10.1109/TKDE.2016.2563436.

52. Blei D, Ng A Jordan M. Latent Dirichlet Allocation. *J Mach Learn Res.* 2003;3:993-1022**.** https://jmlr.org/papers/volume3/blei03a/blei03a.pdf.

53. LINEST function. *Microsoft Support*. 2020. https://support.microsoft.com/en-us/office/linest-function-84d7d0d9-6e50-4101-977a-fa7abf772b6d.

54. Altman DG, Bland JM. How to obtain the P value from a confidence interval. *BMJ*. 2011;343:d2304. PMID: 22803193. DOI:10.1136/bmj.d2304.

55. Heparin sodium- heparin sodium injection, solution: Drug label information. *DailyMed*. U.S. National Library of Medicine. 2020. https://dailymed.nlm.nih.gov/dailymed/drugInfo.cfm?setid=cb1c1e7a-c9ca-4a07-8833-e45ce436d287.

56. Bassereo PP, Cocco D, Bassareo V, et al. Pharmacological treatment of vagal hyperactivity, a rare but potentially fatal cause of sudden cardiac death. *Mini Rev Med Chem*. 2018;18(6):483-489. PMID:28685699. DOI:10.2174/1389557517666170707102040.

57. Information on heparin. Food and Drug Administration. 2017. https://wayback.archive-it.org/7993/20170722214801/https://www.fda.gov/Drugs/DrugSafety/PostmarketDrugSafetyInformationforPatientsandProviders/UCM112597.

58. Acute allergic-type reactions among patients undergoing hemodialysis — Multiple states, 2007–2008. *MMWR*. 2008, February 8; 57(5): 124-125. PMID: 18256585. https://www.cdc.gov/mmwr/preview/mmwrhtml/mm5705a4.htm.

59. Lyn TE. China pig disease caused by new strain: experts. *Reuters*. 2007 June 26. https://www.reuters.com/article/us-china-disease-pig-idUSHKG26819620070626.

60. Barboza D. Virus Spreading Alarm and Pig Disease in China. *New York Times*. 2007, August 16. http://www.nytimes.com/2007/08/16/business/worldbusiness/16pigs.html.

61. Tian K, Yu X, Zhao T, et al. Emergence of fatal PRRSV variants: unparalleled outbreaks of atypical PRRS in China and molecular dissection of the unique hallmark. *PLoS ONE*. 2007;2(6):e526. PMID: 17565379. DOI:10.1371/journal.pone.0000526.

62. Levy P. The Harvard medical system. *Not Running a Hospital*. 2007, January 14. http://runningahospital.blogspot.com/2007/01/harvard-medical-system.html.

63. Schofield A, Magnusson M, Mimno D. Pulling out the stops: rethinking stopword removal for topic models. IN: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, Valencia, Spain, April 3-7, 2017*. Association for Computational Linguistics. 2017; 432–436. https://www.aclweb.org/anthology/E17-2069.

64. Halamka J. What will keep me up at night. *Dispatch from the digital health frontier*. 2007 Nov 1, 2, 19, 20. http://geekdoctor.blogspot.com/2007/11/.

Table 1. Preprocessing/Model Parameters.

Phase	Parameter	Value/Result
Preprocessing	Punctuation and Digit Removal	No
	Lowercase	Yes
	Stop (common) Word Removal	No
	Duplicate Text Removal (Bloatectomy)	Version 2.1
Vectorization	Vectorization Type	Frequency
	Range of *n*-grams	1 through 5
	Maximum Number of Features (terms) to Keep	No Limit
	Minimum Document Frequency (keep terms appearing in at least X documents)	2
	Collocation Detection	Yes
	Threshold (for collocation)* *represents a threshold for forming the phrases (higher means fewer phrases). A phrase of words a and b is accepted if (count(a, b) - min_doc_freq) * N / (count(a) * count(b)) > threshold	2
	Total Number of Features	4,266,455
Feature extraction	Classification Models	• Multinomial Naïve Bayes • Logistic Regression
	Regularization Methods	L2 (Ridge), L1 (Lasso)
	Number of Top Features Extracted (per classification model)	5000
Topic modeling	Topic Modeling	LDA – Sci-kit Learn
	Number of Topics	20
	Vocabulary (significant features)	6,122
	Iterations	15

Table 2. Topics sorted by the maximum word score in the topic, with the top 20 substantive words, the maximum topic match score among admissions, and distribution of the topic match scores among admissions. "Substantive" words had topic scores above the minimum topic score. "Max" means maximum.

Topic				# Admissions in topic score range				
Topic #	Top 20 substantive terms	Max word score	Max topic match score among admissions	≥0.03	≥0.50	≥0.2 to <0.5	≥0.1 to <0.2	≥0.03 to <0.1
18	for, hr, plan, vent, intubated, cont, today, skin, are, family, per, support, increased, off, goal, iv, placed, trach, foley, pain	75372	0.99	1793	505	623	326	339
3	for, hr, pain, bp, are, you, iv, family, time, ccu, per, sats, note, heart, micu, received, skin, if, acute, plan	42070	1.0	2224	912	697	328	287
19	for, are, pain, you, comparison, acute, upper, evaluate, iv, trauma, hospital, if, note, time, large, level, pleural, wbc, read, throughout	39731	1.0	2089	355	880	468	386
7	for, are, pain, pleural, cabg, hr, plan, per, comparison, off, bp, pericardial, time, neo, iv, heart, md, mm, mr, catheter	30722	1.0	1686	589	321	319	457
1	for, are, family, subarachnoid, mm, comparison, pain, iv, occipital, sdh, large, evaluate, plan, cont, acute, craniotomy, per, hr, note, goal	12352	1.0	749	181	235	118	215
4	catheter, pleural, for, pain, jp, [pain-reliever], placed, large, into, pigtail, hr, cont, french, increased, are, pseudoaneurysm, upper, skin, iv, comparison	3523	0.54	683	1	75	180	427
17	for, are, mca, into, time, catheter, arteriogram, occlusion, mm, acute, french, ica, iv, placed, territory, large, cont, comparison, goal, family	3462	0.77	534	39	99	127	269
12	[condition01], section, gynecology, [condition02], dystrophy, cesarean, [anti-thyroid], transabdominal, [event01], lmp, wk, [procedure01], [progesterone], prenatal, [condition03], [condition04], [antispasmodic], enteropathy, [condition05], [condition06]	216	0.22	31	0	1	7	23
11	pentobarb, pentobarbital, cmv, encasement, prison, [condition07], satellite, hematologic, rent, [condition08], [condition09a], [condition09b], [antibiotic], federal, bleach, [device01], allergic, [rare-word01], cluster, [rare-word02]	75	0.11	26	0	0	1	25
5	[rare words, misspelled words]	63	0.05	1	0	0	0	1
15	[rare words, misspelled words]	36	0.13	2	0	0	2	0
16	[rare words, misspelled words]	15	0.11	2	0	0	1	1
6	[rare words, misspelled words]	14	0.02	0	0	0	0	0
10	[rare words, misspelled words]	11	0.06	2	0	0	0	2
0	[rare word]	10	0.04	1	0	0	0	1
2	[rare words, foreign language words, misspelled words]	9	0.12	3	0	0	1	2
14	[rare words, misspelled words]	8	0.03	1	0	0	0	1

Topic				# Admissions in topic score range				
Topic#	Top 20 substantive terms	Max word score	Max topic match score among admissions	≥0.03	≥0.50	≥0.2 to <0.5	≥0.1 to <0.2	≥0.03 to <0.1
9	[rare words, misspelled words]	7	0.07	2	0	0	0	2
13	[rare words, misspelled words]	6	0.06	3	0	0	0	3
8		0	0.00	0	0	0	0	0

Table 3. Summaries of the admissions with the top three topic match scores, for the most common topics. "HD" is hospital day. "Intubated" and "extubated" refer to starting and ending mechanical ventilation.

Topic #: Top 20 substantive terms	
Summary of records with top 3 topic scores	**Comment**
Topic 18: for, hr, plan, vent, intubated, cont, today, skin, are, family, per, support, increased, off, goal, iv, placed, trach, foley, pain	
Admitted on hospital day 1 (HD1) from other hospital, with end stage liver disease, now short of breath. Intubated. Pneumonia. Developed bacteremia. Coagulopathy and anemia due to liver. HD29 severe hypotension, extubated, comfort measures only, died.	Long complex stay.
Admitted HD1 after CPR and intubation. No anticoagulants given; edema. Anemic initially. HD2 multiple chest fractures from CPR. HD1 to HD4, HD7 to HD15, HD16 to HD33 intubated; HD34 to HD35 O2 mask. HD2, HD35 hypotension. HD2 to HD36 AF. HD2 to HD7 pulmonary edema. HD9 surgery on spine; 4 units blood; postoperation hypotension. HD13, HD16, HD20, HD25, HD31 blood transfused. HD16 platelets dropped, stayed low despite removal of all heparin and heparin lines and despite daily platelet transfusions on HD16 to HD34. HD25 HIT+. HD18 edema increased. HD29 bone marrow biopsy. Died HD36.	Long complex stay.
Admitted HD1 with ongoing anemia. Diagnosed leukemia. HD24 tooth pain, extracted, followed by intense pain and treated with antibiotics. Hepatitis B and C diagnosed. Blood transfusions. Chemotherapy. Bacteremia diagnosed and treated; other infections diagnosed over time; several antibiotics tried. HD49 bone marrow transplant. [Immunosuppressant] started, seemed to cause hypertension. HD76 to HD110 intubated. HD76 pulmonary edema. Progressive renal failure, treated with continuous dialysis HD88 to HD98. Died HD111.	Long complex stay.
Topic 3: for, hr, pain, bp, are, you, iv, family, time, ccu, per, sats, note, heart, micu, received, skin, if, acute, plan	
Admitted HD1 for chest pain. Inserted stent; then worse heart beat profile. HD2 went to CCU. Kidney worsened. Discharged HD8.	Heart attack, cardiac catheterization, heparin AE.
Transfused monthly before admission. Admitted HD1 for declined mental status; diagnosed heart attack; started aspirin; not a cardiac catheterization candidate. HD1 started breathing difficulty; new tachycardia. HD2 pulmonary edema; hypotension observed and treated; AF; mask O2. HD3 to HD5 hypotension. HD3 to HD5 pulmonary edema. HD4 to HD? given blood. HD2 to HD4 given heparin. HD1 to HD4 fever. Discharged HD6.	Heart attack, (not) cardiac catheterization, heparin AE.
Admitted HD1 for hypoxia and acute renal failure; diagnosed renal cysts, new AF, urinary tract infection (UTI). HD2 no further SF. HD1 to HD6 UTI. HD? ARF resolved. Discharged HD10.	Hypoxia and kidney failure.
Topic 19: for, are, pain, you, comparison, acute, upper, evaluate, iv, trauma, hospital, if, note, time, large, level, pleural, wbc, read, throughout	
Admitted HD1 for [event02]. Diagnosed ~~rib and spine~~ fractures, dislocations, and muscle injury. Treated with [device02], [device03] and [device04]. Started antihypertensive drug. Discharged HD10 to home.	Bone trauma.

Topic #: Top 20 substantive terms	
Summary of records with top 3 topic scores	**Comment**
[Event03] and went to other hospital. HD1 transferred to BIDMC; diagnosed with fractures; treated with [device02]. Diagnosed chronic kidney failure. Discharged HD4 to home.	Bone trauma, kidney failure.
Admitted HD1 for abdominal pain from [event04]. Diagnosed spleen laceration. Discharged HD3.	Spleen trauma.
Topic 7: for, are, pain, pleural, cabg, hr, plan, per, comparison, off, bp, pericardial, time, neo, iv, heart, md, mm, mr, catheter	
Admitted HD1 for shortness of breath; to get cardiac catheterization. HD4 heart valve replaced; then hypotension treated with [phenylephrine] until next day. HD5 to HD8 heart rhythm abnormal. POD4 moved to step down. HD11 discharged.	Heart failure, cardiac catheterization, heparin AE.
HD1 admitted for heart problem. Started heparin. HD4 replaced heart valve, CABG, placed intra-aortic balloon pump (IABP), started epinephrine, started levophed, gave blood. HD5 stopped IABP, epinephrine, levophed. HD6 new AF. HD7 went to floor. HD12 discharged.	Heart failure, cardiac catheterization, heparin AE.
HD1 admitted for heart surgery; CABG; AF during operation; given [phenylephrine]; ongoing diabetes mellitus, type 2 (DMII). HD2 went to floor. HD3 went to CCU to restart insulin drip. HD5 went to floor. HD9 discharged.	Heart failure, cardiac catheterization, heparin AE.
Topic 1: for, are, family, subarachnoid, mm, comparison, pain, iv, occipital, sdh, large, evaluate, plan, cont, acute, craniotomy, per, hr, note, goal	
HD1 admitted for headache and confusion; diagnosed brain bleed. Blood removed in operating room (OR). HD2 moved to floor. HD4 same place in brain seen to still bleed; in OR removed new blood and stopped bleeding. HD5 went to floor. HD10 discharged.	Brain bleed, brain surgery.
Admitted HD1 for brain surgery. HD2 went to floor. Discharged HD3.	Brain surgery.
HD1 transferred from other hospital that diagnosed brain bleed. HD2, HD3 surgery to remove blood from brain. HD6 went to floor. HD8 Magnetic resonance image (MRI), then seizures for an hour and moved to CCU. HD13, HD14 seizure-free. HD15.	Brain bleed, brain surgery.
Topic 4: catheter, pleural, for, pain, jp, [pain reliever], placed, large, into, pigtail, hr, cont, french, increased, are, pseudoaneurysm, upper, skin, iv, comparison	
HD1 admitted for large drainage from surgery; given fluids; drain replaced. HD4 peripherally inserted central catheter (PICC) line placed for intravenous (IV) fluid. HD25 sclerotherapy to try to stop the leaking. HD34 surgery to stop leak. HD45 stent placed; drainage decreased. HD52 stent migrated and was removed; more stents placed. HD56 discharged to home to care for continuing drainage.	Extensive prolonged drainage after abdominal surgery.

Topic #: Top 20 substantive terms	
Summary of records with top 3 topic scores	**Comment**
HD1 admitted due to [condition10] diagnosed at other hospital; pulmonary emboli and deep vein thrombosis (DVT); heparin started and stopped; AF. HD2 venous filter placed; bilateral pleural effusions; pulmonary embolism, DVT. HD3 catheter inserted in [condition10], turned out to be infected; started antibiotic. HD4 needed extra fluid; continuing AF; edema; stopping heparin. HD5 started O2 mask; AF; edema; pleural effusions bigger; pleural drain placed. HD6 catheter upsized; AF; edema improved. HD7 heparin; AF. HD10 catheter upsized. HD20 discharged to extended care; needs heparin lock flushes of catheter.	Extensive drainage of abdominal infection, already had pleural effusions, thrombi, and AF.
HD1 transferred from other hospital for [condition11] and nearby fluid removal. HD2 inserted drain. HD4 new drain inserted. HD5 stent. HD8 another drain. HD13 fluids decreased; pleural effusions decreased. HD18 hypotensive; septic; drainage increased; antibiotics started. HD23 catheter repositioned. HD26 discharged.	Extensive drainage of abdominal organ, infection.
Topic 17: for, are, mca, into, time, catheter, arteriogram, occlusion, mm, acute, french, ica, iv, placed, territory, large, cont, comparison, goal, family	
HD1 transferred from other hospital for stroke; had received tissue plasminogen activator (tPA). HD2 large brain bleeds. HD3 died.	Brain ischemia; brain bleed.
HD1 transferred from other hospital where taken for signs of stroke; diagnosed brain arteries blocked. HD1 intubated; catheterization lab cleared thrombus; placed stent; antihypertensive after. HD2 extubated. HD9 discharged.	Brain ischemia.
HD1 admitted for stroke symptoms; given tPA; an hour intubated and sedated, stented. HD2 extubated. HD3 went to floor. HD8 discharged.	Brain ischemia.
Topic 12: [condition01], section, gynecology, [condition02], dystrophy, cesarean, [anti-thyroid], transabdominal, [event01], lmp, wk, [procedure01], [progesterone], prenatal, [condition03], [condition04], [antispasmodic], enteropathy, [condition05], [condition06]	
HD1 admitted for abdominal pain; diagnosed with [condition01]; [device05] placed. HD2 [device05] removed; [condition01] resolved. HD3 [condition01] returned; [device05] placed again. HD5 suddenly needed O2. HD6 to HD8 antibiotics for UTI. HD11 discharged.	[Condition01].
HD1 admitted for [condition01]; open surgery to resolve it; also [procedure02]. HD10 discharged.	[Condition01].
HD1 admitted for distressing symptoms; diagnosed [condition01]. [Procedure03] temporarily resolved the condition. HD5 [Procedure04] resolved [condition10]. HD13 discharged.	[Condition01].

Table 4. Summaries of admissions (top scoring for 20-11 and all for the other topics) with topic matching scores for the less common topics. "Unusual" means there were a few or some instances in period 1. "Rare" means there were no instances in period 1.

Topic #: top 20 substantive terms			
[Topic match score] Brief summary of text	**Topic fit**	**AE type**	**Text offers more AE data than codes?**
Topic 11: pentobarb, pentobarbital, cmv, encasement, prison, [condition07], satellite, hematologic, rent, [condition08], [condition09a], [condition09b], [antibiotic], federal, bleach, [device01], allergic, [rare-word01], cluster, [rare-word02]			
[0.11] Admitted HD1 due to [event05]. Given anaphylaxis meds. Discharged HD2.	"Allergic" is more common in the post period.	Non-medical anaphylaxis.	No
[0.08] Admitted HD1 with recent [event06]. Diagnosed [condition9b]. Started antibiotic and developed [condition12], thought to be [condition13] and treated with drugs for [condition13]. Also given [antiviral]. Diagnosed [condition14] so stopped prior antibiotics. In CCU started therapy for [condition14]. HD5 given PICC line for that therapy. Discharged HD6.	"[Condition9a]" and "[condition9b]" are both rare in the text.	Medical therapy AE.	Yes
[0.08] Admitted HD1 due to [event07]. HD2 procedure resolved [event07]; discharged. [No discharge summary]	"[Event07]" is rare in the text.		
[0.08] Admitted HD1 for surgery for [condition07]; had surgery and chest tubes. HD2 transferred to floor. Tubes gradually removed. Discharged HD13.	"[Condition07]" is unusual in text.		
[0.08] Admitted HD1 due to [condition07]; had surgery, drainage tubes, and epidural. HD2 postsurgical [device06] malfunctioned so replaced; tachycardia; on heparin prophylaxis. HD3 replacement [device06] failing the same way as the first. HD5 epidural removed. HD7 one of JP drains removed. Discharged HD8.	"[Condition07]" is unusual in text.	Medical therapy AEs.	Yes, and more information in daily notes than discharge summary.
Topic 5: [rare words, misspelled words]			
[0.05] Admitted HD1 from other hospital for [condition15] and [condition16]. The latter gradually decreased during the stay. HD7 new [condition17]. Discharged HD18. New [condition17] thought to be side effect of combination of therapies; new [condition17] gradually improved.	"[Condition15]" is unusual in the text.	Medical therapy AE.	Codes indicate the outcome, but not the speculated causes.
Topic 15: [rare words, misspelled words]			
[0.13] Admitted HD1 due to reaction to [drug01]. Treated. Discharged HD2.	"[Drug01]" is an unusual word in text.	Medical therapy AE.	Codes say reaction to named drug.
[0.13] Admitted HD1 for distressing symptoms. Diagnosed many conditions. Family refused aggressive treatment. HD13 went to floor. Discharged HD16.	"[Rare-word03]" is rare in the text.		

Topic #: top 20 substantive terms			
[Topic match score] Brief summary of text	**Topic fit**	**AE type**	**Text offers more AE data than codes?**
Topic 16: [rare words, misspelled words]			
[0.11] Admitted HD1 due to [condition18a and b] known DM1 and hypothyroid; [procedure01] done to treat. HD2 orthostatic hypotension. HD3 blood pressure fine; discharged.	"[Condition18a]" and "[condition18b]" are both rare in the text.		
[0.05] Admitted HD1 from outpatient clinic where had received [drug02], then needed rescue. AE attributed to [drug02]. Continued to improve. Discharged HD3.	"[Drug02]" is rare in the text.	Medical therapy AE.	There is a code for this AE, but not clear which of the other codes is the actual AE.
Topic 10: [rare words, misspelled words]			
[0.06] Admitted HD1. [Only 2 notes, both nursing]. Bleeding in brain. Discharged HD2.	"[Rare-word04]" is rare in text and was quoting the patient.		
[0.04] Admitted HD1 incoherent, prior [condition19a and b], other diseases. Has pneumonia. HD2 coherent. Discharged HD7.	"[Condition19a]" and "[condition19b]" are rare in the text.		
Topic 0: [rare word]			
[0.04] Admitted HD1 from other hospital for [event08] following [procedure05]. Started antibiotics. HD2 improved; moved to floor. Discharged HD6.	"[Procedure05]" is unusual in the text.	Medical therapy AE.	Codes specify drug AE, and hint, but don't specify the surgical procedure.
Topic 2: [rare words, foreign language words, misspelled words]			
[0.12] [Patient instructions are in foreign language. Rest of record is in English.]	Foreign language words are unusual or rare.	Medical therapy AE.	Outcomes are in the billing codes but codes don't indicate that one might be a medical therapy AE.
[0.08] [Patient instructions are in foreign language. Rest of record is in English.]	Foreign language words are unusual or rare.		
[0.03] [No foreign language.] Admitted HD1 for [event09] that resulted from loss of consciousness; several fractures and cuts. Cuts sewn. [Device02], [device07] and [device08]; given. Discharged HD4.	"[Rare-word05]" and "[rare-word06]" are rare in the text.		
Topic 14: [rare words, misspelled words]			

Topic #: top 20 substantive terms			
[Topic match score] Brief summary of text	**Topic fit**	**AE type**	**Text offers more AE data than codes?**
[0.03] Admitted HD1 for [procedure06] for [condition20]. POD2 went to floor. POD9 new fever; diagnosed bacterial infection in surgical drainage; given antibiotics. Fistula noted. Discharged HD43.	"[Procedure06]" and "[condition20]" are rare words in text.	Medical therapy AEs.	The order and consequences of events are not noted by the codes.
Topic 9: [rare words, misspelled words]			
Admitted HD1 to treat [condition14]; PICC inserted; treatment was uneventful. HD2 began next phase of therapy; went to floor. HD5 started warmth and redness near PICC line insertions site. HD6 worse; no thrombus found; treated with warm compresses. Discharged HD8 to extended care for continued therapy. Heparin for prophylaxis.	"[Condition14]" unusual in text.	Medical therapy AE.	In discharge summary but not in codes.
Admitted HD1 with recent[event06]. Started antibiotic and developed [condition12], thought to be [condition13], and treated [condition13]. Also given antiviral. Diagnosed [condition14] so stopped prior antibiotics. In CCU started [condition14] therapy. HD5 given PICC line for continued treatment. Discharged HD6.	"[Condition14]" unusual in text.	Medical therapy AE.	In discharge summary but not in codes.
Topic 13: [rare words, misspelled words]			
HD1 had [event10a]; no fracture. Precautionary CCU, showed no issue. HD1 discharged. [No discharge summary.]	"[Event10a]" and "[event10b]" are rare words in the text.		
Admitted HD1 for lightheaded, stomach pain. Anemia. Diagnosed upper gastrointestinal tract ulcer and treated; may have been worsened because of prescribed high [blood thinner] doses. Transfused. Discharged HD6 to home.	"[Rare-word07]" is rare in the text.	Medical therapy AE.	No
Admitted HD1 due to [event02] and lost consciousness; multiple fractures, brain bleed. Discharged HD5.	"[Rare-word08]" is rare in the text.		

Figure 1

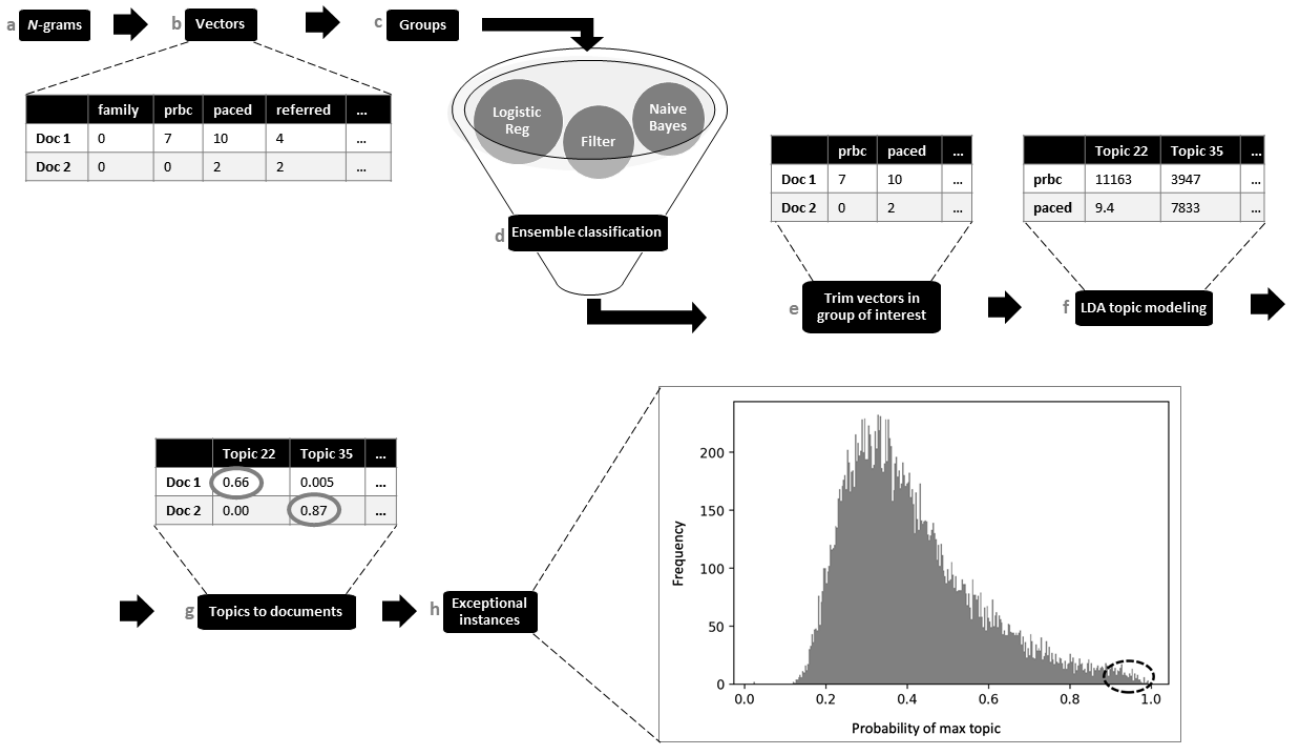


Figure 1. Word selection and topic modeling process with truncated examples.

Figure 2

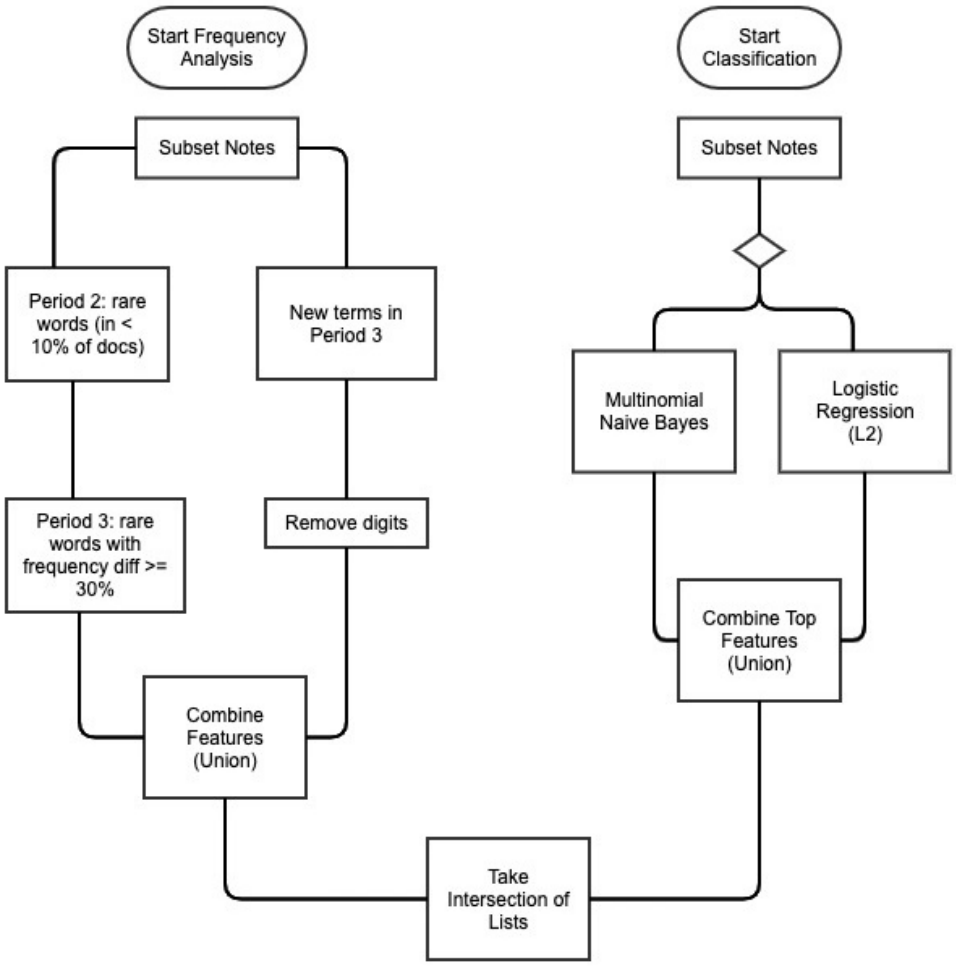


Figure 2. Feature extraction flowchart. This demonstrates the two parallel processes for extracting relevant features prior to topic modeling on the notes: term frequency analysis and binary classification of notes.

Figure 3

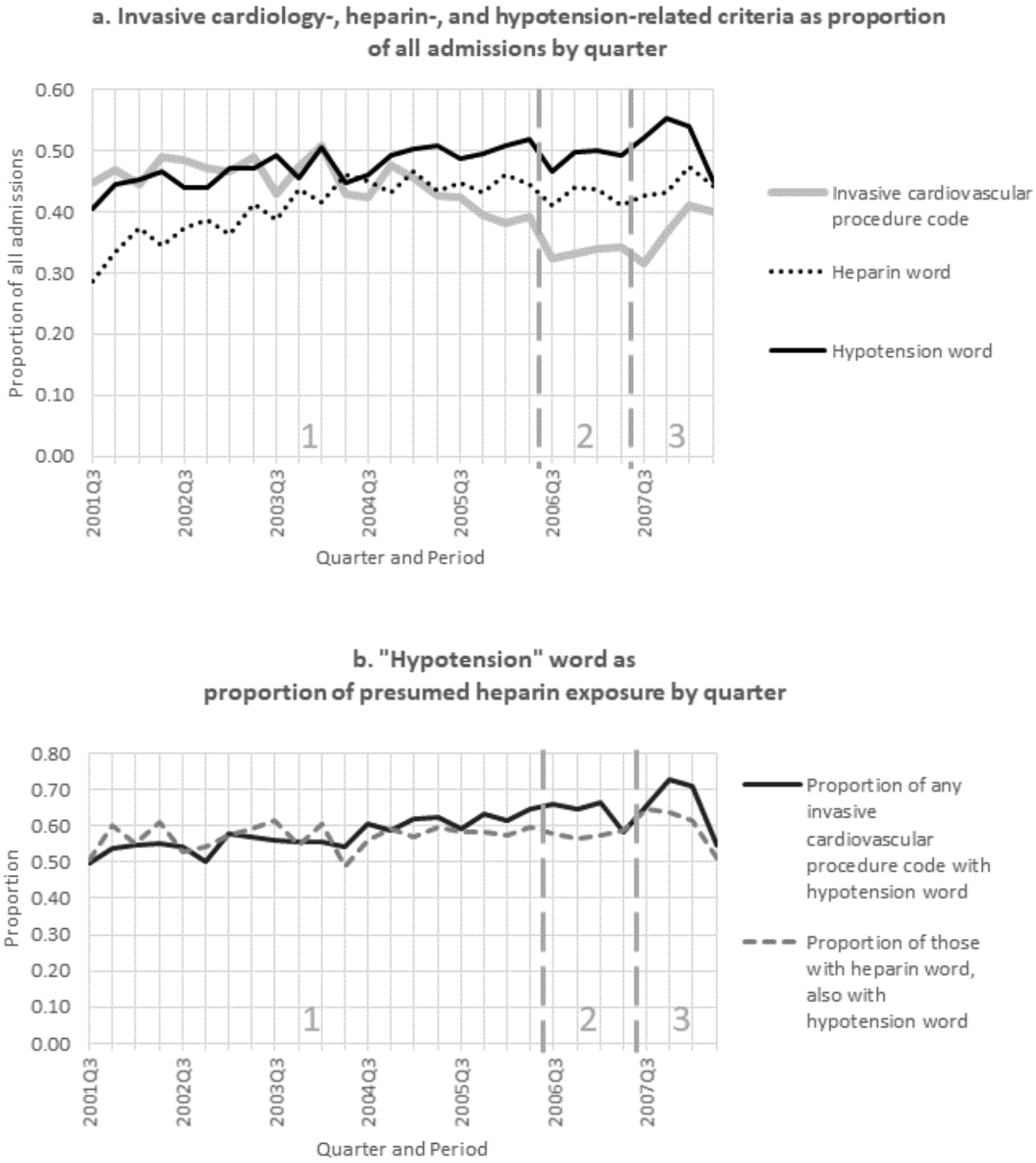


Figure 3. Heparin and hypotension.
Figure 3a. Invasive cardiology-, heparin-, and hypotension-related criteria as proportion of all admissions. Invasive cardiology is presumed to involve heparin treatment. The definitions are listed in "Results eFigures 2 to 6 and figures info". For invasive cardiovascular procedure code, slope = -0.0053

(95%CL -0.0069 to -0.0037, p<0.0001), for heparin word slope = 0.0039 (95% CI 0.0025 to 0.0054, p<0.0001), and for hypotension word slope = 0.0029 (95%CL 0.0017 to 0.0040, p<0.0001).
3b. "Hypotension" word as proportion of presumed heparin exposure. For proportion of any invasive cardiovascular procedure code (presumed to involve heparin), slope = 0.0055 (95%CL 0.0038 to 0.0072, p<0.0001). For proportion of those with "heparin", slope = 0.0013 (95%CL -0.00036 to 0.0030, p=0.12).

Figure 3 notes:

- Invasive cardiovascular code:
  - 3891 Arterial catheterization
  - 3961 Extracorporeal circulation auxiliary to open heart surgery
  - [3965 to 3966]
- "Heparin" word
- "Hypotension" word

Figure 4

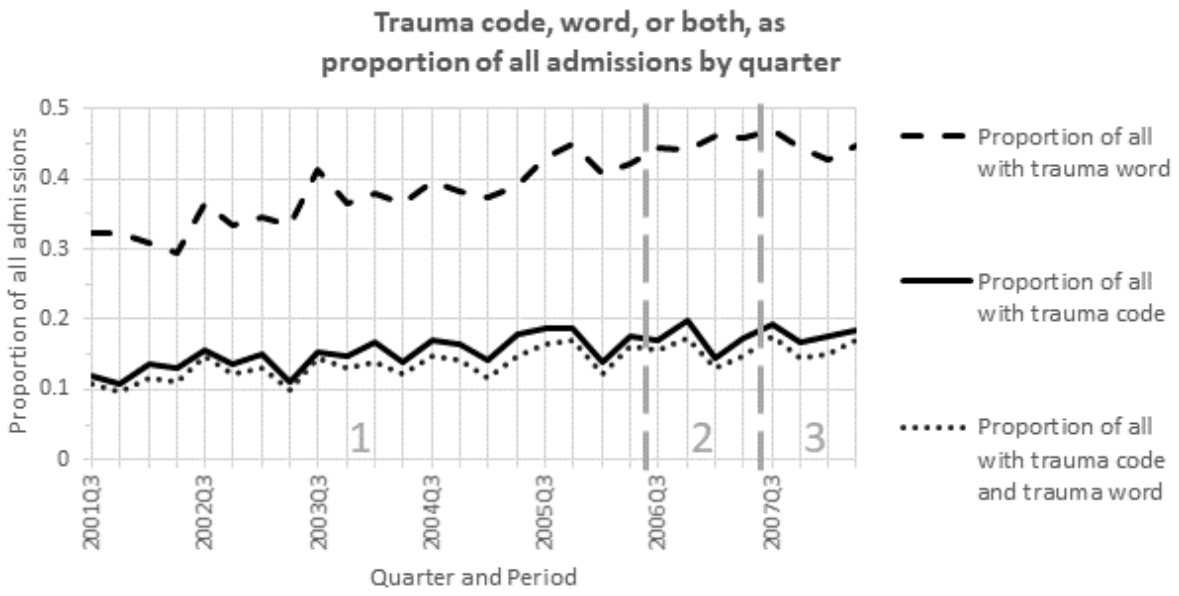


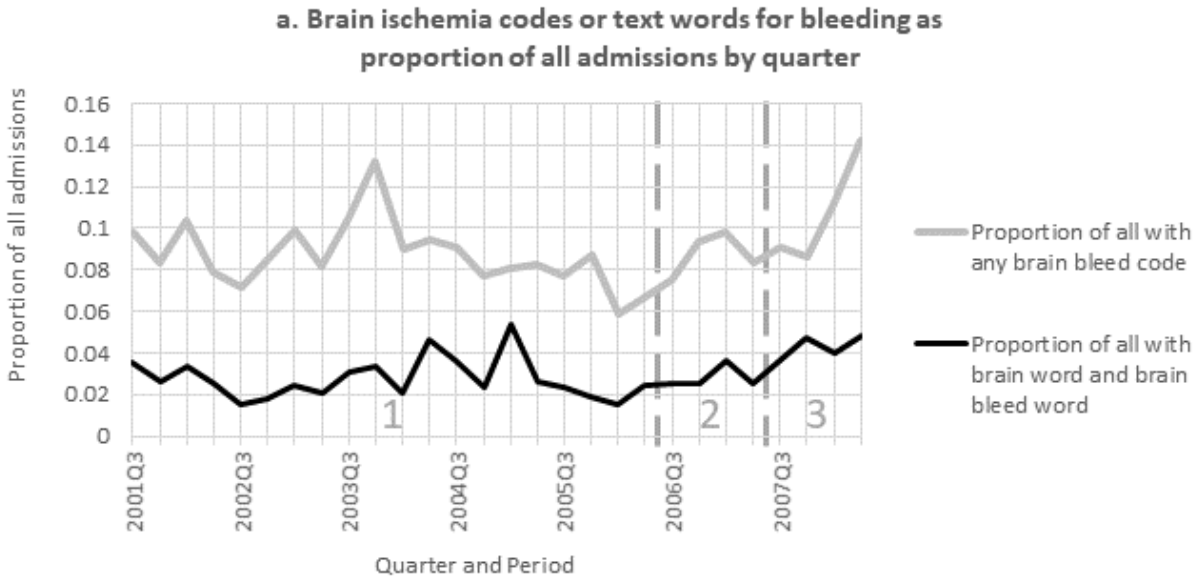
Figure 4. Trauma code, word, or both as proportion of all admissions by quarter. For proportion with trauma code, slope=0.0022 (95%CL 0.0014 to 0.0030), p<0.0001. For proportion with trauma word, slope=0.0057 (95%CI 0.0047 to 0.0067), p<0.0001. For proportion with both trauma code and word, slope=0.0019 (95% CL 0.0012 to 0.0027), p<0.0001.

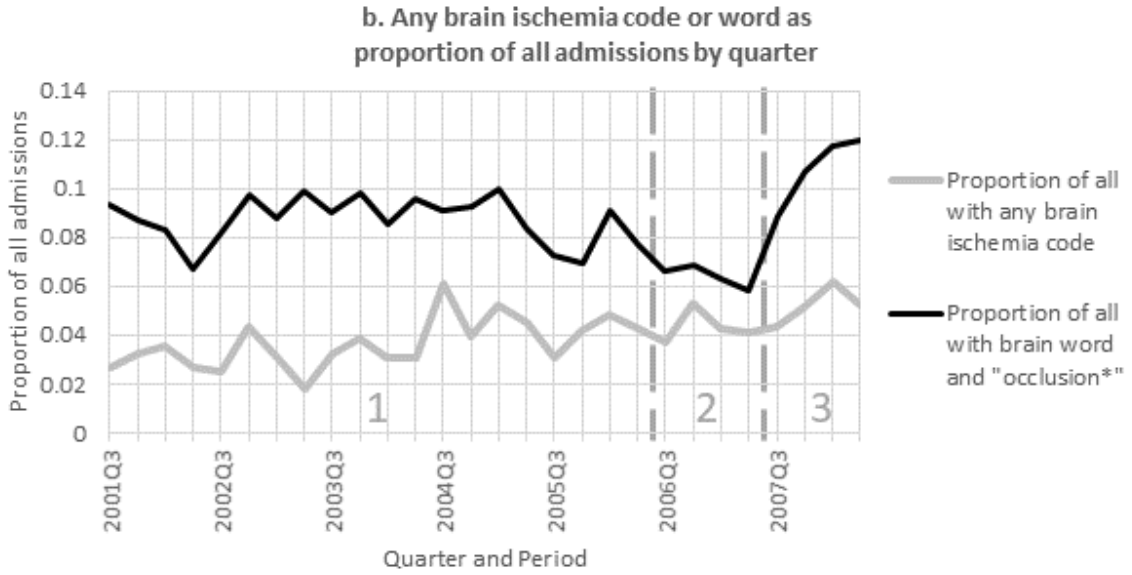
Figure 4 notes:

- Trauma code, any of:
  - Diagnosis code 800* to 829* [Fracture]
  - Diagnosis code 830* to 869* [Dislocations, sprains, strains, and internal injury of cranium, chest, abdomen, pelvis]
  - Diagnosis code 870* to 897* [Open wound]
  - Diagnosis code 900* to 904* Injury to blood vessels
  - Diagnosis code 905* Late effects of musculoskeletal and connective tissue injuries
  - Diagnosis code 9060 to 9064 [Late effects of open wound, superficial injury, contusion, or crushing]
  - Diagnosis code 907* Late effects of injuries to the nervous system
  - Diagnosis code 908* Late effects of other and unspecified injuries
  - Diagnosis code 910* to 924* [Superficial injury and contusion with intact skin surface]
  - Diagnosis code 925* to 929* Crushing injury
  - Diagnosis code 950* to 957* Injury To Nerves And Spinal Cord
  - Diagnosis code 958* to 959* Certain Traumatic Complications And Unspecified Injuries
  - Procedure code [7670 to 7679]
  - Procedure code [7810 to 7819]
  - Procedure code [7900 to 7939]
  - Procedure code [7960 to 7969]
  - Procedure code [7990 to 7999]
- Trauma word, any of:
  - trauma
  - mva

- fall
- mvc
- contusion
- fracture

Figure 5


a. Brain ischemia codes or text words for bleeding as proportion of all admissions by quarter
Proportion of all admissions
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
1
2
3
2001Q3
2002Q3
2003Q3
2004Q3
2005Q3
2006Q3
2007Q3
Quarter and Period
Proportion of all with any brain bleed code
Proportion of all with brain word and brain bleed word


b. Any brain ischemia code or word as proportion of all admissions by quarter
Proportion of all admissions
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
1
2
3
2001Q3
2002Q3
2003Q3
2004Q3
2005Q3
2006Q3
2007Q3
Quarter and Period
Proportion of all with any brain ischemia code
Proportion of all with brain word and "occlusion*"

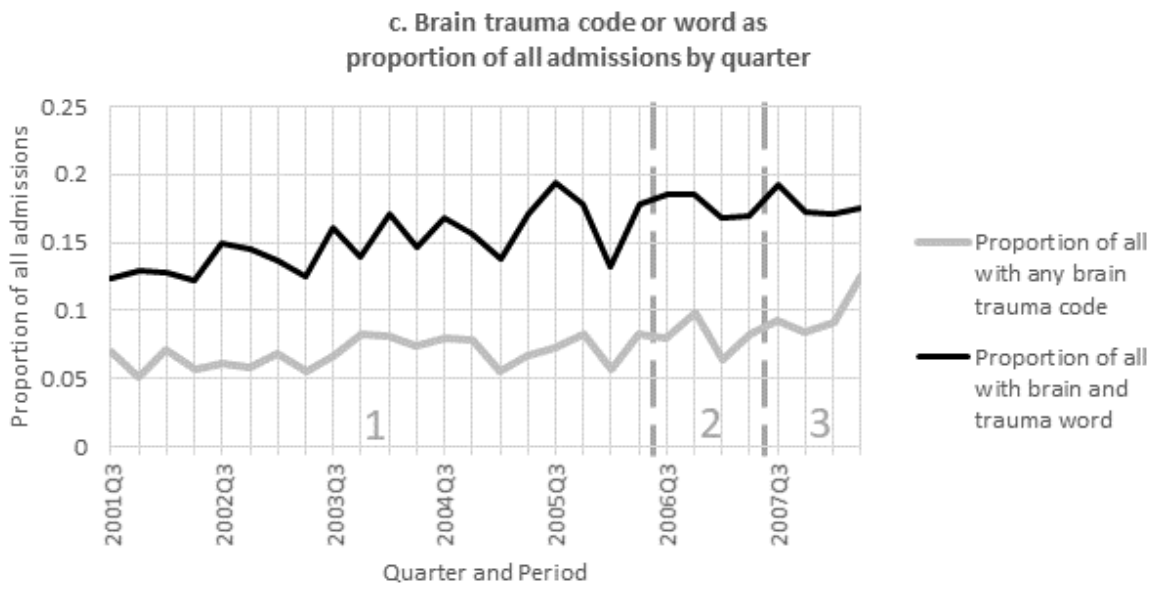


Figure 5. Brain ischemia codes or text words for a. bleeding, b. ischemia, and c. trauma, as proportion of all admissions by quarter. For brain bleed code, slope=0.00022 (95%CI -0.0006 to 0.0010), p=0.61. For brain word and brain bleed word, slope= 0.00039 (95%CI 0 to 0.00085), p=0.10. For brain ischemia code, slope-0.00019 (95%CI 0.00051 to 0.0013), p<0.0001. For brain word and "occlusion*", slope = 0 (95%CI -0.00064 to 0.00080), p=0.84. For brain trauma code, slope=0.0013 (95%CI 0.00073 to 0.0018), p<0.0001. For brain word and "trauma", slope=0.0021 (95%CI 0.0014 to 0.0028), p<0.0001.

Figure 5 notes:

- Brain bleed code is any of:
    - Diagnosis code 3481* Anoxic brain damage
    - Diagnosis code 3484* Compression of brain
    - Diagnosis code 430* Subarachnoid hemorrhage
    - Diagnosis code 431 Subarachnoid hemorrhage
    - Diagnosis code 432 Other and unspecified intracranial hemorrhage
    - Diagnosis code 8042* Closed fractures involving skull or face with other bones with subarachnoid subdural and extradural hemorrhage
    - Diagnosis code 8043* Closed fractures involving skull or face with other bones, with other and unspecified intracranial hemorrhage
    - Diagnosis code 8047* Open fractures involving skull or face with other bones with subarachnoid subdural and extradural hemorrhage
    - Diagnosis code 8048* Open fractures involving skull or face with other bones with other and unspecified intracranial hemorrhage
    - Diagnosis code 852** Subarachnoid subdural and extradural hemorrhage following injury
    - Diagnosis code 853** Therapeutic and unspecified intracranial hemorrhage following injury
    - Procedure code 109 Other cranial puncture
    - Procedure code 110 Intracranial pressure monitoring
    - Procedure code 116 Intracranial oxygen monitoring
    - Procedure code 121 Incision and drainage of cranial sinus
    - Procedure code 123 Reopening of craniotomy site
    - Procedure code 124 Other craniotomy

    - Procedure code 125 Other craniectomy
    - Procedure code 3881 Other surgical occlusion of vessels, intracranial vessels
- Brain bleed word, any of: “IPH”, “aneurysms”, or “embolize”
- Brain word, any of:
    - *occipital
    - *cranio*
    - *cepha*
    - mening*
    - *frontal
    - *tempero*
    - *pariet*
    - brain
    - *arachnoid
    - mca
    - hemiparesis
    - hemiplegia
- Brain ischemia code is any of:
    - Diagnosis code 3481* Anoxic brain damage
    - Diagnosis code 434** Occlusion of cerebral arteries
    - Diagnosis code 435** Transient cerebral ischemia
    - Diagnosis code 4371* Other generalized ischemic cerebrovascular disease
    - Diagnosis code 4376* Nonpyogenic thrombosis of intracranial venous sinus
    - Procedure code 62 Percutaneous angioplasty of intracranial vessel(s)
    - Procedure code 65 Percutaneous insertion of intracranial vascular stent(s)
    - Procedure code 116 Intracranial oxygen monitoring
    - Procedure code 1754 Percutaneous atherectomy of intracranial vessel(s)
    - Procedure code 3811 Endarterectomy, intracranial vessels
- Brain trauma code is any of:
    - Diagnosis code 3485* Cerebral edema
    - Diagnosis code 3484* Compression of brain
    - Diagnosis code 34939 Other dural tear
    - Diagnosis code 800** Fracture of vault of skull
    - Diagnosis code 801** Fracture of base of skull
    - Diagnosis code 803** Other and unqualified skull fractures
    - Diagnosis code 804** Multiple fractures involving skull or face with other bones
    - Diagnosis code 850** Concussion
    - Diagnosis code 851** Cerebral laceration and contusion
    - Diagnosis code 852** Subarachnoid subdural and extradural hemorrhage following injury
    - Diagnosis code 853** Other and unspecified intracranial hemorrhage following injury
    - Diagnosis code 854** Intracranial injury of other and unspecified nature
    - Procedure code 109 Other cranial puncture
    - Procedure code 110 Intracranial pressure monitoring
    - Procedure code 116 Intracranial oxygen monitoring
    - Procedure code 123 Reopening of craniotomy site
    - Procedure code 124 Other craniotomy
    - Procedure code 125 Other craniectomy
    - Procedure code 202 Other craniectomy

Figure 6

Figure 6. Excess draining from postsurgical wounds as proportion of all admissions by quarter. For leaky surgical wound code, slope=0.000027 (95%CI -0.000028 to 0.000082), p=0.34. For leaky surgical wound word and long stay, slope=0.0018 (95%CI 0.0012 to 0.0024), p<0.0001. For wound catheter word and long stay, slope=0.00038 (95%CI -0.00039 to 0.0012), p=0.34. For leaky surgical wound word and wound catheter word and long stay, slope=0.0011 (95%CI 0.00071 to 0.0016), p<0.0001.

Figure 6~~eFigure 4~~ notes:

- Leaky surgical wound word: text has "surg*" and "wound" and ("drain*" or "leak*)
- Long stay: >9 days in hospital admission
- Wound catheter word: “catheter”, “placed”, or “large”

Figure 7

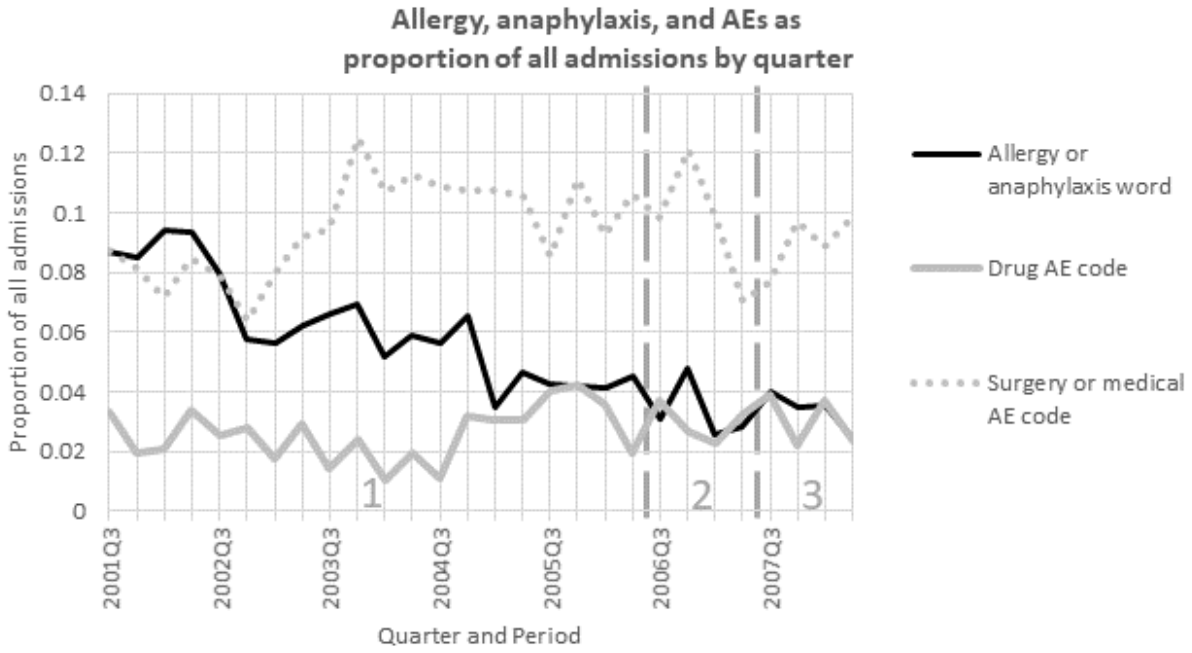


Figure 7. Allergy, anaphylaxis, and AE as proportion of admissions by quarter. For allergy or anaphylaxis word, slope=-0.0022 (95%CI -0.0027 to -0.0018), p<0.0001. For drug AE code, slope=0.00031 (95%CI -0.000079 to 0.00070), p=0.12. For surgery or medical AE code, slope=0.00049 (95%CI -0.00022 to 0.0012), p=0.18.

Figure 7 notes:

- Allergy or anaphylaxis word: "allerg*" or "anaphyl*"
- Drug AE code: 960** to 979** Poisoning By Drugs, Medicinals And Biological Substances
- Surgery or medical AE code: 996** to 999** Complications Of Surgical And Medical Care, Not Elsewhere Classified