Original Paper

Comparison of Rule-Based Natural Language Processing and Diagnosis Code Approaches to Detecting Bleeding in the Electronic Health Record: Algorithm Validation Within a Retrospective Cohort Study

Evan W Draper[1], PharmD; Jan A Anderson[1], PharmD; Surbhi Shah[2], MBBS; Leilia Brown[3], MSN, RN; Margaret R Hanson[4], RN; Kristin D Negley[5], MS, APRN, CNS; Jennifer M Lutz[6], MSN, RN; Andrea E Anderson[1], BSPharm; John G O'Meara[1], PharmD; Nicholas D Jakowenko[1], PharmD; Robert D McBane[7], MD

[1]Department of Pharmacy, Mayo Clinic, Rochester, MN, United States
[2]Division of Hematology and Medical Oncology, Department of Internal Medicine, Mayo Clinic, Phoenix, AZ, United States
[3]Department of Nursing, Mayo Clinic, Jacksonville, FL, United States
[4]Department of Nursing, Mayo Clinic Health System, Mankato, MN, United States
[5]Department of Nursing, Mayo Clinic, Rochester, MN, United States
[6]Department of Nursing, Mayo Clinic, Scottsdale, AZ, United States
[7]Division of Vascular Medicine, Gonda Vascular Center, Department of Cardiovascular Medicine, Mayo Clinic, Rochester, MN, United States

Corresponding Author:

Evan W Draper, PharmD
Department of Pharmacy
Mayo Clinic
201 1st St SW
Rochester, MN 55902
United States
Email: draper.evan@mayo.edu

Abstract

Background: Anticoagulant use is prevalent and increasing. High-quality anticoagulation management balances individual thrombotic and bleeding risks to optimize care. An automated algorithm to accurately extract major bleeding events from the electronic health record (EHR) of a large health care system would be extremely useful for tracking outcomes of anticoagulated patients.

Objective: The study aims to assess the accuracy of a natural language processing (NLP) approach, compared with and in conjunction with a diagnosis code (Dx Code)–based strategy, for detecting major bleeding from the EHR.

Methods: Patients with at least 1 prescription for apixaban, dabigatran, edoxaban, rivaroxaban, or warfarin between January 1, 2019, and December 31, 2023, were analyzed for major bleeding events. Major bleeding was defined similarly to the International Society on Thrombosis and Haemostasis definition. For the NLP algorithm, clinical notes and radiology reports were screened for mentions of bleeding. MedTagger, an open-source, rule-based NLP tool, was then used on the notes. The NLP results were transformed into a feature set, which was used to create a support vector machine model. Three NLP models were created for central nervous system bleeding, critical site bleeding, and bleeding with a hemoglobin drop or transfusions (other major bleeding). Training and test datasets were created using an active learning approach. The model was trained using 80% of the dataset, while the remaining 20% was used for testing. After model finalization, an independent random sample was used to validate the model. Dx Codes were also used to detect major bleeding, similar to previously published approaches. Five unique algorithms were compared: (1) NLP, (2) Dx Code in the primary position, (3) Dx Code in any position, (4) NLP $\cap$ Dx Code in any position, and (5) NLP $\cup$ Dx Code in any position. All 5 algorithms were compared with manual chart review.

Results: The final analysis identified 111,338 patients prescribed outpatient anticoagulation during the study. Manual adjudication was performed for 1189 of the 111,338 (1%) patients. Most of the 5 approaches performed well on the test data, with F_1-scores ranging from 0.545 to 0.959, except for Dx Code–Primary, with an F_1-score of 0.353 for critical site bleeding. The F_1-scores for the validation data ranged from 0.509 to 0.679. The NLP $\cap$ Dx Code in any position approach

yielded excellent precision of 0.864 (95% CI 0.667-0.953) but only 0.559 (95% CI 0.395-0.711) sensitivity. The NLP∪Dx Code in any position approach yielded good sensitivity of 0.824 (95% CI 0.665-0.917) but only 0.368 (95% CI 0.269-0.481) precision.

Conclusions: Electronic tools can screen the EHRs to identify major bleeding outcomes in large patient cohorts within a multisite health care network with high precision but only moderate sensitivity. Additional work is needed to enhance sensitivity while maintaining high precision.

JMIR Cardio 2026;10:e87424; doi: 10.2196/87424

Keywords: anticoagulants; hemorrhage; machine learning algorithms; natural language processing; population surveillance

Introduction

A growing number of patients take antithrombotic therapy, including more than 8 million on anticoagulants in the United States and more than 25 million on antiplatelet agents [1,2]. Overtime, this number is only anticipated to grow with the aging population. These medications carry a risk of major hemorrhage, which adversely impacts quality of life and increases mortality risk. There are an estimated 250,000 anticoagulant-related emergency department (ED) visits annually for bleeding outcomes and for every 100 patients dispensed an anticoagulant, there are nearly 9 ED visits [3]. High-quality anticoagulation delivery requires individualized patient assessment that balances thrombotic and bleeding risks to optimize management. This decision-making process is nuanced and dynamic, with overlap between bleeding and thrombotic risk factors as patients age.

High-quality delivery of anticoagulation for large health care networks involving tens of thousands of patients can be challenging. When considering ideal oversight of large anticoagulation networks, tools capable of monitoring hard outcomes are essential for quality improvement, and central to this process is an accurate and near-real–time assessment of bleeding events.

Given the widespread use of antithrombotic agents, manual chart abstraction for bleeding outcomes ascertainment is neither timely, cost-effective, nor practical. Methodologies using diagnosis code (Dx Code)–based major bleeding detection are widely used [4-8] but may lack sensitivity, specificity, and overall accuracy [9-11]. Electronic tools capable of automated and near-real–time identification of bleeding events could facilitate quality improvement and the refinement of safer anticoagulation delivery.

Established in 2019, the Mayo Clinic Anticoagulation and Antiplatelet Stewardship Quality Committee has been charged with oversight of antithrombotic delivery across the multisite institution. Since 2019, this has included more than 115,000 patients receiving a direct oral anticoagulant and more than 40,000 patients treated with warfarin. To facilitate the management of this large patient population, tools capable of providing near-real–time detection of bleeding and thrombosis are required. The objective of this study was to assess the accuracy of a natural language processing (NLP) approach compared with and in conjunction with a Dx Code–based strategy for detecting major bleeding.

Methods

Patient Population

An algorithm validation within a retrospective cohort study design was used to evaluate the accuracy of NLP, Dx Codes, and NLP plus Dx Codes to identify major bleeding in the electronic health record (EHR). Patients with at least one outpatient prescription for apixaban, dabigatran, edoxaban, rivaroxaban, or warfarin (January 1, 2019, to December 31, 2023) were included in the analysis. Data were collected from a centralized EHR that contains detailed information about all inpatient hospitalizations, outpatient visits, radiology images and reports, laboratory results, pathology results, and patient correspondence. The indication for anticoagulation was estimated based on Dx Codes in the EHR. Demographic information, indication, and prescribed medication were provided for all manually reviewed patients. Any missing values were categorized as unknown.

Bleeding Definition

Major bleeding was defined similar to the ISTH (International Society on Thrombosis and Haemostasis) and included symptomatic bleeding into a critical area or organ, and bleeding causing a hemoglobin decrease of at least 2 g/dL or transfusions of at least 2 units of packed red blood cells (PRBC) [12]. Symptomatic bleeding into a critical area or organ was divided into 2 categories: central nervous system (CNS) bleeding and critical site bleeding. CNS bleeds were defined as either intracranial or intraspinal bleeding. Critical site bleeding included intraocular, retroperitoneal, pericardial, hemothorax [13,14], and intra-articular bleeding. Bleeding with a hemoglobin drop or transfusions was referred to as “other major bleeding.” Fatal bleeding is included in the ISTH bleeding definition. However, a Dx Code specific to fatal bleeding does not exist, making it impossible to evaluate specifically for fatal bleeding without assuming the cause of death or manually reviewing the chart. Fatal bleeding events are likely to be found by the NLP approach and Dx Codes, but neither approach will specifically label the event as a fatal bleed. Mortality occurring within the same month as a bleeding event was collected and is described.

The ISTH definition of symptomatic bleeding with a 2 g/dL or greater change in hemoglobin or transfusion of 2 or more units of PRBC (“other major bleeding”) does not specify the timing for comparing hemoglobin results or for PRBC transfusions. For the purposes of this study, a qualifying drop in hemoglobin was defined by 2 methods.

The first method compared a hemoglobin result with any other hemoglobin result within the preceding 30 days. The second method compared a hemoglobin result to the immediately prior hemoglobin result. If either method was true, then the hemoglobin change met the definition. Using both approaches captures a slow, occult bleed with months between hemoglobin results as well as a more acute drop in hemoglobin values. For the transfusion requirement definition, 2 or more units of PRBC had to be administered within a 3-day time increment. If there was a 4-day separation between transfusions, that would not meet the definition. Other studies have proposed a 48-hour limit on the hemoglobin change and did not specify a time frame for the transfusions [15]. The temporal thresholds for transfusions and hemoglobin change for this study were determined through active learning and by examining bleeding events within the EHR.

Event Identification and Adjudication

To compare the accuracy of NLP and Dx Codes to detect bleeding outcomes, manual chart review for event validation was performed. Positive bleeding events for training, testing, and validation were manually and independently reviewed in the EHR by at least 2 study team members (nurses, pharmacists, or physicians) to create the gold standard against which the NLP and Dx Code approaches were compared. Months in which no mentions of bleeding terms were noted throughout the EHR were assumed to be months without a bleeding event, and no EHR manual review was completed. Months in which bleeding terms were found or Dx Codes were present were manually reviewed in the EHR by 1 study team member. Obvious nonevents, such as "no intracranial hemorrhage present," were categorized as nonevents by only 1 study team member. Ambiguous situations such as "differential diagnosis includes hemarthrosis, septic arthritis…" were reviewed by at least 2 study team members when the initial review could not confidently rule out a bleeding event. The date and type of the bleed (CNS, critical site, or other major bleed) were recorded, and each month adjudicated was classified as a bleeding event or not a bleeding event. Adjudication by month was chosen after an initial 3-day rolling window proved performance-intensive. Cohen $\varkappa$ was calculated by evaluating interteam member rating reliability on the validation sample. Specific sentences or notes were not annotated for NLP purposes. A bleeding event could occur in the outpatient or inpatient setting. Event disagreements were resolved through group discussion. Only the incident bleeding event was included for patients with multiple events. While prescribed anticoagulation was an inclusion criterion for the analysis, temporary anticoagulant interruption prior to the bleeding event did not exclude the patient. For event adjudication, surgical bleeding leading to a drop in hemoglobin or transfusion was excluded. The NLP algorithm does not have the functionality to exclude surgical bleeding, and future work would be required to align the algorithm with manual review. Additional details on event adjudication are available in the Section S1 of Multimedia Appendix 1.

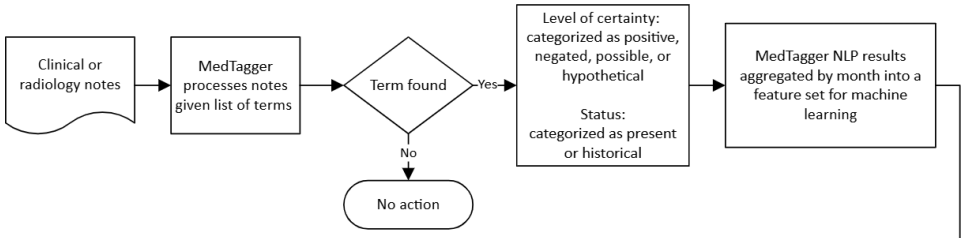
NLP

MedTagger is an open-source, rule-based NLP program from Open Health Natural Language Processing (OHNLP) Consortium that detects specific terms and then uses more than 700 rules to determine the context of each term [16]. MedTagger classifies the context using the following elements: experiencer: patient or other individual; status: present or historical; and level of certainty: positive, negative, possible, or hypothetical. The text was not preprocessed before using MedTagger. A MedTagger version based on the OHNLP Toolkit (v1.0.82) was used.

All patient notes were searched for CNS bleeding and critical site bleeding using regular expressions. The incident date and 3 days following a 2 g/dL or greater drop in hemoglobin or 2 or more units of PRBC were also searched using regular expressions. Search terms were specific to each bleeding type. Additional information on the search terms is provided in Section S2 of Multimedia Appendix 1. Notes with bleeding mentions found by regular expressions were further processed with MedTagger NLP. Exclusionary terms were added to the algorithm for NLP refinement. Additional details on MedTagger search terms and excluded note sections are available in Section S3 of Multimedia Appendix 1.

A feature set for each algorithm was created, and machine learning using a support vector machine model was created to identify bleeds. Support vector machine was selected as the model after evaluating logistic regression, random forest, and gradient-boosted machine models. Adjudicated data were divided into training (80%) and testing sets (20%), stratified by patient to ensure no data-leakage. Months without a bleeding event were combined with months with a CNS bleed and used to create the CNS bleed model. The same months without a bleeding event were used to create the critical site bleed and other major bleed models. The feature set aggregated NLP results by month, and NLP results from clinical notes and radiology notes were handled as separate features. Machine learning parameters were chosen using grid search and cross-validation on the training set. Additional information on model training is provided in Section S4 of Multimedia Appendix 1. A visual representation of the approach is provided in Figure 1.

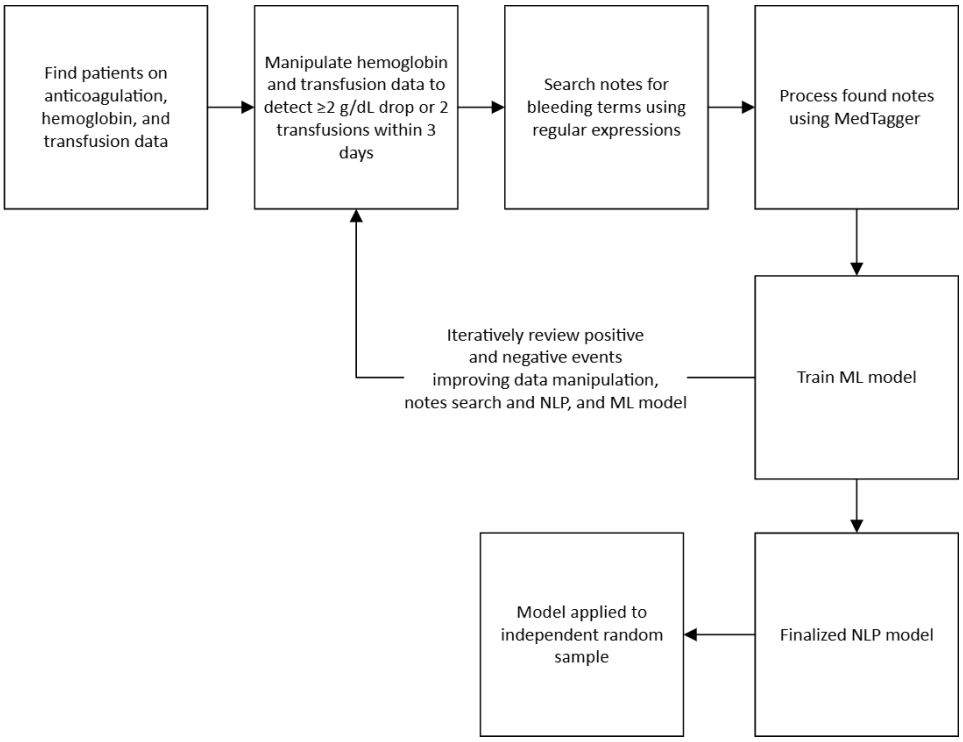
Figure 1. MedTagger process. MedTagger is an open-source, rule-based natural language processing (NLP) tool that categorizes medical terms with regard to certainty (ie, positive, negated, or hypothetical) and status (ie, present or history of). This work used MedTagger to categorize mentions of bleeding found in clinical and radiology notes in an attempt to find major bleeding events among patients on anticoagulation. The MedTagger results are aggregated to create a feature set, which is subsequently used in machine learning. MRN: medical record number; *each combination of NLP results creates an additional feature; **one row per MRN per month.



MRN	Month start date	Positive present radiology	Negated present radiology	Positive history of radiology	...*	Positive present clinical
1234	5/1/2021	1	0	0	...	2
1234	4/1/2020	2	0	0	...	1
1234	4/1/2022	1	0	0	...	1
1234	4/1/2021	1	0	0	...	0
1234	7/1/2021	17	0	0	...	0
...**	...	...	...	...	...	...

The NLP approach was refined iteratively by reviewing high and low probability cases similar to active learning (Figure 2) [17]. This approach was used to select cases for manual review and adjudication during model development. This approach does not suffer as much from class imbalance compared to random sampling. After model development and finalization, an independent random sample of patients was manually adjudicated and run through the model.

Figure 2. Natural language processing (NLP) model development. This retrospective cohort study of Mayo Clinic patients prescribed oral anticoagulation from January 1, 2019, to December 31, 2023, assessed the feasibility and accuracy of an NLP and machine learning (ML) approach to detect major bleeding from the electronic health record (EHR). The model development process iteratively refined the search terms, the method of combining discrete hemoglobin and transfusion data, and the feature creation process. The finalized NLP model was applied to an independent, random sample and was compared with 2 different diagnosis code approaches.



Dx Code

Bleeding events were also identified using Dx Codes [8]. Dx Codes associated with ED visits or hospital admissions were included. The primary Dx Code ("Dx Code–Primary") or any Dx Code associated with the encounter ("Dx Code–Any Position") were evaluated. Specific *ICD-10* (*International Statistical Classification of Diseases, Tenth Revision*) codes used are provided in Section S5 of Multimedia Appendix 1. Data were aggregated by month based on the admission date.

NLP Combined With Dx Code

The Dx Code–Any Position approach and the NLP approach were combined in 2 ways: "NLP∩Dx Code Any" requires both the NLP and Dx Code–Any Position results to be positive, while "NLP∪Dx Code Any" allows either the NLP or the Dx Code–Any Position approach to be positive.

Algorithm and Comparison Summary

Five approaches for three different bleeding types were created: (1) NLP, (2) Dx Code–Primary, (3) Dx Code–Any, (4) NLP∩Dx Code Any, and (5) NLP∪Dx Code Any. For model test data, the 5 approaches were evaluated for each of the 3 different bleeding types (CNS bleeding, critical site bleeding, and other major bleeding). For model validation on a random sample of data, the 3 bleeding types were combined due to the low volume of events.

Statistical Analysis

Accuracy measures of sensitivity, specificity, precision, and F_1-score, with 95% CIs, were calculated for each approach compared against the gold standard of a manually adjudicated review. The methodology to calculate the 95% CIs was SciPy's implementation of the Wilson without continuity correction for sensitivity, specificity, and precision [18]. Given no clear best approach for calculating a CI for F_1-score [19], a bootstrap methodology using 5000 samples was used to calculate the 95% CI for F_1-score. The F_1-score is a measurement that evaluates both sensitivity and precision in a single metric. Cohen κ was calculated by evaluating interteam member rating reliability and used the bootstrap methodology. The above analyses were performed for each type of bleeding as well as for a combined "all bleeding" analysis. All analyses were performed using Python (version 3.10.18).

Ethical Considerations

The project was deemed exempt by the Mayo Clinic Institutional Review Board in the context of a quality improvement project. Patient information was stored in a secure environment within the Mayo Clinic Cloud. Study team members were provided with access to the minimum amount of identifiable information necessary to verify the bleeding event during event adjudication.

Results

The final analysis identified 111,338 total patients prescribed outpatient anticoagulation over the study time frame. Clinical notes and radiology reports for the included patients were searched for bleeding terms, identifying 264,551 clinical notes and radiology reports for CNS bleeds, 73,973 for critical site bleeds, and 581,120 for other major bleeding. The total number of unique notes with bleeding terms found is 773,779. These notes and reports were processed using MedTagger (Figure 3). A cohort of 604 patients was manually adjudicated during active learning and was used for model training and testing. Each patient and month reviewed contributed to model development. A total of 1001 patient-months were manually reviewed, with 643 months without a bleeding event, 183 months with CNS bleeding, 68 months with critical site bleeding, and 107 months with other major bleeding. An additional random sample of 585 patients was used to validate the final model. The validation sample yielded a total of 30,757 patient-months, with 30,723 months without a bleeding event, 7 months with a CNS bleed, 7 months with a critical site bleed, and 20 months with other major bleed. Cohen κ between the reviewers for the random sample was 0.66 (95% CI 0.38-0.88). Demographic information, indications, and medications prescribed for all patients manually reviewed are provided in Table 1. The patients had a mean age of 70 (SD 14.4) years, and 57% were male patients. The majority of patients were White individuals. The 2 most common indications for anticoagulant therapy were atrial fibrillation and thromboembolism. The 2 most common anticoagulants used were apixaban and warfarin.

Figure 3. Pipeline and adjudicated reviews flow diagram. A natural language processing (NLP) and machine learning–based approach to detect major bleeding events from the electronic health record (EHR) was developed and evaluated. The approach processed 111,338 patients and 773,779 unique notes. Patients on oral anticoagulation were first identified. Notes mentioning bleeding were found and processed by MedTagger. Manual reviews of bleeding events were performed to evaluate the accuracy of the approach. CNS: central nervous system.

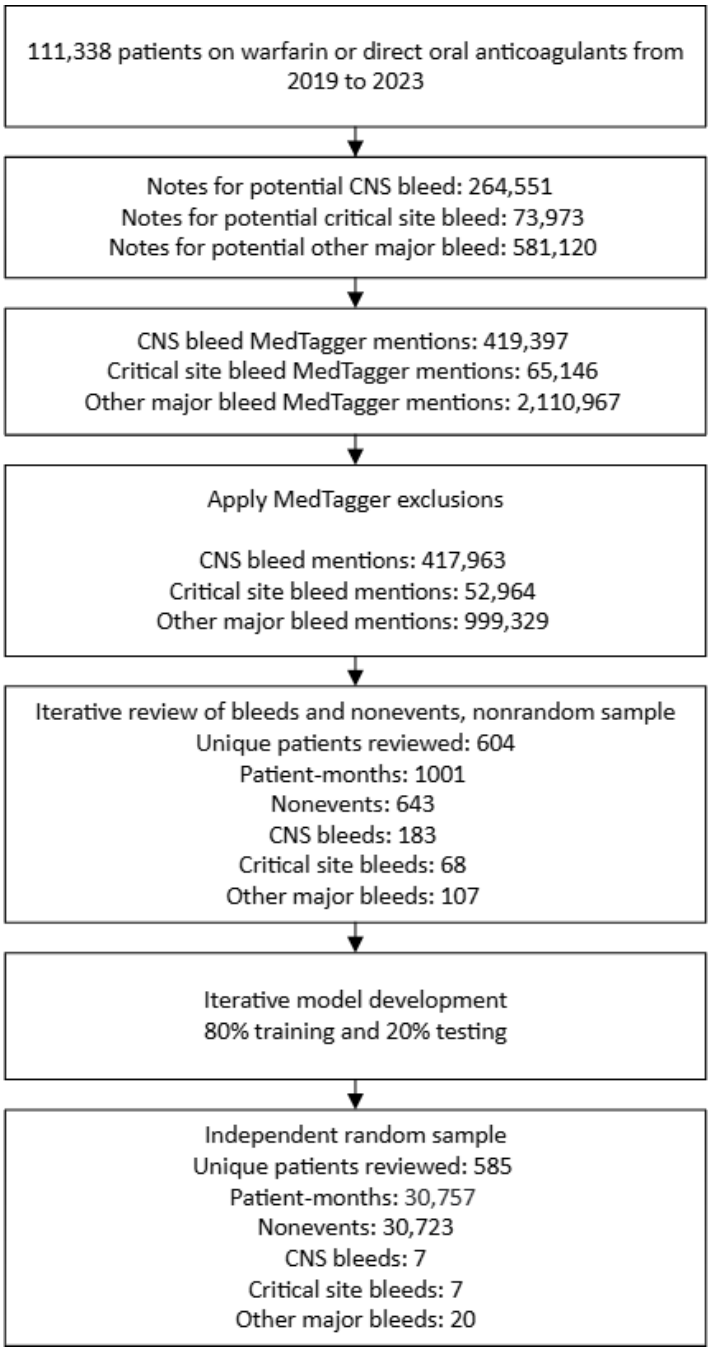


Table 1. Baseline demographics and clinical characteristics.

Characteristic	Patients (N=1189)
Age (y), mean (SD)	70.0 (14.4)
Sex: female, n (%)	514 (43.2)
Race, n (%)	
White	1108 (93.2)
Black	29 (2.4)
Asian	20 (1.7)
American Indian/Alaskan Native	5 (0.4)
Other/unknown	27 (2.3)

Characteristic	Patients (N=1189)
Ethnicity, n (%)	
Not Hispanic/Latino	1142 (96.0)
Hispanic/Latino	25 (2.1)
Not disclosed/unknown	22 (1.9)
Anticoagulant indication, n (%)	
Atrial fibrillation	638 (53.7)
Thromboembolism	399 (33.6)
Other/unknown	111 (9.3)
Heart valve prosthesis or repair	23 (1.9)
Peripheral artery disease	18 (1.5)
Anticoagulant used, n (%)	
Apixaban	630 (53.0)
Warfarin	342 (28.8)
Rivaroxaban	208 (17.5)
Dabigatran	8 (0.7)
Edoxaban	1 (0.1)

Performance metrics and 95% CIs for the testing data and individual bleeding types (Table 2) compare the performance of the 5 different approaches. CNS bleeding algorithms performed the best with F_1-scores of 0.946 (95% CI 0.885-0.989) for NLP, 0.959 (95% CI 0.904-1.000) for Dx Code–Primary, and 0.947 (95% CI 0.886-0.989) for Dx Code–Any, 0.959 (95% CI 0.906-1.000) for NLP∩Dx Code Any, and 0.935 (95% CI 0.868-0.986) for NLP∪Dx Code Any.

Table 2. Performance metrics for individual bleeding types for test data.

Bleeding classification method	Sensitivity (95% CI)	Specificity (95% CI)	Precision (95% CI)	F_1-score (95% CI)
CNS[a] bleeds				
NLP[b]	0.946 (0.823-0.985)	0.982 (0.938-0.995)	0.946 (0.823-0.985)	0.946 (0.885-0.989)
Dx Code[c]–Primary	0.946 (0.823-0.985)	0.991 (0.952-0.998)	0.972 (0.858-0.995)	0.959 (0.904-1.000)
Dx Code–Any	0.973 (0.862-0.995)	0.974 (0.925-0.991)	0.923 (0.797-0.973)	0.947 (0.886-0.989)
NLP∩Dx Code Any[d]	0.946 (0.823-0.985)	0.991 (0.952-0.998)	0.972 (0.858-0.995)	0.959 (0.906-1.000)
NLP∪Dx Code Any[e]	0.973 (0.862-0.995)	0.965 (0.913-0.986)	0.900 (0.769-0.960)	0.935 (0.868-0.986)
Critical site bleeds				
NLP	0.643 (0.388-0.837)	0.983 (0.942-0.995)	0.818 (0.523-0.949)	0.720 (0.471-0.897)
Dx Code–Primary	0.214 (0.076-0.476)	1.000 (0.969-1.000)	1.000 (0.439-1.000)	0.353 (0.000-0.632)
Dx Code–Any	0.429 (0.214-0.674)	0.983 (0.942-0.995)	0.750 (0.409-0.929)	0.545 (0.235-0.769)
NLP∩Dx Code Any	0.429 (0.214-0.674)	1.000 (0.969-1.000)	1.000 (0.610-1.000)	0.600 (0.286-0.828)
NLP∪Dx Code Any	0.643 (0.388-0.837)	0.967 (0.918-0.987)	0.692 (0.424-0.873)	0.667 (0.421-0.857)
Other major bleeds				
NLP	0.818 (0.615-0.927)	0.958 (0.906-0.982)	0.783 (0.581-0.903)	0.800 (0.650-0.913)
Dx Code–Primary	0.545 (0.347-0.731)	0.975 (0.929-0.991)	0.800 (0.548-0.930)	0.649 (0.437-0.809)
Dx Code–Any	0.636 (0.43-0.803)	0.942 (0.884-0.971)	0.667 (0.454-0.828)	0.651 (0.452-0.800)
NLP∩Dx Code Any	0.591 (0.387-0.767)	0.983 (0.941-0.995)	0.867 (0.621-0.963)	0.703 (0.500-0.851)
NLP∪Dx Code Any	0.864 (0.667-0.953)	0.917 (0.853-0.954)	0.655 (0.473-0.801)	0.745 (0.595-0.864)

[a]CNS: central nervous system.
[b]NLP: natural language processing.
[c]Dx Code: diagnosis code.
[d]NLP∩Dx Code Any is the intersection (∩) of NLP and Dx Code Any, where both approaches must be positive to qualify as a bleed.
[e]NLP∪Dx Code Any is the union (∪) of NLP and Dx Codes where either approach can be positive to qualify as a bleed.

Based on F_1-scores, the NLP approach performed the best for critical site bleeds, with 0.720 (95% CI 0.471-0.897), and for other major bleeds, with a score of 0.800 (95% CI 0.650-0.913).

Performance metrics for the validation data and combined bleeds (Table 3) demonstrate decreased performance compared with the test data. F_1-scores ranged from 0.509 (95% CI 0.388-0.618) to 0.679 (95% CI 0.524-0.812) for

validation compared with up to 0.959 for the test data. The best precision was for the NLP ∩ Dx Code Any with 0.864 (95% CI 0.667-0.953). However, this approach only yielded 0.559 (95% CI 0.395-0.711) sensitivity. Model features are available in Section S6 of Multimedia Appendix 1.

Table 3. Performance metrics for the random validation sample.

Bleeding identification method	Sensitivity (95% CI)	Specificity (95% CI)	Precision (95% CI)	F_1-score (95% CI)
NLP[a]	0.676 (0.508-0.809)	0.999 (0.999-1.000)	0.511 (0.370-0.650)	0.582 (0.441-0.703)
Dx Code[b]–Primary	0.618 (0.450-0.761)	1.000 (0.999-1.000)	0.618 (0.450-0.761)	0.618 (0.462-0.738)
Dx Code–Any	0.706 (0.538-0.832)	0.999 (0.999-0.999)	0.453 (0.327-0.585)	0.552 (0.415-0.667)
NLP ∩ Dx Code Any[c]	0.559 (0.395-0.711)	1.000 (1.000-1.000)[d]	0.864 (0.667-0.953)	0.679 (0.524-0.812)
NLP ∪ Dx Code Any[e]	0.824 (0.665-0.917)	0.998 (0.998-0.999)	0.368 (0.269-0.481)	0.509 (0.388-0.618)

[a]NLP: natural language processing.
[b]Dx Code: diagnosis code.
[c]NLP ∩ Dx Code Any is the intersection (∩) of NLP and Dx Code Any, where both approaches must be positive to qualify as a bleed.
[d]The specificity, rounded to 6 decimal places, is 0.999902 (0.999713-0.999967).
[e]NLP ∪ Dx Code Any is the union (∪) of NLP and Dx Code Any, where either approach can be positive to qualify as a bleed.

Mortality in the training and test data occurred within the same month as the bleed outcome for 22 out of 183 (12%) CNS bleeds, 2 out of 68 (3%) critical site bleeds, and 5 out of 107 (5%) other major bleeds. Mortality in the validation data occurred within the same month, 2 out of 7 (29%) of CNS bleeds, 0 out of 7 (0%) critical site bleeds, and 2 out of 20 (10%) other major bleeds. While causality assignment for fatal bleeding can be challenging, the temporal association between the bleeding event and death suggests a potential contribution to this outcome.

Discussion

Principal Findings

The principal finding of the current study is that electronic tools can screen electronic medical records for major bleeding outcomes from large patient cohorts within a multisite health care network with high precision and moderate sensitivity. In validation, the NLP ∩ Dx Code Any approach may have the most clinical utility with the highest precision of 0.864 (95% CI 0.667-0.953) and moderate sensitivity of 0.559 (95% CI 0.395-0.711). Given the high volume of patients on anticoagulant medications, an electronic tool that maximizes precision might provide the most useful measure for groups focused on quality improvement and process oversight.

The clinical utility of these findings is notable for several reasons. One could envision the use of this information to improve patient outcomes. First, appropriate use criteria could be assessed to ensure anticoagulants are appropriately prescribed. For example, patients with severe renal impairment, liver impairment, or thrombocytopenia may be at particular risk for bleeding outcomes. When tracked electronically, "our practice alerts" could then be delivered to providers. Second, duplicate antithrombotic delivery could be scrutinized and corrected. This could include discontinuation of antiplatelets where appropriate. Third, dose adjustments of direct anticoagulants could be reviewed such as the "ABC" criteria for apixaban or dose reductions for secondary venous thromboembolism prevention after 6 months. Fourth, this tool could be combined with electronic measures of International Normalized Ratio (INR) stability for those patients receiving warfarin therapy. While engaging the practice after a major bleed or intracranial bleed has occurred is not ideal, the use of this tool could have a major impact on practice change to prevent the next bleeding event. Fifth, these tools can provide a real-world, in-house assessment of new antithrombotic strategies. As the factor XI inhibitors are Food and Drug Administration (FDA)–approved and incorporated into clinical practice, adverse outcomes, compared with traditional anticoagulants, can be assessed in real-time. Clearly, more work is needed to refine this and other tools to maximize utility. These types of tools may then enable incorporation into clinical practice for quality oversight.

Comparison to Prior Work

Multiple recent studies have evaluated the ability to detect and categorize bleeding events using structured [15,20], unstructured [21], or a combination of structured and unstructured data [22,23]. Coumau et al [23] evaluated various approaches for detecting major bleeding as well as clinically relevant nonmajor bleeding. Their top approach was a combination of structured and unstructured data and yielded a sensitivity of 0.84, a precision of 0.51, and an F_1-score of 0.64. Gaspar et al [21] performed document-level categorization of discharge summaries of older adults and compared an NLP approach with a rule-based, Dx Code approach. The NLP approach outperformed the rule-based, Dx Code approach in the ability to categorize a document as "no bleeding," "clinically significant bleeding," "severe bleeding," and "history of bleeding" with an F_1-score of 0.80. An et al [20] applied a previously published algorithm [15] using structured data to detect major bleeding and clinically relevant nonmajor bleeding. They report a sensitivity of 0.91 and specificity of 0.92 for detecting major bleeding and a sensitivity of 0.52 and specificity of 0.92 for detecting clinically relevant nonmajor bleeding. They used a conditional sampling approach for the validation.

Older studies have used similar approaches to scrutinize clinical outcomes among large cohorts of anticoagulated

patients. Cunningham et al [9] used Dx Codes in addition to transfusion data to identify bleeding related to anticoagulation and achieved an 89% positive predictive value. Negative predictive value was not reported. Arnason et al [10] reviewed patient charts with a "bleeding" Dx Code and determined if a major bleed occurred. Accuracy measures with 95% CIs were reported. Definite bleeding had a sensitivity of 93% (95% CI 90%-96%) and specificity of 88% (95% CI 83%-91%). Major bleeding had a sensitivity of 94% (95% CI 91%-96%) and specificity of 83% (95% CI 78%-87%). Tirschwell and Longstreth [11] used Dx Codes to detect hemorrhagic stroke in administrative claims data, including primary diagnosis only, first 2 diagnoses, and all diagnoses. The best algorithm differed by stroke type, and for intracranial hemorrhage and subarachnoid hemorrhage, using only the primary diagnosis maximized agreement. Intracranial hemorrhage had sensitivity of 82% (95% CI 66%-92%), specificity of 93% (95% CI 86%-97%), and $\varkappa$ 0.74 (0.58-0.90), while subarachnoid hemorrhage had a sensitivity of 98% (95% CI 90%-100%), specificity of 92% (95% CI 84%-96%), and $\varkappa$ 0.87 (95% CI 0.71-1.00). While Dx Code approaches can be automated, they may not have sufficient accuracy metrics to be used in practice or within a learning health system. For example, a method with a sensitivity of 98% and a specificity of 92% applied to a population of 100,000 and assuming a 1% bleeding rate would result in 9780 false positives and 20 false negatives and a precision of 11%.

Others have assessed strategies that do not use Dx Codes for bleeding detection. Lee et al [24] studied bleeding events related to clopidogrel usage. Ten different approaches, 3 rule-based, 3 score-based, and 4 machine learning–based approaches were analyzed. While the highest precision was 97%, the sensitivity was low at 29%. A machine learning approach performed best, with a precision of 89%, a sensitivity of 97% (sensitivity in this study may not be accurate due to nonrandom sampling), and an F_1-score of 92%. Taggart et al [25] searched for "clinically relevant bleeding" in the general adult intensive care unit population using the MIMIC-III (Medical Information Mart for Intensive Care) database. Using NLP results, they created rule-based and machine learning approaches for detecting bleeding. "Clinically relevant bleeding" was defined as "as symptomatic bleeding that required treatment." The rule-based approach had a sensitivity of 91%, specificity of 85%, and precision of 63%. The machine learning–based approach had a sensitivity of 94%, specificity of 72%, and precision of 49%. Shung et al [26] assessed gastrointestinal bleeding in the ED. An early large language model (LLM) was the best approach (sensitivity of 93%, specificity of 63%, and precision of 84%) compared with either an NLP machine learning model and a discrete data model. More recent studies have focused on bleeding mention detection at the sentence or note level [27,28]. Further studies are needed to determine how sentence or note level detection would translate into patient level detection of bleeding events.

LLMs for information extraction present a potential future step beyond rule-based NLP that may have a superior ability to distinguish between bleeding events and no bleeding. Limitations of LLM approaches include potentially high cost, processing time and power requirements, and data privacy concerns. While recent advances suggest that LLMs and deep learning language models may be superior to rule-based NLP [29-31], 1 recent study found that rule-based NLP remained superior [32]. One potential future direction of this work is to compare and/or combine the NLP$\cap$Dx Code Any approach with an LLM approach.

Study strengths include the use of a rule-based NLP system to detect major bleeding for over 100,000 patients prescribed outpatient anticoagulation. This proof of concept demonstrates that large-scale detection of bleeding with NLP is feasible. Train, test, and validation cases were adjudicated by multiple reviewers to ensure accuracy. Two commonly used approaches were compared with a novel NLP approach that combines structured data of hemoglobin and transfusions to approximate the ISTH major bleed definition and scrutinizes both clinical and radiology notes for bleeding detection. An open-source NLP program that can be applied to large numbers of patients and notes was used at minimal expense.

Limitations

The study has several limitations. First, only the incident bleeding event per patient was included. Yet, it is well known that patients can experience more than one bleeding event. Indeed, those patients with recurrent bleeding may be fundamentally different compared to those with an isolated event. By restricting the outcome to the initial bleed without further refinement, important clinical variables may be missed. Second, while considerable time was spent to thoroughly validate 1189 patients, this represents only 1% of the total population at risk. Whether this was an adequately representative sample is not known. Third, patients who suffered a major bleed were not necessarily on anticoagulation at the time. There may be fundamental differences between anticoagulation-related major bleeding and bleeding events occurring without anticoagulation. Fourth, in the manual adjudication, patients experiencing a bleeding complication related to surgery were excluded. These represent a unique and important cohort; however, uncontrollable variables related to surgical technique, surgical site, and extent are difficult to control for and unique to surgical populations. Surgical bleeding is a topic unto itself and not easily comprehensible when combined with nonsurgical bleeding. Fifth, the approach was not externally validated and relies on infrastructure that may be unique to Mayo Clinic. Sixth, the Cohen $\varkappa$ metric on the validation sample was 0.66 (95% CI 0.38-0.88), which is lower than desired. Adjudicating bleeding events is challenging, and discrepancies between reviewers were resolved by physician review or group consensus.

Conclusions

A rule-based NLP pipeline processing tool was developed and validated for analysis of more than 100,000 patients over a period of 5 years. These kinds of tools are essential

for developing quality improvement programs with the structured delivery of high-quality anticoagulation for large patient populations. While the "NLP∩Dx Code" in any position approach yielded excellent precision, the sensitivity of this approach needs to be improved. Future directions will combine this rule-based and Dx Code approach with new LLM-based approaches to information extraction.

Acknowledgments

Che Ngufor and Shrinath Patel reviewed machine learning scripts and provided technical guidance. David Hodge reviewed the statistical plan.

Generative AI Use

Generative AI was not used in the drafting of this manuscript.

Funding

The authors declared no financial support was received for this work.

Data Availability

The datasets generated or analyzed during this study are not publicly available due to the presence of protected health information (PHI) in clinical notes and radiology reports. Summary counts of natural language processing findings and/or code are available from the corresponding author upon reasonable request.

Authors' Contributions

Conceptualization: EWD, SS, RDM
Data curation: EWD (lead), JAA, SS, LB, MRH, KDN, JML, AEA, JGO, NDJ, RDM
Formal analysis: EWD
Investigation: EWD
Supervision: SS, RDM
Writing – original draft: EWD
Writing – review & editing: EWD, JAA, SS, LB, MRH, KDN, JML, AEA, JGO, NDJ, RDM

Conflicts of Interest

None declared.

Multimedia Appendix 1

Additional information regarding event adjudication, note search terms, MedTagger search terms, model training, diagnosis codes, and feature lists.
[DOCX File (Microsoft Word File), 38 KB-Multimedia Appendix 1]

References

1. Yates NY, Hale SA, Clark NP. The impact of clinical pharmacy services on direct oral anticoagulant medication selection and dosing in the ambulatory care setting. J Pharm Pract. Jun 2024;37(3):671-676. [doi: 10.1177/08971900231166555] [Medline: 36989436]
2. Gupta M, Gulati S, Nasir K, Sarraju A. Aspirin use prevalence for cardiovascular disease prevention among U.S. adults from 2012 to 2021. Ann Intern Med. Aug 2024;177(8):1139-1141. [doi: 10.7326/M24-0427] [Medline: 38914002]
3. Geller AI, Shehab N, Lovegrove MC, Weidle NJ, Budnitz DS. Bleeding related to oral anticoagulants: trends in US emergency department visits, 2016-2020. Thromb Res. May 2023;225:110-115. [doi: 10.1016/j.thromres.2023.03.010] [Medline: 37062120]
4. Siontis KC, Zhang X, Eckard A, et al. Outcomes associated with apixaban use in patients with end-stage kidney disease and atrial fibrillation in the United States. Circulation. Oct 9, 2018;138(15):1519-1529. [doi: 10.1161/CIRCULATIONAHA.118.035418] [Medline: 29954737]
5. Reynolds MR, Gunnarsson CL, Hunter TD, et al. Health outcomes with catheter ablation or antiarrhythmic drug therapy in atrial fibrillation: results of a propensity-matched analysis. Circ Cardiovasc Qual Outcomes. Mar 1, 2012;5(2):171-181. [doi: 10.1161/CIRCOUTCOMES.111.963108] [Medline: 22373904]
6. Jun M, James MT, Manns BJ, et al. The association between kidney function and major bleeding in older adults with atrial fibrillation starting warfarin treatment: population based observational study. BMJ. Feb 3, 2015;350:h246. [doi: 10.1136/bmj.h246] [Medline: 25647223]
7. White RH, Beyth RJ, Zhou H, Romano PS. Major bleeding after hospitalization for deep-venous thrombosis. Am J Med. Nov 1999;107(5):414-424. [doi: 10.1016/s0002-9343(99)00267-3] [Medline: 10569295]
8. Ngufor C, Yao X, Inselman JW, et al. Identifying treatment heterogeneity in atrial fibrillation using a novel causal machine learning method. Am Heart J. Jun 2023;260:124-140. [doi: 10.1016/j.ahj.2023.02.015] [Medline: 36893934]

9. Cunningham A, Stein CM, Chung CP, Daugherty JR, Smalley WE, Ray WA. An automated database case definition for serious bleeding related to oral anticoagulant use. Pharmacoepidemiol Drug Saf. Jun 2011;20(6):560-566. [doi: 10.1002/pds.2109] [Medline: 21387461]
10. Arnason T, Wells PS, van Walraven C, Forster AJ. Accuracy of coding for possible warfarin complications in hospital discharge abstracts. Thromb Res. 2006;118(2):253-262. [doi: 10.1016/j.thromres.2005.06.015] [Medline: 16081144]
11. Tirschwell DL, Longstreth WT. Validating administrative data in stroke research. Stroke. Oct 2002;33(10):2465-2470. [doi: 10.1161/01.str.0000032240.28636.bd] [Medline: 12364739]
12. Schulman S, Kearon C, Subcommittee on Control of Anticoagulation of the Scientific and Standardization Committee of the International Society on Thrombosis and Haemostasis. Definition of major bleeding in clinical investigations of antihemostatic medicinal products in non-surgical patients. J Thromb Haemost. Apr 2005;3(4):692-694. [doi: 10.1111/j.1538-7836.2005.01204.x] [Medline: 15842354]
13. Hauk L. Bleeding with anticoagulants: ACC releases expert consensus decision pathway to guide management. Am Fam Physician. Jul 1, 2018;98(1):57-58. [Medline: 30215949]
14. Tomaselli GF, Mahaffey KW, Cuker A, et al. 2020 ACC expert consensus decision pathway on management of bleeding in patients on oral anticoagulants: a report of the American College of Cardiology Solution Set Oversight Committee. J Am Coll Cardiol. Aug 4, 2020;76(5):594-622. [doi: 10.1016/j.jacc.2020.04.053] [Medline: 32680646]
15. Hartenstein A, Abdelgawwad K, Kleinjung F, Privitera S, Viethen T, Vaitsiakhovich T. Identification of International Society on Thrombosis and Haemostasis major and clinically relevant non-major bleed events from electronic health records: a novel algorithm to enhance data utilisation from real-world sources. Int J Popul Data Sci. 2023;8(1):2144. [doi: 10.23889/ijpds.v8i1.2144] [Medline: 38414540]
16. Liu H, Bielinski SJ, Sohn S, et al. An information extraction framework for cohort identification using electronic health records. AMIA Jt Summits Transl Sci Proc. 2013;2013:149-153. [Medline: 24303255]
17. Settles B. Active learning literature survey. University of Wisconsin–Madison; 2009. URL: https://minds.wisc.edu/server/api/core/bitstreams/8a78cf83-0702-4157-aeff-3cacb62f9ad5/content [Accessed 2025-03-10]
18. proportion_ci. SciPy. URL: https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats._result_classes.BinomTestResult.proportion_ci.html#r67c9937f472a-3 [Accessed 2026-04-01]
19. Lam KFY, Gopal V, Qian J. Confidence intervals for the F1 score: a comparison of four methods. arXiv. Preprint posted online on Sep 26, 2023. [doi: 10.48550/ARXIV.2309.14621]
20. An J, Hartenstein A, Choi SK, et al. Validation of electronic health record-based International Society on Thrombosis and Haemostasis bleeding algorithms in a US health system. Res Pract Thromb Haemost. Apr 2026;10(4):106616. [doi: 10.1016/j.rpth.2026.106616] [Medline: 42232248]
21. Gaspar F, Zayene M, Coumau C, et al. Natural language processing and ICD-10 coding for detecting bleeding events in discharge summaries: comparative cross-sectional study. JMIR Med Inform. Aug 29, 2025;13:e67837. [doi: 10.2196/67837] [Medline: 40882207]
22. Zheng NS, Keloth VK, You K, et al. Detection of gastrointestinal bleeding with large language models to aid quality improvement and appropriate reimbursement. Gastroenterology. Jan 2025;168(1):111-120. [doi: 10.1053/j.gastro.2024.09.014] [Medline: 39304088]
23. Coumau C, Gaspar F, Zayene M, et al. Detection of antithrombotic-related bleeding in older inpatients: multicenter retrospective study using structured and unstructured electronic health record data. J Med Internet Res. Jan 29, 2026;28:e77809. [doi: 10.2196/77809] [Medline: 41610431]
24. Lee HJ, Jiang M, Wu Y, et al. A comparative study of different methods for automatic identification of clopidogrel-induced bleedings in electronic health records. AMIA Jt Summits Transl Sci Proc. 2017;2017(185–192):185-192. [Medline: 28815128]
25. 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. Oct 5, 2018;1(6):e183451. [doi: 10.1001/jamanetworkopen.2018.3451] [Medline: 30646240]
26. Shung D, Tsay C, Laine L, et al. Early identification of patients with acute gastrointestinal bleeding using natural language processing and decision rules. J Gastroenterol Hepatol. Jun 2021;36(6):1590-1597. [doi: 10.1111/jgh.15313] [Medline: 33105045]
27. Pedersen JS, Laursen MS, Rajeeth Savarimuthu T, et al. Deep learning detects and visualizes bleeding events in electronic health records. Res Pract Thromb Haemost. May 2021;5(4):e12505. [doi: 10.1002/rth2.12505] [Medline: 34013150]
28. Li R, Hu B, Liu F, et al. Detection of bleeding events in electronic health record notes using convolutional neural network models enhanced with recurrent neural network autoencoders: deep learning approach. JMIR Med Inform. Feb 8, 2019;7(1):e10788. [doi: 10.2196/10788] [Medline: 30735140]

29. Wang X, Huang L, Xu S, Lu K. How does a generative large language model perform on domain-specific information extraction?—A comparison between GPT-4 and a rule-based method on band gap extraction. J Chem Inf Model. Oct 28, 2024;64(20):7895-7904. [doi: 10.1021/acs.jcim.4c00882] [Medline: 39375999]
30. Patel PV, Davis C, Ralbovsky A, et al. Large language models outperform traditional natural language processing methods in extracting patient-reported outcomes in inflammatory bowel disease. Gastro Hep Adv. 2024;4(2):100563. [doi: 10.1016/j.gastha.2024.10.003] [Medline: 39877865]
31. Lybarger K, Yetisgen M, Uzuner Ö. The 2022 n2c2/UW shared task on extracting social determinants of health. J Am Med Inform Assoc. Jul 19, 2023;30(8):1367-1378. [doi: 10.1093/jamia/ocad012] [Medline: 36795066]
32. Patra BG, Lepow LA, Kasi Reddy Jagadeesh Kumar P, et al. Extracting social support and social isolation information from clinical psychiatry notes: comparing a rule-based natural language processing system and a large language model. J Am Med Inform Assoc. Jan 1, 2025;32(1):218-226. [doi: 10.1093/jamia/ocae260] [Medline: 39423850]

Abbreviations

CNS: central nervous system
Dx Code: diagnosis code
ED: emergency department
EHR: electronic health record
FDA: Food and Drug Administration
***ICD-10*:** *International Statistical Classification of Diseases, Tenth Revision*
INR: International Normalized Ratio
ISTH: International Society on Thrombosis and Haemostasis
LLM: large language model
MIMIC-III: Medical Information Mart for Intensive Care
NLP: natural language processing
OHNLP: Open Health Natural Language Processing
PRBC: packed red blood cells

Edited by Andrew Coristine; peer-reviewed by Marie-Annick Le-Pogam, Sophia Ananiadou; submitted 09.Nov.2025; final revised version received 23.Sep.2026; accepted 24.Sep.2026; published 20.Nov.2026

Please cite as:
Draper EW, Anderson JA, Shah S, Brown L, Hanson MR, Negley KD, Lutz JM, Anderson AE, O'Meara JG, Jakowenko ND, McBane RD
Comparison of Rule-Based Natural Language Processing and Diagnosis Code Approaches to Detecting Bleeding in the Electronic Health Record: Algorithm Validation Within a Retrospective Cohort Study
JMIR Cardio 2026;10:e87424
URL: https://cardio.jmir.org/2026/1/e87424
doi: 10.2196/87424

© Evan W Draper, Jan A Anderson, Surbhi Shah, Leilia Brown, Margaret R Hanson, Kristin D Negley, Jennifer M Lutz, Andrea E Anderson, John G O'Meara, Nicholas D Jakowenko, Robert D McBane. Originally published in JMIR Cardio (https://cardio.jmir.org), 20.Nov.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Cardio, is properly cited. The complete bibliographic information, a link to the original publication on https://cardio.jmir.org, as well as this copyright and license information must be included.