A comparative cross-sectional study of Natural Language Processing and ICD-10 Coding for detecting bleeding events in discharge summaries

Frederic Gaspar, Mehdi Zayerne, Claire Coumau, Elliott Bertrand, Marie Bettex, Marie Annick Le Pogam, Chantal Csajka

Submitted to: Journal of Medical Internet Research
on: October 22, 2024

Table of Contents

A comparative cross-sectional study of Natural Language Processing and ICD-10 Coding for detecting bleeding events in discharge summaries

Frederic Gaspar[1, 2, 3*]; Mehdi Zayerne[4]; Claire Coumau[1, 2, 5]; Elliott Bertrand[4]; Marie Bettex[6]; Marie Annick Le Pogam[6*]; Chantal Csajka[2, 5, 3*]

[1]Center for Research and Innovation in Clinical Pharmaceutical Sciences Lausanne CH
[2]School of Pharmaceutical Sciences University of Geneva Geneva CH
[3]Effixis SA Lausanne CH
[4]Institute of Pharmaceutical Sciences of Western Switzerland University of Geneva, University of Lausanne, Geneva and Lausanne CH
[5]Center for Primary Care and Public Health (Unisanté) Department of Epidemiology and Health Systems University of Lausanne Lausanne CH
[*]these authors contributed equally

Corresponding Author:
Chantal Csajka

Abstract

Background: Bleeding adverse drug events (ADEs), particularly among older adult patients on antithrombotic therapy, are a significant concern in hospital settings. These events often go undetected using traditional rule-based methods relying on structured data from electronic medical records, underscoring the need for more effective detection approaches.

Objective: This study aimed to develop and evaluate a natural language processing (NLP) model to accurately detect and categorise bleeding events in older adult inpatients' discharge summaries. Specifically, it would identify ADEs related to antithrombotic therapy and compare the NLP model's performance with Boolean algorithms based on International Classification of Diseases, 10th Revision (ICD-10) codes.

Methods: Clinicians manually annotated 400 discharge summaries, comprising 65,706 sentences, into four categories: 'no bleeding', 'clinically significant bleeding', 'severe bleeding' and 'history of bleeding'. These annotations were used to train and validate two detection models: an NLP model using binary logistic regression and support vector machine classifiers and a rule-based model using ICD-10 codes specific to bleeding ADEs. We assessed both models' performance using accuracy, precision, recall, F1 score and the area under the curve (AUC) from receiver operating characteristic (ROC) analysis. Manual annotations served as the gold standard.

Results: The NLP model outperformed the rule-based model, especially in identifying 'clinically significant' and 'severe bleeding'. The NLP model achieved macro-averages of 0.81 for accuracy and 0.80 for the F1 score. It also demonstrated high precision in distinguishing current bleeding ADEs from past ones, with a strong true positive rate and minimal false positives.

Conclusions: This study highlights a significant advance in using artificial intelligence for healthcare, with the NLP model surpassing traditional ICD-10 coding for detecting bleeding ADEs in electronic medical records. The NLP model provides a more precise tool for clinical decision-making involving older adult patients on antithrombotic therapy.

(JMIR Preprints 22/10/2024:67837)
DOI: https://doi.org/10.2196/preprints.67837

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

A comparative cross-sectional study of Natural Language Processing and ICD-10 Coding for detecting bleeding events in discharge summaries

Frédéric Gaspar[1 2 3], Mehdi Zayerne[4], Claire Coumau[1 2 3], Zachary Schillaci[4], Marie Bettex[5], Elliott Bertrand[4], Marie-Annick Le Pogam*[5], Chantal Csajka*[# 1 2 3] and the SwissMADE study group.

* Contributed equally

[1] Center for Research and Innovation in Clinical Pharmaceutical Sciences, Lausanne University Hospital and University of Lausanne, Lausanne, Switzerland.

[2] School of Pharmaceutical Sciences, University of Geneva, Geneva, Switzerland.

[3] Institute of Pharmaceutical Sciences of Western Switzerland, University of Geneva, University of Lausanne, Geneva and Lausanne, Switzerland.

[4] Effixis SA, Lausanne, Switzerland

[5] Center for Primary Care and Public Health (Unisanté), Department of Epidemiology and Health Systems, University of Lausanne, Lausanne, Switzerland.

SwissMADE study group: Patrick E. Beeler, Bernard Burnand, Nicola Colic, Nathalie Casati, Vasiliki Foufi, Jean Philippe Goldman, Christophe Gaudet-Blavignac, Pierre-Olivier Lang, Angela Lisibach, Christian Lovis, Monika Lutters, Fabio Rinaldi and Arnaud Robert.

Conflicts of Interest: The authors declared no competing interests for this work.

Supporting Information: This study was supported by Swiss National Science Foundation grant No. 407440_167381.

Article type: Original article

Words: 4459 (main text), (291; abstract)

[#] corresponding author: Prof Chantal Csajka, Center for Research and Innovation in Clinical Pharmaceutical Sciences, Lausanne University Hospital and University of Lausanne, Lausanne, Switzerland, Rue du Bugnon 19, 1011 Lausanne/ Switzerland, Tel.: +4121 314 42 63, E-mail: chantal.csajka@chuv.ch

Abstract

Background: Bleeding adverse drug events (ADEs), particularly among older adult patients on antithrombotic therapy, are a significant concern in hospital settings. These events often go undetected using traditional rule-based methods relying on structured data from electronic medical records, underscoring the need for more effective detection approaches.

Objectives: This study aimed to develop and evaluate a natural language processing (NLP) model to accurately detect and categorise bleeding events in older adult inpatients' discharge summaries. Specifically, it would identify ADEs related to antithrombotic therapy and compare the NLP model's performance with Boolean algorithms based on International Classification of Diseases, 10th Revision (ICD-10) codes.

Methods**:** Clinicians manually annotated 400 discharge summaries, comprising 65,706 sentences, into four categories: 'no bleeding', 'clinically significant bleeding', 'severe bleeding' and 'history of bleeding'. These annotations were used to train and validate two detection models: an NLP model using binary logistic regression and support vector machine classifiers and a rule-based model using ICD-10 codes specific to bleeding ADEs. We assessed both models' performance using accuracy, precision, recall, F1 score and the area under the curve (AUC) from receiver operating characteristic (ROC) analysis. Manual annotations served as the gold standard.

Results**:** The NLP model outperformed the rule-based model, especially in identifying 'clinically significant' and 'severe bleeding'. The NLP model achieved macro-averages of 0.81 for accuracy and 0.80 for the F1 score. It also demonstrated high precision in distinguishing current bleeding ADEs from past ones, with a strong true positive rate and minimal false positives.

Conclusions**:** This study highlights a significant advance in using artificial intelligence for healthcare, with the NLP model surpassing traditional ICD-10 coding for detecting bleeding ADEs in electronic medical records. The NLP model provides a more precise tool for clinical decision-making involving older adult patients on antithrombotic therapy.

Keywords: decision support systems (clinical), deep learning, electronic medical record, haemorrhage, international classification of diseases, ICD, machine learning, natural language processing

Introduction

Adverse drug events (ADEs) are a significant patient safety issue, particularly among older adult inpatients. Globally, ADEs are estimated to affect 10% to 40% of hospitalised patients, contributing to increased morbidity, mortality and healthcare costs [1-3]. Among older adult patients, who are often treated using complex medication regimens, the risk of ADEs is even higher due to age-related physiological changes and a higher prevalence of polypharmacy [4, 5]. Data on the incidence and impact of ADEs in Switzerland's hospitals are sparse, however, making it difficult to fully assess the problem's scope [6].

Antithrombotic therapy, commonly prescribed to prevent thrombotic events, significantly increases the risk of bleeding by inhibiting normal clotting mechanisms. Studies have shown that approximately 36% of older adult inpatients on antithrombotic therapy experience bleeding complications, which can lead to extended hospital stays and increased morbidity and mortality [5]. The widespread use of polypharmacy in this population further compounds the risk of drug interactions, contributing to ADEs [7]. In Swiss hospitals, the timely and accurate detection of bleeding events is considered crucial to improve patient outcomes and ensuring safer care [8].

Electronic medical records provide an opportunity to automate the detection of ADEs like bleeding. Bleeding events are commonly identified through structured data, particularly via the diagnostic codes in the International Classification of Diseases (ICD), which are frequently used for billing purposes. However, ICD codes often lack the specificity required to capture the complexity and nuances of bleeding ADEs [9-11]. Research has shown that ICD codes frequently underreport ADEs, with sensitivities below 50% in many cases, leading to an incomplete picture of patient safety [12]. Additionally, coding algorithms for detecting ADEs usually exhibit low sensitivity and precision, and there is no universally accepted set of ICD-10 codes or algorithms that ensures the consistent identification of bleeding ADEs in administrative data [13]. Although a manual review of medical records can be more accurate, it is labour-intensive and impractical for widespread use [10].

Natural language processing (NLP), a branch of artificial intelligence, provides a scalable solution to the automated extraction and classification of information on bleeding ADEs from unstructured text, such as inpatient discharge summaries and clinical notes [14, 15]. These notes often contain detailed narrative descriptions of clinical events, like "non-glomerular microhaematuria" or "no visible bleeding at the anamnesis", which billing codes might miss [16]. NLP models can detect key clinical information buried within these narrative notes, providing more accurate insights into patients' conditions than frequently used methods like ICD coding. Previous studies have demonstrated that NLP can detect ADEs from clinical notes with accuracies as high as 85% to 90%, significantly outperforming standard methods [17-20]. By leveraging NLP and integrating it into hospital workflows, healthcare professionals can improve the surveillance of ADEs, make more timely interventions and, ultimately, provide more responsive, personalised patient care [21].

In this study, conducted within the framework of the Swiss Monitoring of Adverse Drug Events (SwissMADE) project [8], we hypothesised that an NLP-based approach would be more effective than ICD code-based algorithms for detecting and categorising bleeding ADEs among older adult inpatients receiving antithrombotic therapy at Lausanne University Hospital. The primary objective was to develop an NLP model capable of identifying bleeding ADEs from the discharge summaries of older adult inpatients hospitalised in 2015 and 2016 and to categorise these events based on their timing (i.e. before admission or during the hospital stay) and severity (clinically significant haemorrhage or severe haemorrhage). The secondary objective was to compare the NLP model's performance against standard ICD-10-based algorithms and identify the most effective automated method for detecting bleeding ADEs in Switzerland's healthcare context.

Methods

Study design

We conducted a secondary analysis of unstructured data in the electronic medical records investigated by the SwissMADE study, a multicentre, cross-sectional study that used retrospective medical data from four large Swiss hospitals [8].

Study population and dataset selection

The dataset comprised the discharge summaries of patients aged 65 or older who were hospitalised for more than 24 hours in 2015 and 2016 and received at least one antithrombotic medication during their stay. These summaries also included administrative data, such as ICD-10-GM diagnostic codes (i.e. ICD 10th version, German modification). A detailed description of the SwissMADE study's methods has been published previously [8].

Of 7,513 discharge summaries examined, an unsupervised machine learning approach identified 400 as likely to contain bleeding ADEs (Figure S1 in supplementary materials). This approach involved text scanning, thematic aggregation, and data extraction. The study generated unique sentence embeddings by integrating Bidirectional Encoder Representations from Transformers (BERT) into the Sentence Transformer library [22]. Techniques such as Uniform Manifold Approximation and Projection (UMAP) [23] and Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) [24] were then applied to organise these embeddings into 'clusters' of bleeding ADEs. This methodology was instrumental in selecting the 400 discharge summaries most relevant to the study.

Annotation of clinical documents

The 400 discharge summaries were first annotated by clinicians and then divided into a training set (280 summaries) and a test set (120 summaries). The distribution of summaries was randomised to ensure the sample remained representative of the overall population of hospitalised patients.

Three clinicians independently annotated each discharge summary using four predefined labels:

(i) Presence of severe bleeding: we used this label when a discharge summary explicitly identified severe bleeding, either by using the term 'severe' or by describing conditions that meet the criteria for severe bleeding, such as fatal bleeding, bleeding at critical sites (e.g. intracranial, intraspinal), a drop in haemoglobin of ≥ 20 g/L or transfusion of ≥ 2 units of blood, as defined by the

International Society on Thrombosis and Haemostasis (ISTH) [25].

(ii) Presence of clinically significant bleeding: we used this label when bleeding was mentioned in the clinical documentation but did not meet the criteria for severe bleeding.

(iii) History of bleeding: we applied this label when a discharge summary mentioned bleeding in the patient's medical history before their hospital admission.

(iv) Absence of bleeding: we used this label when a discharge summary did not mention bleeding.

A fourth clinician resolved any disagreements, and this classification was used as the gold standard for training the machine learning model. Fleiss' kappa coefficient, calculated from 30 summaries, showed a 96% agreement between clinicians, allowing us to shift to a single-reviewer approach. Only discharge summaries signed by an attending physician were included, ensuring data credibility.

Development of the NLP-based classifier

The development method comprised three phases: segmenting discharge summaries into sentences, classifying those sentences and aggregating them at the document level.

Phase 1: Segmentation

Sentences were segmented from their discharge summaries using the pre-trained French spaCy model [19]. To reduce the impact of abnormal or incomplete sentences, any sentence containing fewer than three characters was omitted.

Phase 2: Classification Process

Class imbalance within the dataset (i.e. few summaries mentioning bleeding ADEs versus many without an event) was addressed by implementing a three-stage sequential classification process: first, binary classification; second, multi-class classification; and third, a final binary classification. This structured approach allowed us to manage class frequency imbalances carefully, ensuring accurate categorisation. Figure 1 illustrates the multi-stage classification process for identifying bleeding ADEs in clinical narratives.

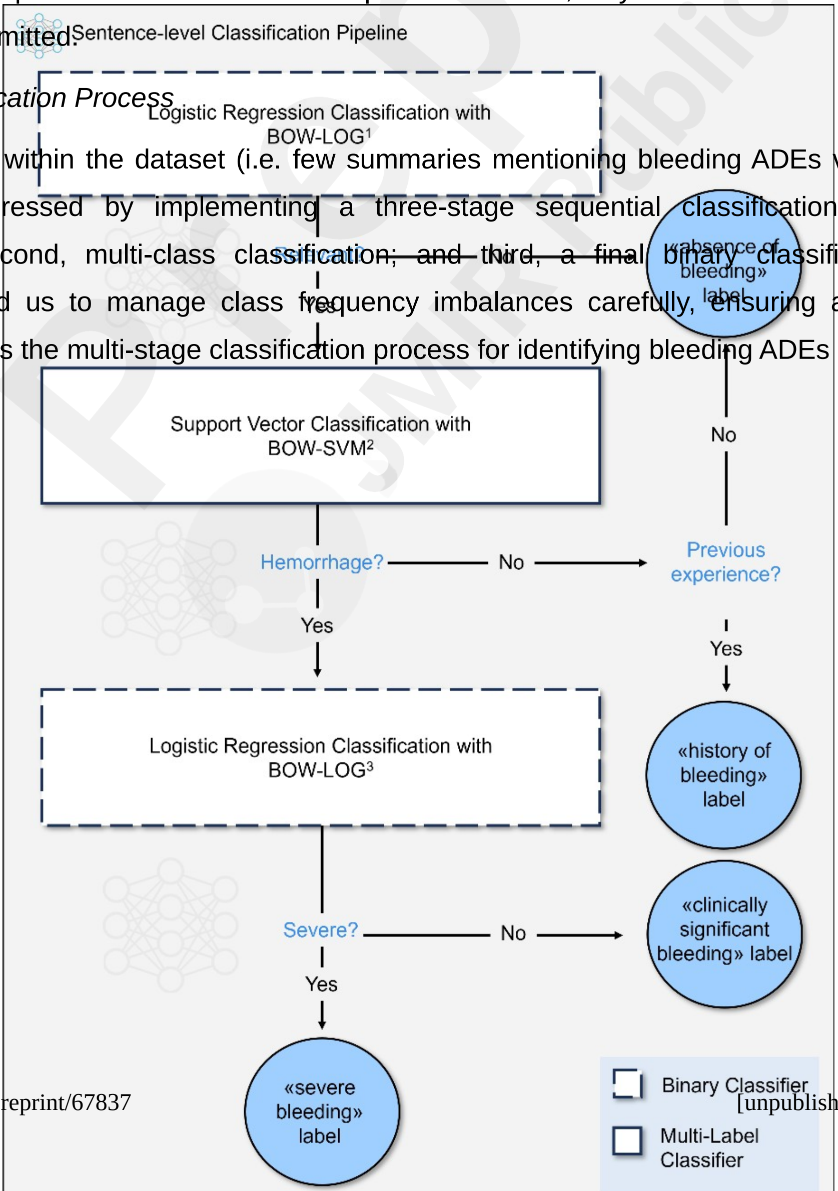
Sentence-level Classification Pipeline
Logistic Regression Classification with BOW-LOG¹
Relevant?
No
«absence of bleeding» label
Yes
Support Vector Classification with BOW-SVM²
No
Hemorrhage?
No
Previous experience?
Yes
Yes
Logistic Regression Classification with BOW-LOG³
«history of bleeding» label
Severe?
No
«clinically significant bleeding» label
Yes
«severe bleeding» label
Binary Classifier
Multi-Label Classifier

Figure 1: Sentence-level classification pipeline for bleeding event detection. [1] BOW-LOG : Logistic regression classification using a Bag of Words model, [2] BOW-SVM : Support Vector Machine classification using Bag of Words, [3] BOW-LOG : Second logistic regression classification using Bag of Words to assess the severity of the bleeding.

Stage one used a binary logistic regression model to classify sentences as either containing haemorrhage-related information (labelled ‘relevant’ and given a value of 1) or not (labelled ‘irrelevant’ and given a value of 0). This reduced the number of non-relevant sentences. Stage two used a Support Vector Machine (SVM) classifier to further divide the relevant sentences into three categories: ‘irrelevant’, ‘antecedent’ or ‘haemorrhage-related’. Stage three applied a second binary classification to sentences previously identified as ‘haemorrhage-related’, categorising them as either ‘clinically significant’ (value of 0) or ‘severe’ (value of 1). The entire process used Bag-of-Words (BoW) encoding to convert the text into a format suitable for machine learning algorithms.

Phase 3: Document Aggregation

We aggregated the sentence-level classification results at the discharge summary level by grouping sentences under their corresponding ‘document id’ and combining the predictions using a union-like operation. If all the sentences in a document were labelled ‘irrelevant’, we classified the entire document as ‘irrelevant’.

Otherwise, we assigned the document labels such as ‘antecedent’, ‘clinically significant haemorrhage’ or ‘severe haemorrhage’, depending on the sentences it contained. Unlike sentences, we allowed documents to have multiple labels.

Rule-based classifier development

In parallel with the NLP approach, we developed a rule-based classifier using the ICD-10-GM codes to detect bleeding ADEs. This classifier enabled us to compare the analysis with the NLP methods. We began by compiling a comprehensive list of ICD-10 diagnostic codes related to bleeding, drawing on subdivisions defined by the ISTH and codes identified in previous studies [26-28].

We thoroughly explored ICD-10 ontologies to identify additional codes for terms like ‘bleeding’ and ‘haemorrhage’. A multidisciplinary panel of physicians, pharmacologists, pharmacists and statisticians reviewed and expanded this list, adding codes for conditions such as haemodynamic instability, drug-induced bleeding and contusions. We then categorised these codes into two mutually exclusive groups based on the ISTH’s severity criteria: ‘clinically significant bleeding’ and ‘severe bleeding’. The classification considered factors such as the site of bleeding.

The complete list of codes used appears in In tables S1 and S2, additional materials. However, unlike the NLP approach, the rule-based method was limited to just two labels due to the absence of specific ICD-10 codes for identifying a patient’s history of bleeding. Moreover, the rule-based approach did not account for timing, making it difficult to determine whether bleeding occurred before or during the hospital stay.

Model evaluation and comparison

We conducted a comparative analysis of the rule-based classifier and the NLP method to assess their effectiveness and accuracy in identifying haemorrhagic events from discharge summaries. Using the independent test dataset to ensure unbiased assessments, we applied standardised evaluation metrics, namely precision, recall, specificity and F1 score, focusing on each method’s ability to detect clinically significant and severe bleeding. We used a receiver operating characteristic (ROC) curve to evaluate the NLP model’s diagnostic capacity, measuring its overall performance through its area under the curve (AUC) [29]. We also calculated Cohen’s kappa to facilitate a comparative analysis of the methods’ detection accuracies [30].

All statistical analyses were performed using Python software (version 3.9), thus ensuring a robust computational environment. We calculated both micro- and macro-averages to provide a comprehensive evaluation of the classifiers’ performances. Micro-averages were computed at the sentence level, measuring overall performance across all sentences, while macro-averages were calculated at the document level, giving equal weight to each document regardless of the number of sentences it contained.

Results

A total of 400 discharge summaries were analysed, comprising 65,706 annotated sentences. Of these, 47,100 sentences were allocated to the training set and 18,606 to the test set. Detailed demographic and clinical characteristics of the hospital stays associated with each dataset are presented in Table 1.

Table 1: Training and testing set patients' characteristics.

Variable	Training set	Test set
Discharge summaries (n)	280	120
Sentences (n)	47100	18606
Unique patients (n)	270	120
Length of stay, median days [range]	15 [1–82]	13 [1–60]
Female, n (%)	111 (41%)	89 (47%)
Age, median years [range]	81 [65–98]	79 [65–97]
ICU admissions, n (%)	17 (6%)	4 (3%)
Modes of admission		
Emergency, n (%)	214 (76%)	89 (74%)
Planned, n (%)	49 (18%)	20 (17%)
Internal transfer, n (%)	15 (5%)	6 (5%)
Transfer within 24 hours, n (%)	2 (1%)	5 (4%)

Sentence-level analysis revealed a predominance of 'irrelevant' annotations, reflecting the large amount of information in discharge summaries unrelated to bleeding. However, class distribution was more balanced at the document level, demonstrating the complexity of clinical documentation where multiple annotations often coexist within a single summary. Table 2 provides the detailed distribution of these categories.

Table 2: Training and test set class balance at the sentence and document levels.

Classification label	Sentence-level[a]		Document-level[b]	
	Training set, n (%)	Test set, n (%)	Training set, n (%)	Test set, n (%)
Irrelevant (absence of bleeding)	45897 (97.45%)	18118 (97.38%)	103 (36.79%)	44 (36.67%)
History of bleeding	154 (0.33%)	58 (0.31%)	67 (23.93%)	22 (18.33%)
Clinically significant bleeding	900 (1.91%)	373 (2.00%)	141 (50.36%)	60 (50.00%)
Severe bleeding	149 (0.32%)	57 (0.31%)	77 (27.50%)	31 (25.83%)

[a] Sentence-level: Indicated the frequency and proportion of each classification label per individual sentence; [b] Document-level: Indicated the frequency and proportion of documents containing at least one instance of the respective classification label.

The NLP model demonstrated strong classification capabilities, achieving over 85% accuracy across all categories at the document level. It also showed robust performance, with a precision exceeding 72% across categories and a recall of 98% for 'irrelevant' instances. F1 scores indicated balanced performance despite class imbalances, highlighting the model’s ability to manage diverse data distributions. A detailed summary of the performance metrics for our multi-label classification model is provided in Table 3.

Table 3: Detailed performance metrics of the multi-label classification model.

Metric	Irrelevant	History of bleeding	Clinically significant bleeding	Severe bleeding	Macro-average[a]	Micro-average[b]
Accuracy	0.83	0.68	0.86	0.89	0.81	0.84
Precision	0.81	0.72	0.87	0.92	0.83	0.85
Recall	0.98	0.88	0.59	0.31	0.69	0.71
F1 Score	0.89	0.65	0.88	0.70	0.78	0.80

As Figure 2 shows, the ROC curve analysis further highlighted the model’s diagnostic accuracy. In stage one, the model achieved an AUC of 0.91 for classifying sentences as either ‘irrelevant’ or ‘potentially haemorrhage-related’, effectively filtering out irrelevant data. In stage two, it refined these classifications into the ‘irrelevant’ and ‘antecedent’ categories, with AUCs of 0.88 and 0.83, respectively. Stage three focused on distinguishing ‘clinically significant haemorrhage’ from ‘severe haemorrhage’, achieving an AUC of 0.94. Overall, the ROC curves demonstrated the model’s consistently high performance across every stage, with elevated AUC values reflecting its strong ability to discriminate between classes.

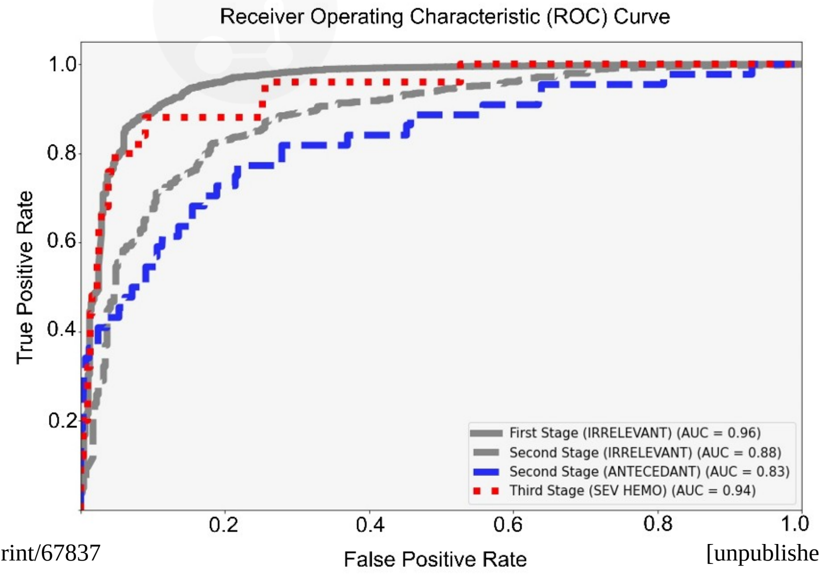

Figure 2: ROC curves showing the diagnostic performance of the multi-stage classification model at various thresholds, illustrating the trade-off between the true positive rate (sensitivity) and the false positive rate (1 - specificity).

The rule-based classifier, while simpler than the multi-label NLP model, showed great precision in identifying ‘clinically significant bleeding’ events. Table 4 provides a detailed comparison of the algorithm’s metrics and those of the NLP model.

Table 4: Detailed performance metrics of the rule-based classification model.

Metric	Irrelevant	Antecedent[a]	Clinically significant bleeding	Severe bleeding	Macro-average	Micro-average[b]
Accuracy	0.81	-	0.86	0.74	0.80	-
Precision	0.80	-	0.94	0.50	0.75	-
Recall	0.95	-	0.77	0.03	0.58	-
F1 Score	0.87	-	0.84	0.06	0.59	-

[a] Metrics for the ‘Antecedent’ category are not provided due to the absence of corresponding ICD-10 codes; [b] The 'Micro-average' was not calculated as the rule-based model uses ICD-10 codes linked to hospital stays.

The rule-based classifier achieved a precision score of 0.94 for ‘clinically significant bleeding’, highlighting its accuracy in detecting these events. However, its performance in identifying ‘severe bleeding’ was significantly weaker, with a recall of only 0.03. This low recall indicates that while the model could detect severe haemorrhage when present, it also frequently missed such events.

For ‘clinically significant bleeding’, the classifier relied heavily on frequently used ICD-10 codes, such as K92.2 (Gastrointestinal haemorrhage, unspecified), R31 (Haematuria, unspecified) and K26.4 (Gastric ulcer, acute with haemorrhage), which contributed to its high precision. In contrast, codes associated with ‘severe bleeding’, including R57.1 (Shock due to haemorrhage) and I85.3 (Oesophageal varices with bleeding), were less common in the dataset, resulting in poorer performance for this category. The rule-based model achieved an F1 score of 0.84 for ‘clinically significant bleeding’ but only 0.06 for ‘severe bleeding’, underscoring the disparity in its ability to handle these two categories.

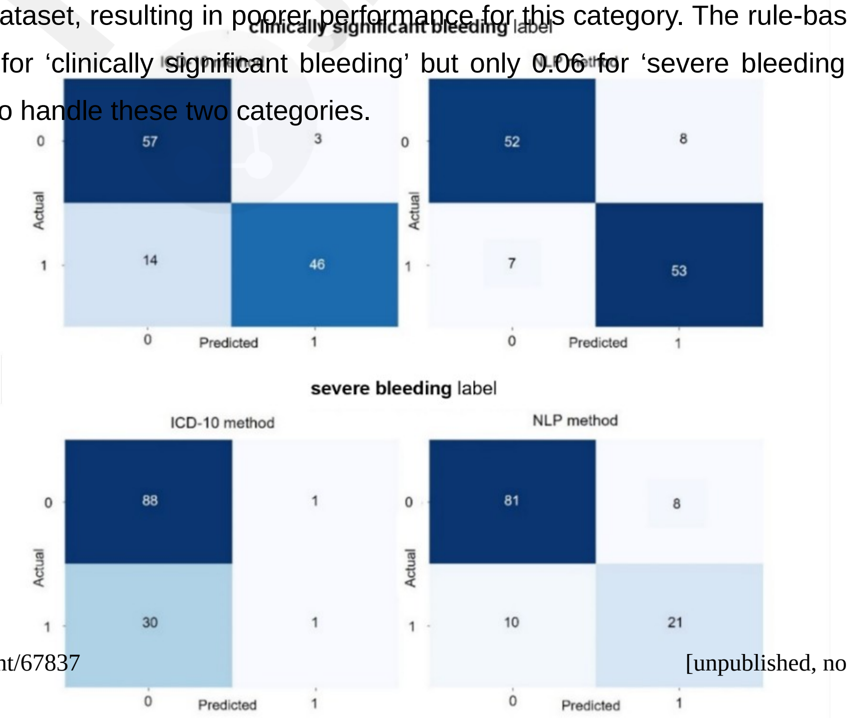

Figure 3: Comparative analysis of confusion matrices in haemorrhage detection using NLP and ICD coding methods. This figure contrasts the performance of the NLP model and the ICD coding approach, showing counts of true positives, true negatives, false positives and false negatives, providing a comparative evaluation of both methods.

Discussion

This study developed and validated an NLP-based model for detecting haemorrhagic events from clinical narratives, achieving a high level of accuracy. The model demonstrated 91% accuracy at the sentence level and 88% at the document level. Compared to traditional ICD-10 code-based methods, the NLP model provided more nuanced and precise detection of haemorrhagic events, effectively capturing details that ICD-10 codes often miss, particularly in cases involving secondary conditions or multiple types of bleeding ADEs. These findings underscored NLP's potential to improve the detection and management of adverse events and make healthcare more efficient.

One of the NLP model's primary strengths lies in its high precision and recall, even when dealing with class imbalances, such as differentiating between 'clinically significant' and 'severe haemorrhages'. The model excelled at interpreting complex clinical data, including negations (e.g. "no evidence of bleeding") and secondary conditions, which traditional methods based solely on ICD-10 codes typically struggle to identify [9]. By incorporating these nuances into its analysis, the NLP model provided a more comprehensive understanding of the patient's clinical condition. The ability to process unstructured clinical narratives offers significant potential for real-time decision-making, allowing clinicians to identify bleeding events more accurately and promptly. This capability could lead to more effective monitoring and treatment strategies, ultimately improving patient outcomes in hospital settings [10, 11, 31].

In comparison to the existing literature, the NLP model's performance stands out. The model's accuracy, precision, recall and F1 scores compared favourably with other models in the literature, such as those using deep learning methods like biLSTM-CRF or transformer-based models like BERT [32]. While these models have demonstrated strong performance in similar applications, the present study's use of multi-label classification distinguishes it, as it enables the detection of overlapping conditions—clinically significant and severe haemorrhages—within a single document. Previous research has also highlighted the challenges of handling information across time, a limitation common to many NLP models. Although our model still has difficulties in this area, especially distinguishing between recent and past events, it marks an important step toward addressing this issue. These challenges suggest that further refinements to temporal reasoning models are needed to improve their accuracy in identifying past or contemporary events [33, 34].

The limitations of ICD-10 coding, particularly its inability to capture nuanced clinical events, are well-documented in the literature. Similar to findings by Johnson *et al*. [35], our study revealed that reliance on a few key ICD-10 codes, such as K92.2 (Gastrointestinal haemorrhage, unspecified) and R57.1 (Shock due to haemorrhage), could contribute to low recall for severe haemorrhages. While useful for administrative purposes, these codes lack the specificity needed to accurately classify complex clinical conditions, especially when secondary or co-occurring conditions are present. Thus, ICD-10 coding limitations highlight the need for advanced approaches, such as NLP, that ensure more accurate detection and classification of complex clinical events [17].

One notable strength of our NLP model was its superior performance in handling negation, such as phrases like "no source of bleeding" [36]. Many previous studies have struggled with accurately interpreting negation, leading to overestimations of ADEs. By successfully distinguishing between positive and negative statements, our model significantly reduced false positives, enhancing its clinical utility [37]. The capacity to manage linguistic nuances is essential in clinical environments, where misinterpretation can lead to unnecessary interventions [38].

Another key advantage of our study was the use of a multi-label classification approach, which provides a flexible, accurate method of identifying multiple overlapping conditions, a significant plus over standard single-label classifiers or rule-based methods [39]. This could be particularly important in clinical cases where a single patient presents with both clinically significant and severe bleeding or where a history of bleeding must be considered alongside current bleeding events [40]. This feature of the NLP model demonstrates its adaptability to complex real-world scenarios [41].

The present study had some limitations that should be acknowledged. The dataset used consisted solely of discharge summaries from Lausanne University Hospital, which may limit its generalisability. Data from a single, large, tertiary care institution may not represent the variety of clinical settings and regions in which the model could be deployed. Additionally, the over-representation of certain ICD-10 codes, such as K92.2, likely contributed to the model's high precision for detecting clinically significant haemorrhages. The dataset's class imbalance, particularly the lower frequency of cases of severe haemorrhage, may have biased the results, underscoring the need for more diverse and balanced data sources [42]. Expanding the baseline dataset could help improve the model's robustness and ability to generalise across different hospital environments [43].

Although the model performed well overall, its accuracy dropped to 70% for detecting severe haemorrhages. This decrease was largely due to an over-reliance on numerical data, such as haemoglobin and haematocrit levels, which the model frequently misinterpreted as indicative of severe bleeding. This led to a number of false positives, particularly when laboratory values came close to terms related to bleeding. This issue suggests that while the NLP model was effective in many cases, it may require additional refinement to better differentiate between clinically relevant data and incidental mentions of numerical values. Additionally, the model had difficulty to interpret time data, making it difficult to distinguish between recent and past bleeding events. This limitation is critical for many clinical applications where understanding the timing of an event is essential for accurate diagnosis and treatment [44].

NLP models, particularly those using deep learning or transformer-based architectures, require significant computational resources for training and deployment. Although our model was relatively efficient, scalability remains challenging, particularly for real-time clinical applications requiring continuous updates and large datasets. Furthermore, although integrating large language models like GPT-3 or BERT holds the promise of improved performance, it also introduces concerns around computational cost and whether sensitive patient data can be handled securely [40, 42, 45-47]. These practical challenges will have to be addressed before the widespread adoption of NLP models in clinical settings [44, 48].

Conclusion and perspectives

Despite some challenges, this study contributed to the growing body of research supporting the use of natural language processing (NLP) in healthcare, specifically for the detection of adverse drug events like haemorrhages. The model developed in this study outperformed standard International Classification of Diseases, 10th Revision (ICD-10) coding systems by capturing nuanced clinical information that is often missed, such as negations and secondary conditions. The use of multi-label classification added a layer of flexibility, allowing the model to handle complex clinical cases with overlapping symptoms and conditions. This advance has great potential for improving real-time clinical decision-making and enhancing patient care.

Looking ahead, future research should focus on refining the model by expanding the dataset to include clinical records from multiple hospitals and medical settings, thus ensuring greater diversity and representativeness. Incorporating additional data sources, such as laboratory values, imaging reports and progress notes, could further enhance the model’s accuracy and robustness. Integrating more advanced NLP techniques, such as transformer-based models (e.g. BERT, GPT), may also improve the model’s ability to handle complex language structures and implicit temporal information.

Real-world validation of the model in different healthcare settings will be essential for assessing its scalability and generalisability. This should include testing across different types of medical records, such as emergency department notes or outpatient summaries, to evaluate its performance in various clinical environments. Lastly, future studies should explore combining structured and unstructured data, such as laboratory results and ICD codes, to create a more comprehensive diagnostic tool for predicting adverse drug events like bleeding. Such an approach could provide a more comprehensive and more accurate picture of a patient’s condition, ultimately improving clinical outcomes.

References

1. Cook, D.J., et al., *The attributable mortality and length of intensive care unit stay of clinically important gastrointestinal bleeding in critically ill patients.* Crit Care, 2001. **5**(6): p. 368-75.
2. Krahenbuhl-Melcher, A., et al., *Drug-related problems in hospitals: a review of the recent literature.* Drug Saf, 2007. **30**(5): p. 379-407.
3. Berger, J.S., et al., *Bleeding, mortality, and antiplatelet therapy: results from the Clopidogrel for High Atherothrombotic Risk and Ischemic Stabilization, Management, and Avoidance (CHARISMA) trial.* Am Heart J, 2011. **162**(1): p. 98-105 e1.
4. Classen, D.C., et al., *Computerized surveillance of adverse drug events in hospital patients. 1991.* Qual Saf Health Care, 2005. **14**(3): p. 221-5; discussion 225-6.
5. Kanagaratnam, L., et al., *[Serious Adverse Drug Reaction and Their Preventability in the Elderly Over 65 Years].* Therapie, 2015. **70**(5): p. 477-84.
6. Beeler, P.E., T. Stammschulte, and H. Dressel, *Hospitalisations Related to Adverse Drug Reactions in Switzerland in 2012-2019: Characteristics, In-Hospital Mortality, and Spontaneous Reporting Rate.* Drug Saf, 2023. **46**(8): p. 753-763.
7. Long, S.J., et al., *What is known about adverse events in older medical hospital inpatients? A systematic review of the literature.* Int J Qual Health Care, 2013. **25**(5): p. 542-54.
8. Gaspar, F., et al., *Automatic Detection of Adverse Drug Events in Geriatric Care: Study Proposal.* JMIR Res Protoc, 2022. **11**(11): p. e40456.
9. Wilchesky, M., R.M. Tamblyn, and A. Huang, *Validation of diagnostic codes within medical services claims.* J Clin Epidemiol, 2004. **57**(2): p. 131-41.
10. Bates, D.W., et al., *Detecting adverse events using information technology.* J Am Med Inform Assoc, 2003. **10**(2): p. 115-28.
11. Hohl, C.M., et al., *ICD-10 codes used to identify adverse drug events in administrative data: a systematic review.* J Am Med Inform Assoc, 2014. **21**(3): p. 547-57.
12. Hazlehurst, B., et al., *Detecting possible vaccination reactions in clinical notes.* AMIA Annu Symp Proc, 2005. **2005**: p. 306-10.
13. Shehab, N., et al., *Assessment of ICD-10-CM code assignment validity for case finding of outpatient anticoagulant-related bleeding among Medicare beneficiaries.* Pharmacoepidemiol Drug Saf, 2019. **28**(7): p. 951-964.
14. Nadkarni, P.M., L. Ohno-Machado, and W.W. Chapman, *Natural language processing: an introduction.* Journal of the American Medical Informatics Association, 2011. **18**(5): p. 544-551.
15. Yim, W.W., et al., *Natural Language Processing in Oncology: A Review.* JAMA Oncol, 2016. **2**(6): p. 797-804.
16. Mehta, N. and A. Pandit, *Concurrence of big data analytics and healthcare: A systematic review.* International journal of medical informatics, 2018. **114**: p. 57-65.
17. Li, R., 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, 2019. **7**(1): p. e10788.
18. Tang, B., et al. *Clinical entity recognition using structural support vector machines with rich features*. in *Proceedings of the ACM sixth international workshop on Data and text mining in biomedical informatics*. 2012.
19. Hao, T., et al., *Health natural language processing: methodology development and applications.* JMIR medical informatics, 2021. **9**(10): p. e23898.
20. Hossain, E., et al., *Natural Language Processing in Electronic Health Records in relation to healthcare decision-making: A systematic review.* Comput Biol Med, 2023. **155**: p. 106649.
21. Locke, S., et al., *Natural language processing in medicine: a review.* Trends in Anaesthesia and Critical Care, 2021. **38**: p. 4-9.
22. Grootendorst, M., *BERTopic: Neural topic modeling with a class-based TF-IDF procedure.* arXiv preprint arXiv:2203.05794,

2022.

23. McInnes, L., J. Healy, and J. Melville, *Umap: Uniform manifold approximation and projection for dimension reduction.* arXiv preprint arXiv:1802.03426, 2018.
24. Birant, D. and A. Kut, *ST-DBSCAN: An algorithm for clustering spatial–temporal data.* Data & knowledge engineering, 2007. **60**(1): p. 208-221.
25. Schulman, S., et al., *Definition of major bleeding in clinical investigations of antihemostatic medicinal products in non-surgical patients.* J Thromb Haemost, 2005. **3**(4): p. 692-4.
26. Walther, D., et al., *Hospital discharge data is not accurate enough to monitor the incidence of postpartum hemorrhage.* PLoS One, 2021. **16**(2): p. e0246119.
27. Hartenstein, A., et al., *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.* International Journal of Population Data Science, 2023. **8**(1).
28. Joos, C., et al., *Accuracy of ICD-10 codes for identifying hospitalizations for acute anticoagulation therapy-related bleeding events.* Thromb Res, 2019. **181**: p. 71-76.
29. Metz, C.E. *Basic principles of ROC analysis*. in *Seminars in nuclear medicine*. 1978. Elsevier.
30. Sim, J. and C.C. Wright, *The kappa statistic in reliability studies: use, interpretation, and sample size requirements.* Phys Ther, 2005. **85**(3): p. 257-68.
31. Sun, W., et al., *Data Processing and Text Mining Technologies on Electronic Medical Records: A Review.* J Healthc Eng, 2018. **2018**: p. 4302425.
32. Mitra, A., et al., *Bleeding Entity Recognition in Electronic Health Records: A Comprehensive Analysis of End-to-End Systems.* AMIA Annu Symp Proc, 2020. **2020**: p. 860-869.
33. Zhou, L. and G. Hripcsak, *Temporal reasoning with medical data—a review with emphasis on medical natural language processing*. Journal of biomedical informatics, 2007. **40**(2): p. 183-202.
34. Zhou, L., et al., *A temporal constraint structure for extracting temporal information from clinical narrative.* J Biomed Inform, 2006. **39**(4): p. 424-39.
35. Johnson, S.A., et al., *A comparison of natural language processing to ICD-10 codes for identification and characterization of pulmonary embolism*. Thromb Res, 2021. **203**: p. 190-195.
36. Pedersen, J.S., et al., *Deep learning detects and visualizes bleeding events in electronic health records.* Res Pract Thromb Haemost, 2021. **5**(4): p. e12505.
37. Taggart, M., et al., *Comparison of 2 Natural Language Processing Methods for Identification of Bleeding Among Critically Ill Patients*. JAMA Netw Open, 2018. **1**(6): p. e183451.
38. Zeng, Z., et al., *Natural Language Processing for EHR-Based Computational Phenotyping*. IEEE/ACM Trans Comput Biol Bioinform, 2019. **16**(1): p. 139-153.
39. Jagannatha, A., et al., *Overview of the First Natural Language Processing Challenge for Extracting Medication, Indication, and Adverse Drug Events from Electronic Health Record Notes (MADE 1.0).* Drug Saf, 2019. **42**(1): p. 99-111.
40. Haupt, C.E. and M. Marks, *AI-generated medical advice—GPT and beyond.* Jama, 2023. **329**(16): p. 1349-1350.
41. Krishnan, R., P. Rajpurkar, and E.J. Topol, *Self-supervised learning in medicine and healthcare.* Nature Biomedical Engineering, 2022. **6**(12): p. 1346-1352.
42. Dave, T., S.A. Athaluri, and S. Singh, *ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations.* Frontiers in Artificial Intelligence, 2023. **6**: p. 1169595.
43. Sallam, M., et al., *ChatGPT applications in medical, dental, pharmacy, and public health education: A descriptive study highlighting the advantages and limitations.* Narra J, 2023. **3**(1): p. e103-e103.

44. Deng, J., A. Zubair, and Y.-J. Park, *Limitations of large language models in medical applications.* Postgraduate Medical Journal, 2023. **99**(1178): p. 1298-1299.

45. Chen, Z., et al., *MEDITRON-70B: Scaling Medical Pretraining for Large Language Models.* arXiv preprint arXiv:2311.16079, 2023.

46. Yang, X., et al., *GatorTron: A Large Language Model for Clinical Natural Language Processing.* medRxiv, 2022: p. 2022.02.27.22271257.

47. Nori, H., et al., *Capabilities of gpt-4 on medical challenge problems.* arXiv preprint arXiv:2303.13375, 2023.

48. Head, C.B., et al., *Large language model applications for evaluation: Opportunities and ethical implications.* New Directions for Evaluation, 2023. **2023**(178-179): p. 33-46.

Supplementary Files