Automated identification of cancer-associated thrombosis events *via* natural language processing: a systematic review of the literature

Aidan Boyne,[1] Emily Zhou,[2] Ang Li[1]

[1]Section of Hematology-Oncology, Department of Medicine, Baylor College of Medicine, Houston, TX; [2]McGovern Medical School, University of Texas Health Science Center, Houston, TX, USA

ABSTRACT

Accurate identification of cancer-associated thrombosis (CAT) in electronic health records is essential for disease surveillance, trial design, and the development of risk stratification models. Manual chart review is impractical at scale, while ICD-code or similar coding system-based extraction often misses events, fails to distinguish incident from prevalent events, or equates bland and tumor thrombi. Natural language processing (NLP) is a promising solution, offering scalable extraction of structured CAT events directly from clinical text. We systematically reviewed the literature for NLP-based pipelines for CAT identification according to PRISMA guidelines. PubMed, Embase, and Web of Science were queried for English language studies from 2010-2025. NLP methodology, training strategy, and model performance were extracted from relevant studies. Seven studies, implementing NLP approaches ranging from lexicon or rules-based pipelines to transformer models, met inclusion criteria. Most studies developed and evaluated models using text from a single institution, and only one distinguished incident CAT events from prevalent events. Reported metrics varied between studies, though models tended to exhibit higher specificity and negative predictive value than sensitivity and positive predictive value. Overall, current NLP systems for CAT identification have achieved desired results to assist but not supplant manual chart review. Major limitations include small and institution-specific datasets, absence of external validation, and lack of distinction between incident vs prevalent events. Development of generalizable models will require large, multi-institutional datasets as the field moves towards transformer-based models, and standardized evaluation metrics with shared benchmark test sets are needed for an unbiased measure of progress.

Key words: natural language processing; venous thromboembolism; neoplasm.

Corresponding author: Ang Li, Baylor College of Medicine, One Baylor Plaza, 011DF, Houston, TX 77030, USA.
E-mail: ang.li2@bcm.edu

Contributions: all authors contributed equally to the composition of this work including review of available literature, drafting the manuscript, and providing critical edits.

Conflict of interest: The authors declare no conflict of interest.

Funding: AL, a CPRIT Scholar in Cancer Research, was supported by Cancer Prevention and Research Institute of Texas (RR190104), NIH NHLBI (K23 HL159271, R01 HL180402), American Society of Hematology (ASH) Scholar Award, and a Career Development Award from Conquer Cancer, the ASCO Foundation.

Received: 5 January 2026.
Accepted: 20 February 2026.

Publisher's note: all claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article or claim that may be made by its manufacturer is not guaranteed or endorsed by the publisher.

Bleeding, Thrombosis and Vascular Biology 2026; 5(s1):437
doi:10.4081/btvb.2026.437

Introduction

Cancer-associated thrombosis (CAT), a subset of venous thromboembolisms (VTEs), is a major cause of morbidity and mortality in patients being treated for malignancy.[1] With the global expansion of electronic health record (EHR) systems, longitudinal documentation of CAT has become more readily available. Accurate identification of initial and recurrent CAT in the EHR facilitates thrombosis surveillance, development of predictive and decision support tools, and therapeutic trials.

Manually extracting occurrences of CAT from free-form medical text, however, is time consuming and unfeasible at a large scale. Automated extraction using ICD-codes, while scalable, often cannot differentiate prevalent from incident events and struggles to distinguish bland from tumor thrombus. Natural language processing (NLP) offers a solution, allowing for extraction of structured data (i.e., the time and anatomical location of a thrombotic event) from the unstructured medical note.[2]

NLP encompasses a broad spectrum of models and techniques. Lexicon-based approaches and rules-based algorithms try to sequentially match patterns, words, or key phrases and are highly interpretable but struggle to generalize across datasets and adapt to subtle changes or errors in text. Convolutional neural networks (CNNs) encode the text by converting sequences into feature maps, extracting discriminative patterns modeled after those found in labeled training data.[3,4] CNNs are computationally efficient and effective at capturing localized patterns, but they lack an explicit mechanism for modeling sequential structure beyond a specific limited area of input. As a result, standard CNN-based models often struggle to capture global word order and long-range dependencies in text.

Recurrent neural networks (RNNs), including variants such as long short-term memory (LSTM) networks and gated recurrent units (GRUs) explicitly model sequential dependencies in text.[5] By processing tokens iteratively, RNNs maintain a hidden state that captures contextual information across time, enabling improved modeling of long-range dependencies compared to CNNs. However, their sequential nature limits parallelization and can pose challenges for training on long documents.

More recently, language models based on the transformer architecture have become the foundation of modern NLP.[6] Transformers replace recurrence with self-attention mechanisms that directly model interactions between all tokens in a sequence, allowing efficient capture of global context and long-range dependencies. Encoder-based transformer models, such as BERT and its derivatives, are typically pretrained using self-supervised objectives and then fine-tuned for downstream tasks such as entity recognition or event classification.[7]

In parallel, large language models (LLMs), typically based on decoder or encoder-decoder transformer architectures and trained on massive corpora, have shown the ability to perform a variety of NLP tasks in zero-shot or few-shot settings through prompt-based interaction.[8] While such models can sometimes perform labeling or extraction tasks without task-specific fine-tuning, their use in clinical contexts introduces challenges related to computational cost, reproducibility, and sensitivity to prompt wording.[9] The various NLP approaches employed by studies included in the review are outlined visually in Figure 1.

Prior work using NLP for identification of VTE spans the full gamut of techniques which have been applied to varying degrees of success[10]. However, only a small subset focuses on

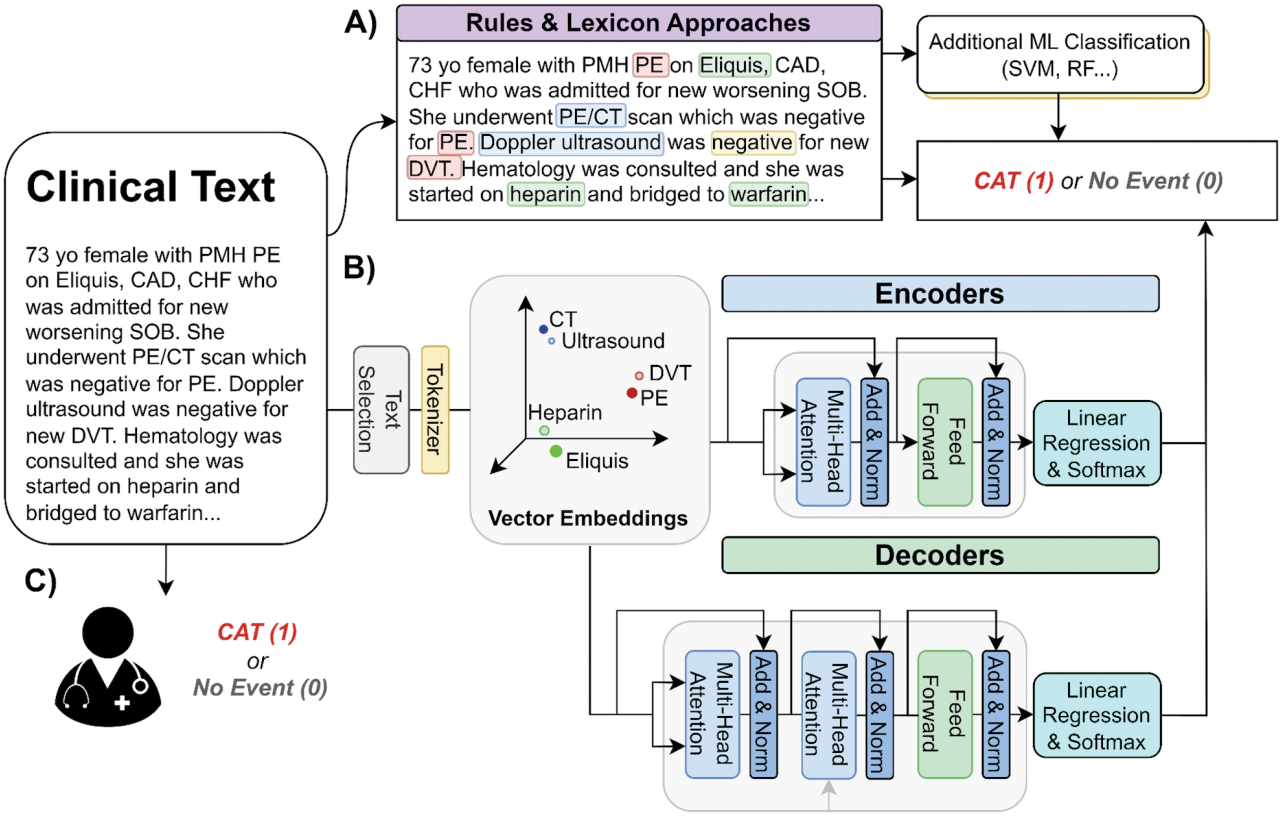

Figure 1. NLP methodologies of included studies. A) Rules and lexicon based approaches. These approaches employ dictionaries to identify key words in clinical notes, using branching logic and decision trees to determine the relationship between entities of interest. In the example above, words relating to DVT, anticoagulants, diagnostic studies, and negating words (e.g. not, negative…) are tagged and either passed to a predetermined ruleset or a machine learning algorithm for note classification. B) Transformer based approaches. Before being passed to a transformer model, notes are pre-processed to select text likely to contain CAT related information, often using similar techniques as the approaches outlined in section A. The selected text is tokenized, a process in which words and phrases are encoded numerically for processing by the transformer. Tokenized text is embedded in a vector space, which groups conceptually similar words and phrases together and encodes relationships between different model inputs. Finally, the vectorized data is passed to an encoder model (such as BERT) or a decoder model (such as one of the many consumer-grade large language models). Encoder models are designed to process bidirectional relationships in the text, including negation, temporality, and entity-event relationships. Encoders output a deterministic, fixed-length vector which can be interpreted as a probability of a new CAT event within the note. Decoder models output autoregressive text, predicting the next word or sequence of words based on the input and previous model output. The presence of a new CAT event can then be extracted from the output text. C) Manual note review provides a gold standard annotation for the evaluation of automated methods. PMH, past medical history; PE, pulmonary embolism; CAD, coronary artery disease; CHF, congestive heart failure; SOB, shortness of breath; CT, computed tomography; DVT, deep vein thrombosis; CAT; cancer associated thrombosis; SVM, support vector machine; RF, random forest.

CAT, and even fewer provide comprehensive and transparent training and validation datasets or open access to the model code for broader use. In this review, we evaluate the methodology, performance, and practical usability of current NLP systems developed for the identification of CAT events.

Methods

A systematic review following the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 guidelines was performed (Figure 2).[11] PubMed, Embase, and Web of Science were queried for peer-reviewed human studies published from 2010 to 2025 in English. References within included studies were manually screened. A search was performed in all three databases using terms "venous thromboembolism" OR "pulmonary embolism" OR "venous thrombosis" AND "neoplasm" OR "cancer" OR "malignancy" AND "natural language processing" OR "machine learning" OR "artificial intelligence" OR "text mining" in the title, abstract, or key words (Appendix A). This search strategy ensures that only NLP models built in cancer populations were selected and reviewed.

Two reviewers (AB and EZ) independently screened titles and abstracts for inclusion. Review articles, poor quality studies, studies without sufficient data, articles describing only predictive models, studies in primarily non-cancer populations, and all studies with irrelevant, duplicate, or redundant information were excluded from review. Full-text of studies meeting the above criteria were sought for retrieval. Country of origin, clinical setting (inpatient vs. ambulatory or both), dataset size and composition, NLP methodology, validation strategy and patient vs. note-level validation, gold standard comparator, and perform-

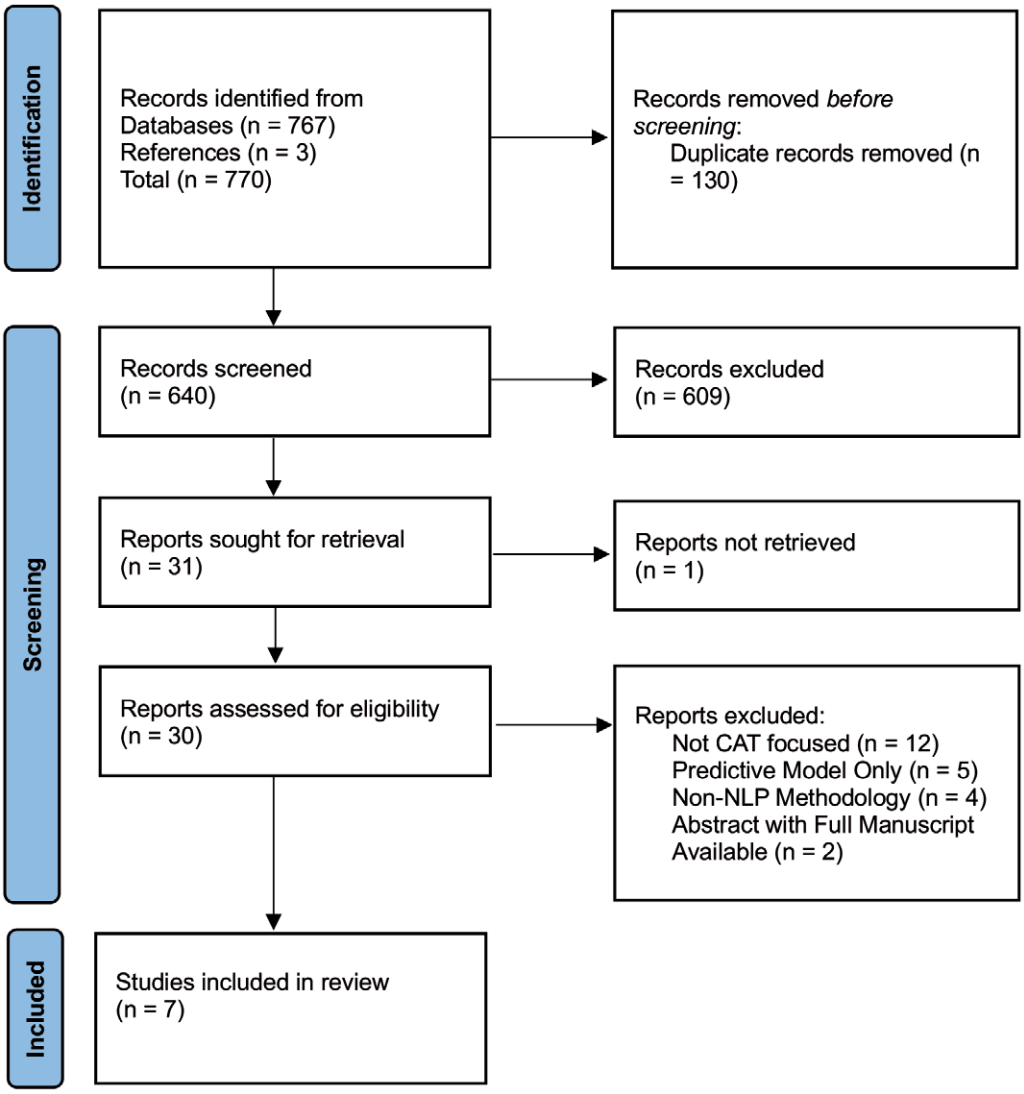

Figure 2. PRISMA diagram.

ance metrics were extracted when available. Availability of the code or model was also assessed.

Results

Included studies

A total of 7 studies (3 articles and 4 abstracts) were included (Table 1).[12-18] The majority of studies were conducted at large academic US medical centers (n=6), with one study conducted in Canada (n=1). Most studies included text originating in both the inpatient and outpatient setting. The use of medical notes varied, however, with 4 studies utilizing full clinical notes and radiology reports for CAT prediction, and 3 using only text from radiology reports.

NLP techniques

The studies represented a diverse set of approaches for automated extraction of CAT phenotype from text. Two studies used lexicon approaches with or without additional processing by machine learning (ML) models. More recent studies used proprietary ML-NLP pipelines (n=3) and transformer-based (n=2) approaches. Only one of the models was made publicly available for external use.

Model training and validation

Four of the models were not trained in a strict sense – rather, the authors reviewed notes for common patterns, words, and phrases which were included and/or emphasized in the final NLP model. Training approaches for the remaining three models were split between balanced (n=1) and unbalanced (n=2). Validation of the models was performed using data from a single institution in the majority of cases (n=6): only a single model was evaluated on both an internal and external dataset.[13] Most models also used a pre-selected validation set to assess performance (n=6), while one model used bootstrapping.[16] CAT identification was assessed on a per-note basis (e.g., given the text of a single radiology report or medical note, is CAT reported?) in three papers, and on a per-patient basis (e.g., from all notes from a single patient in a given period, is there a report of CAT at least once?) in four (Table 1). Only one study distinguished between incident and prevalent CAT during model assessment.[13] All studies used manual chart review as the gold standard comparator.

NLP annotation platforms

An essential step for model training and testing is the manual annotation of text corresponding to CAT. Several research groups and commercial entities have developed graphical user interfaces (GUIs) to facilitate annotation of clinical text data. The VTE-BERT model was trained using data annotated using the NLPMed-Portal (https://nlpmed.demo.angli-lab.com/), and the PINES model (no published performance results available) was trained using the CEDAR annotation interface (https://cedars.io/).[19] These platforms support closed system annotation of protected health information and offer automated annotation assistance through trained NLP models.

Table 1. Methodologies of included studies. All included studies analyzed notes in English. NLP, natural language processing; UMLS, unified medical language system; LLM, large language model; CLAMP, clinical language annotation, modeling, and processing; RR, radiology reports; MN, medical notes; VTE, venous thromboembolism. *External validation of CLAMP model developed by Li *et al.*[16]

First author	Format	Country	NLP technique	Clinical setting	Text source	Event level	Training approach	Training set	Validation methodology	Validation set
Chen[12]	Article	United States	Support vector machine using NLP derived UMLS terms	Ambulatory & inpatient	RR & MN	Note	Unbalanced	VTE: 502 Control: 201	Internal (no dedicated validation set)	Validated on training set
Jafari[13]	Article	United States	Fine-tuned Bio_ClinicalBERT	Ambulatory & inpatient	RR & MN	Patient	Balanced	VTE: 391 Control: 324	Internal (pre-selected) & External	Internal: 69 VTE, 319 Control External: 49 VTE, 326 Control
He[14]	Abstract	Canada	Fine-tuned LLM (unspecified)	Ambulatory & inpatient	RR	Note	Unbalanced	5500 Notes (VTE prevalence not specified)	Internal (pre-selected)	All patients treated in 2019 (N not specified)
Jagasia[15]	Abstract	United States	Rules Based Lexicon	Not specified	RR	Patient	-	-	Internal (pre-selected)	15 VTE, 148 Control
Li[16]	Article	United States	Customized Proprietary NLP Pipeline (CLAMP)	Ambulatory & inpatient	RR	Patient	-	-	Internal (bootstrapped)	894 VTE, 7626 Control
Avery[17,*]	Abstract	United States	Customized Proprietary NLP Pipeline (CLAMP)	Ambulatory & inpatient	RR & MN	Patient	-	-	External	275 Incident VTE, 389 Historical VTE, 76 Control
Subramanian[18]	Abstract	United States	MD Anderson NLP Pipeline	Inpatient	RR & MN	Note	-	-	Internal (pre-selected)	129 VTE, 1170 Control

Annotation tools extend beyond these VTE-specific examples. The U.S. National Library of Medicine provides the NLM Visual Tagging Tool and MetaMap, two downloadable applications for manual annotation of biomedical text.[20,21] They have been used extensively in the development of public biomedical NLP datasets. Commercial annotation platforms may also be used for text annotation. Common options include Label Studio, which includes a free self-hosted option, and John Snow Labs. Limited performance data is available for these proprietary pipelines and comparison would be difficult given the broad use cases of these models compared to models developed specifically for CAT or VTE detection. The GUIs for these annotation platforms are depicted in Figure 3.

Model performance

The included studies reported a range of variably reported metrics – sensitivity (equivalent to recall), specificity, positive predictive value (PPV, equivalent to precision), negative predictive value (NPV), area under the receiver operating curve (AUC, equivalent in the binary case to the c-statistic). Sensitivity ranged from 0.73 to 0.92 in validation datasets, with three models achieving a value of over 0.90.[13,15,16] The highest PPV of 0.95 was achieved by Jafari *et al*.[13] on an internal validation set, while other reporting studies achieved values from 0.80 to 0.89. Specificity and NPV exceeded 0.95 in all studies that reported these values.[13,16-18] Individual study results are detailed in Table 2. In studies reporting sensitivity and specificity, models were almost universally equally or more specific than sensitive and achieved equal or higher NPV than PPV in all cases. Furthermore, we examined whether the model was tested in an artificially balanced environment (50% positive and 50% negative), or with a test set more accurately reflecting the low prevalence of CAT (5% positive and 95% negative) in a typical clinical environment. Given the low prevalence of CAT as the expected outcome, NPV and specificity would be high by default. Therefore, PPV and sensitivity are key metrics of our evaluation.

Discussion

A practical example

To frame the discussion, it is helpful to describe a real-world scenario. Suppose investigators aim to estimate CAT incidence

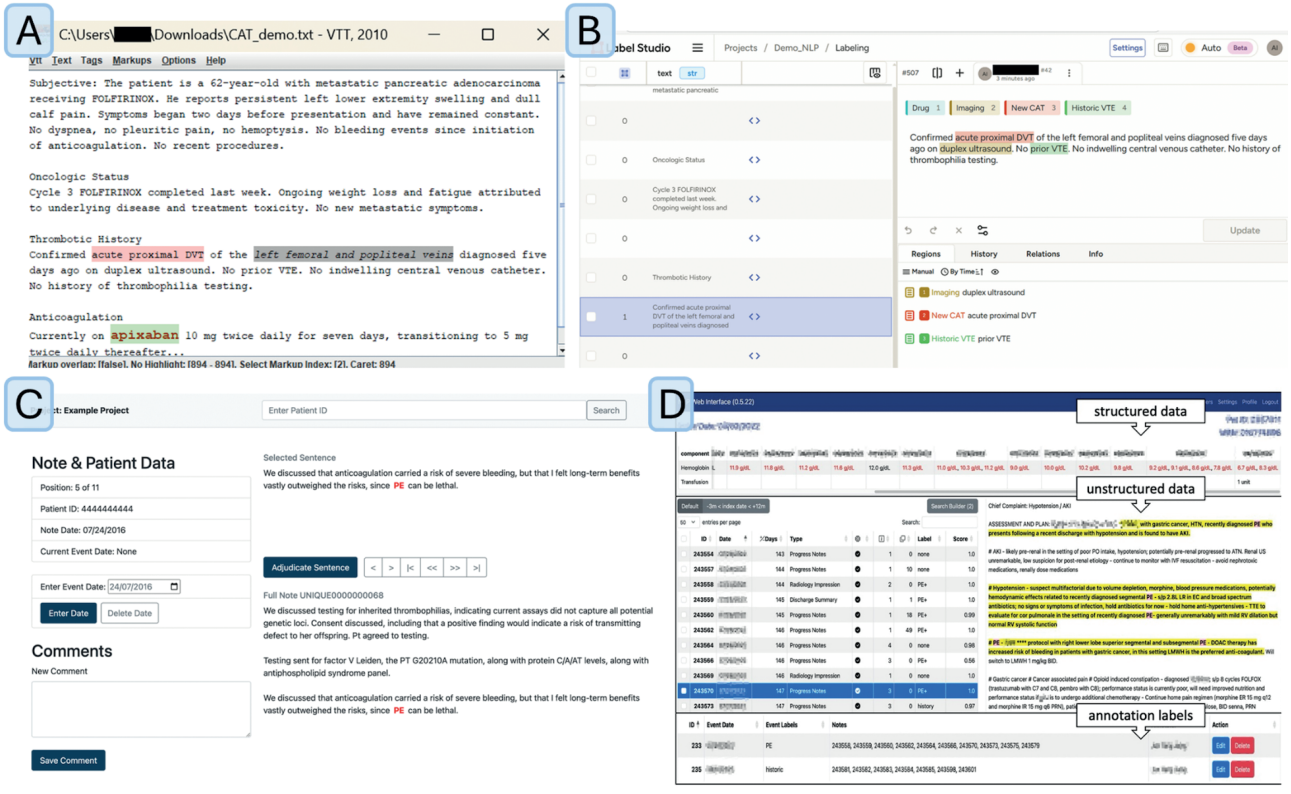

Figure 3. Graphic user interfaces for selected natural language processing annotation platforms. A) National Library of Medicine Visual Tagging Tool (NLM-VTT), a downloadable, desktop-based option for entity and part of speech tagging.[20] B) Label Studio (© 2025 HumanSignal, Inc.) is a commercial platform allowing for coordinated cloud-based annotation across large teams. The platform also includes options to upload models for assisted classification or fine tuning. C) CEDARS is an open source platform developed at Memorial Sloan Kettering specifically for annotation of VTE events with a focus on temporal relation between events (https://cedars.io).[19] D) NLP-MED is another platform geared toward VTE annotation, allowing for use of assisted annotation using pretrained or fine-tuned models as well as incorporation of structured data and temporal relations into the annotation process (https://nlpmed.demo.angli-lab.com/).

Table 2. Performance of included studies. Characteristics of the corresponding validation set are provided for context when interpreting the results. #External validation of CLAMP model developed by Li *et al.*[16] *Model Repository: https://huggingface.co/ang-li-lab/VTE-BERT. PPV, positive predictive value; NPV, negative predictive value; AUC, area under the receiver operating curve; NLP, natural language processing; ICD, international classification of diseases code; DVT, deep vein thrombosis; PE, pulmonary embolism.

Author	Event level	Validation set comparator		Standard	Sensitivity (recall)	Specificity	Precision (PPV)	NPV	AUC (c-statistic)	Model availability	Format
		VTE	Control								
Chen[12]	Note	502	201	Manual Review	-	-	-	-	0.74	Not available	Article
Jafari[13]	Patient	Internal: 69 External: 49	319 326	Manual Review	Training: 0.98 Internal: 0.91 External: 0.92	-	Training: 0.95 Internal: 0.95 External: 0.85	Training: 0.97 Internal: 0.9 8 External: 0.99	-	Open source*	Article
He[14]	Note	All patients treated in 2019 (N not specified)		Manual Review	-	-	-	-	DVT: 0.76 PE: 0.99	Not available	Abstract
Jagasia[15]	Patient	14	148	Manual Review	0.93	-	-	-	-	Not available	Abstract
Li16	Patient	894	7626	Manual Review	0.90	0.97	0.80	0.99	0.93	Closed source	Article
Avery[17,#]	Patient	Incident: 275 Prevalent: 389	76	Manual Review	0.73	-	0.82	-	-	Closed source	Abstract
Subramanian[18]	Note	129	1170	Manual Review	0.85	0.99	0.89	0.98	-	Not available	Abstract

in patients receiving two different immunotherapies for lung cancer from clinical notes. What is desirable for them to know in this situation? First and most obvious is the presence or absence of a CAT event. The model must be able to, on a note-by note basis, identify if a CAT has occurred.

Now consider a patient admitted with lung cancer who does not have a CAT during their admission, i.e., they should be classified as 'negative' for CAT by the model. During their hospitalization, they will likely accumulate many notes and radiology reports in their chart. In this case, even if a highly specific model can correctly identify 99 of 100 documents as negative, a single false positive will cause the whole admission to be marked as positive for CAT on a patient level. Despite high specificity on a per-note basis, the model returns an incorrect result for the investigators. Thus, for the end user, it is most relevant to know the performance on a per-patient rather than per-note basis.

A second scenario to examine is a patient with a remote history of DVT now admitted with lung cancer. Her record may contain many references to VTE, but none of these represent CAT or relate to an event in the current admission. Avoiding false positives in this case is challenging, and identifying newly incident CAT even more difficult. Herein lies the importance of distinguishing incident from prevalent events. Capturing incidence has the additional benefit of allowing temporal analyses, such as assessing associations between immunotherapy initiation and timing of subsequent CAT event.

A final consideration is portability. Users often want to deploy CAT-detection tools on their institution's data, which may contain distinctive language patterns, dot-phrases, and reporting conventions. A model that performs well only on data similar to its training corpus may adapt poorly to the new setting, and performance is likely to decline substantially. To ensure a model is able to transfer from institution to institution without performance loss, generalizability must be assessed using external validation sets.

In summary, the ideal model should perform well on the patient level, distinguish prevalent from incident CAT events despite low event frequency, and ideally, provide temporal information about event onset that can be used in conjunction with structured date data to effectively answer clinical questions. In practical terms, one might liken the performance of an ideal model to that of a capable senior medical student, but with far greater speed, scalability, and availability. Currently, the VTE-BERT model meets most of the criteria, though it is not yet capable to detect future recurrent events.[13]

The current state of NLP for CAT

Significant progress has been made over the past decade. From simple rules or lexicon-based models to ML pipelines with nuanced natural entity recognition that account for word position within sentences, and now to encoder-based approaches such as BERT and decoder-based LLMs, the field has undeniably advanced. A direct comparison between these models, however, is made difficult due to the variation between reported metrics and datasets.

Regardless, several transformer-based approaches are achieving excellent results in internal validation, approaching or exceeding NPV and PPV of 0.90 on a per-patient level.[13,22]

For certain use cases, such as building a cohort for retrospective review, this performance may be sufficient (provided the incorrect classifications are not biased in one way or another). For many other cases, however, human review is still a necessity.

There were few studies with external validation. The NLP pipeline based on the CLAMP platform (Licensed by Melax Technologies Inc.), initially customized by Li *et al*., was used again to classify notes at a separate institution by Avery *et al*. 2022.[16,17] Here, we see a notable drop in model sensitivity (0.90 to 0.73) but similar PPV (0.80 to 0.82). Contrast this with the transformer-based model where external validation revealed a stable sensitivity (0.91 to 0.92) but inferior PPV (0.95 to 0.85).[13] With only two models to compare, it is difficult to draw firm conclusions about differences in generalizability. Nonetheless, the current evidence suggests that transformer-based architectures will prove to be more generalizable as context windows lengthen and the size of datasets available for model training increases.[23]

An important but unresolved challenge: recurrent CAT detection

Identifying new CAT events is essential, as a new event often triggers further workup and an escalation of management for disease progression. No included studies fully addressed this problem, and the single study that differentiated incident from prevalent CAT did so at the binary level without attempting to resolve multiple CAT events across a patient's record. The problem is challenging, as it requires models to both disambiguate references to recent and distant historical events and pinpoint the exact timing of new events. Several real-world aspects increase the complexity of this task. One of the most common challenges arises from physicians using the "copy forward" function: phrases such as "patient with PE", or "DVT noted on ultrasound" may persist when no new event has occurred. Moreover, multiple events may appear within the same note, and terms like "acute" and "new" may not be modified in copy forwarded notes.

To tackle those situations and meet the requirements of an accurate model for decision making support tool, a recurrence detection model requires capabilities beyond a simple note-level classification. The model must digest temporal data, reconcile evidence across note over time, perform cross-document event linking, and recognize errors within the data. Traditional rule-based methods and even classical deep neural network models struggle to solve this task without explicit temporal supervised signals. Transformer based approaches, especially those with long context windows, have a level of understanding of clinical language and reasoning capabilities. As a result, transformers are better positioned to encode the longitudinal data. Yet no public models or dataset currently provided for such model to learn recurrence reliably.

Given the clinical importance of finding new thrombotic events particularly for treatment decision, follow up and risk modeling, the inability of current NLP models to capture recurrence events shows a major limitation. Developing a recurrence aware CAT system will require longitudinally annotated dataset, clear and accurate definition for recurrence versus new events, and evaluation frameworks that operate at the patient-level rather than the note level alone.

Moving forward

To reach this ideal model, a variety of considerations are required - first is training. Balanced data is needed to produce an accurate model, however positive events (CAT) are relatively rare. Constructing large enough datasets is therefore a difficult task. Conglomerating data across institutions is both promising and necessary, but it brings its own problems in terms of patient privacy in data transfer and model weights. Federated learning and synthetic training data are some viable strategies with demonstrated efficacy in similar machine learning tasks that may help mitigate this risk.[24] In the case of LLMs which require a very large training corpus, fine tuning existing general-purpose models is a versatile and proven solution to improve performance on domain-specific tasks.[25] However, data privacy restrictions and computational costs often limit the options for deployment of such a large model. Therefore, small locally hosted models, such as those generated via distillation, may be necessary.[26]

The second is model type. In this review, we observed increased performance in transformer based *vs* lexicon or rules only, similar to models in broader VTE detection.[10] As available compute power increases and becomes more accessible, transformer-based models will only become more feasible. That is not to say rules-based techniques are useless - all the models presented in this review rely at least in some capacity on an initial filtering or preprocessing step to reduce the size and dimensionality of the input data. For example, a predefined rule set may restrict input for a transformer-based model to only the HPI and assessment portion of notes written by an MD, with at least 1 keyword for VTE.

Finally, as we develop more powerful and versatile models, it is important that we have a standardized way to compare them. As seen in this manuscript, the measures reported by authors vary. Of course, as we included abstracts (as the published, full-length literature on NLP powered CAT identification is sparse), studies may not have been able to report the full array of metrics they computed. Nonetheless, testing done on a real-world dataset using random sampling and reporting at the least sensitivity, specificity, and PPV would facilitate a better comparison between models. External validation of each model on an open-source and standardized dataset with existing annotations (e.g., MIMIC IV) would further make inter-model comparison more meaningful.[27] Finally, open-source distribution of NLP models on code repositories such as GitHub and HuggingFace to facilitate external use and validation is highly encouraged.

References

1. Li A, Zhou E. Trends and updates on the epidemiology of cancer associated thrombosis: a systematic review. Bleeding Thromb Vasc Biol 2024;3:108.
2. Elvas LB, Almeida A, Ferreira JC. Natural language processing in medical text processing: A scoping literature review. Int J Med Inf 2025;204:106049.
3. Kim Y. Convolutional neural networks for sentence classification. arXiv:1408.5882v2.
4. Meystre SM, Savova GK, Kipper-Schuler KC, Hurdle JF. Extracting information from textual documents in the elec-

tronic health record: a review of recent research. Yearb Med Inform 2008;17:128-44.
5. Goldberg Y. Modeling with recurrent networks. In: Goldberg Y, ed. Neural network methods for natural language processing. Cham, Springer; 2017.
6. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. 2023. arXiv:1706.03762v7
7. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. 2019. arXiv:1810.04805.
8. Brown TB, Mann B, Ryder N, et al. Language models are few-shot learners. 2020. arXiv:2005.14165.
9. Liu P, Yuan W, Fu J, et al. Pre-train, prompt, and predict: a systematic survey of prompting methods in natural language processing. ACM Comput Surv 2023;55:195.
10. Lam BD, Chrysafi P, Chiasakul T, et al. Machine learning natural language processing for identifying venous thromboembolism: systematic review and meta-analysis. Blood Adv 2024;8:2991-3000.
11. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021;372:n71.
12. Chen Y, Carroll RJ, Hinz ERM, et al. Applying active learning to high-throughput phenotyping algorithms for electronic health records data. J Am Med Inform Assoc 2013; 20:e253-9.
13. Jafari O, Ma S, Lam BD, et al. Development and validation of venous thromboembolism–bidirectional encoder representations from transformers (VTE-BERT) natural language processing model. J Thromb Haemost 2025. Online ahead of print.
14. He JC, Hirsch I, Li Y, et al. Impact of applying machine learning to the electronic medical record on prediction of cancer-associated thrombosis. JCO Oncol Pract 2024;20: 409.
15. Jagasia S, Krauze AV. Developing a word lexicon from electronic health records for natural language processing analysis of free-text reports for patients with venous thromboembolism. Int J Radiat Oncol 2023;117:e469.
16. Li A, da Costa WL, Guffey D, et al. Developing and optimizing a computable phenotype for incident venous thromboembolism in a longitudinal cohort of patients with cancer. Res Pract Thromb Haemost 2022;6:e12733.
17. Avery J, Martens KL, Nguyen D, et al. Utilization of natural language processing in venous thromboembolism identification. Blood 2022;140:7860-1.
18. Subramanian NG, Pleitez HG, Nguyen D, et al. Diagnostic performance of natural language processing in detection of acute cancer VTE. J Clin Oncol 2023;41:e19062.
19. Singh R, Mantha S. PINES (progressive inference networked episodic service). Available from: https://pines.ai
20. Williamson J. Development of visual tagging tool.
21. Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proc AMIA Symp 2001:17-21.
22. Maghsoudi A, Zhou E, Guffey D, et al. A Transformer natural language processing algorithm for cancer associated thrombosis phenotype. Blood 2023;142:S1267.
23. Yuan K, Yoon CH, Gu Q, et al. Transformers and large language models are efficient feature extractors for electronic health record studies. Commun Med 2025;5:83.
24. Peng L, Luo G, Zhou S, et al. An in-depth evaluation of federated learning on biomedical natural language processing for information extraction. Npj Digit Med 2024;7:127.
25. Guluzade A, Heiba N, Boukhers Z, et al. ELMTEX: fine-tuning LLMs for structured clinical information extraction. A case study on clinical reports. In: Bellazzi R, Juarez Herrero JM, Sacchi L, Zupan B, eds. Artificial intelligence in medicine. AIME 2025. Cham, Springer.
26. Hsieh CY, Li CL, Yeh CK, et al. Distilling step-by-step! Outperforming larger language models with less training data and smaller model sizes. 2023. arXiv:2305.02301v2.
27. Lam BD, Ma S, Kovalenko I, et al. Using a transformer language model to curate a pulmonary embolism dataset from the Medical Information Mart for Intensive Care IV: MIMIC-IV-Ext-PE. Res Pract Thromb Haemost 2025;9: 102896.