





ORIGINAL ARTICLE

Developing and optimizing a computable phenotype for incident venous thromboembolism in a longitudinal cohort of patients with cancer

Ang Li MD, MS¹   | Wilson L. da Costa Jr MD, MPH, PhD²  | Danielle Guffey MS³ | Emily M. Milner BS⁴ | Anthony K. Allam BS⁴ | Karen M. Kurian BS⁴ | Francisco J. Novoa MD⁴ | Marguerite D. Poche BS⁴ | Raka Bandyo MS^{5,6} | Carolina Granada MD¹ | Courtney D. Wallace BS⁶ | Neil A. Zakai MD, MS⁷  | Christopher I. Amos PhD^{2,3}

¹Section of Hematology-Oncology, Baylor College of Medicine, Houston, Texas, USA

²Section of Epidemiology and Population Science, Baylor College of Medicine, Houston, Texas, USA

³Institute for Clinical and Translational Research, Baylor College of Medicine, Houston, Texas, USA

⁴School of Medicine, Baylor College of Medicine, Houston, Texas, USA

⁵Bluetree Network Inc, Madison, Wisconsin, USA

⁶Harris Health System, Houston, Texas, USA

⁷Departments of Medicine and Pathology and Laboratory Medicine, Larner College of Medicine at the University of Vermont, Burlington, Vermont, USA

Correspondence

Ang Li, Section of Hematology-Oncology, Baylor College of Medicine, One Baylor Plaza, MS:307, Room 610D, Houston, TX 77030, USA.

Email: ang.li2@bcm.edu

Funding information

Cancer Prevention and Research Institute of Texas, Grant/Award Number: RP210037, RPI160097 and RR190104

Handling Editor: Dr Suzanne Cannegieter

Abstract

Background: Research on venous thromboembolism (VTE) that relies only on the International Classification of Diseases (ICD) can misclassify outcomes. Our study aims to discover and validate an improved VTE computable phenotype for people with cancer.

Methods: We used a cancer registry electronic health record (EHR)-linked longitudinal database. We derived three algorithms that were ICD/medication based, natural language processing (NLP) based, or all combined. We then randomly sampled 400 patients from patients with VTE codes ($n = 1111$) and 400 from those without VTE codes ($n = 7396$). Weighted sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) were calculated on the entire sample using inverse probability weighting, followed by bootstrapped receiver operating curve analysis to calculate the concordance statistic (c statistic).

Results: Among 800 patients sampled, 280 had a confirmed acute VTE during the first year after cancer diagnosis. The ICD/medication algorithm had a weighted PPV of 95% and a weighted sensitivity of 81%, with a c statistic of 0.90 (95% confidence interval [CI], 0.89–0.91). Adding Current Procedural Terminology codes for inferior vena cava filter removal or early death did not improve the performance. The NLP algorithm had a weighted PPV of 80% and a weighted sensitivity of 90%, with a c statistic of 0.93 (95% CI, 0.92–0.94). The combined algorithm had a weighted PPV of 98% at the higher cutoff and a weighted sensitivity of 96% at the lower cutoff, with a c statistic of 0.98 (95% CI, 0.97–0.98).

Conclusions: Our ICD/medication-based algorithm can accurately identify VTE phenotype among patients with cancer with a high PPV of 95%. The combined algorithm should be considered in EHR databases that have access to such capabilities.

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial-NoDerivs](https://creativecommons.org/licenses/by-nc-nd/4.0/) License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2022 The Authors. *Research and Practice in Thrombosis and Haemostasis* published by Wiley Periodicals LLC on behalf of International Society on Thrombosis and Haemostasis (ISTH).

KEYWORDS

administrative claims, health care, natural language processing, neoplasms, venous thromboembolism, venous thrombosis

Essentials

- Venous thromboembolism (VTE) is an important complication to study among patients with cancer.
- It is challenging to identify VTE in epidemiology studies using electronic health records.
- We validated a VTE phenotype algorithm using billing codes and natural language processing (NLP).
- The new algorithm combining billing codes and NLP radiology reports provided optimal prediction.

1 | INTRODUCTION

A computable phenotype is defined as a clinical condition that can be ascertained by means of a computerized query using a defined set of data elements and logical expressions without chart review or interpretation by a clinician.¹ In the context of pragmatic clinical trials and epidemiology studies, defining and validating important computable phenotypes is both critical and challenging.² Specifically within the realm of hematology/oncology research, many have tried to create a computable phenotype for venous thromboembolism (VTE) using administrative claims data with varying success.³ Even fewer algorithms exist to identify acute VTE events among longitudinal cohorts with distinct index dates and repeated radiology scans such as those with incident cancer diagnosis.⁴

While most of the existing work in VTE phenotype relies on using diagnostic codes within administrative claims or billing data, the full reliance on International Classification of Diseases (ICD) codes in such data can lead to significant bias. Fortunately, advances in clinical informatics and integrated electronic data warehouses have provided us with an expanded armamentarium. VTE is often recorded in patients' electronic health records (EHRs) both as structured data (billing codes, radiology codes) and unstructured data (radiology reports, clinical notes). The vastness and complexity of these data provide a challenge for researchers in obtaining consistent results, particularly when attempting to correctly identify and describe specific VTE events. The gold standard for this process is manual chart review, though this method presents a significant limitation when thousands of patient charts must be reviewed. The development of an algorithm that is capable of rapidly and correctly identifying VTE in a modern EHR data environment (as opposed to administrative claims data) would be an important tool for patient identification and outcome ascertainment in pragmatic clinical trials.

In our current study, we present various algorithms that leverage both structured and unstructured data elements in an EHR database linked to our local cancer registry. We demonstrate the use of ICD codes, medications, Current Procedural Terminology (CPT) codes, and natural language processing (NLP) algorithms to identify VTE events with high positive predictive value (PPV) and sensitivity in a longitudinal cohort of patients with cancer. We also present

an internal validation study with random sampling to estimate the overall performance of the algorithm in comparison with other previously published studies.

2 | METHODS

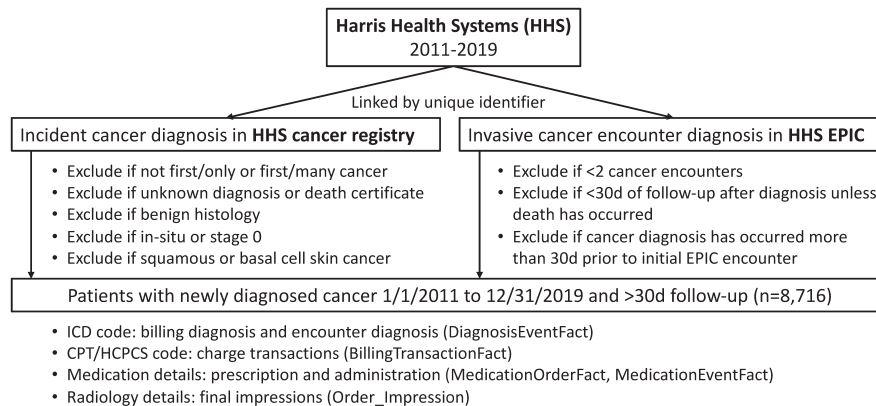
2.1 | Study design and population

We performed a retrospective cohort study at Baylor College of Medicine (BCM) and Harris Health System (HHS), the largest safety-net health care system that provides care for the underserved and uninsured patients in the Houston metropolitan area. To construct the cohort, we identified and linked ambulatory and hospitalized patients through unique identifiers from both the institutional Cancer Registry and *Epic* Clarity and Caboodle data warehouse. As shown in [Figure 1](#), a patient was considered eligible if he/she had an incident cancer diagnosis with confirmed histology and invasive staging. Furthermore, the patient must have two or more cancer-related encounters and >30 days of continuous follow-up. Due to the lack of medical insurance in the majority of patients in this cohort, there was very little loss to follow-up or clinical visits outside of the HHS facilities. The study was approved by the BCM Institutional Review Board.

2.2 | Definition of VTE phenotype gold standard

The gold-standard VTE phenotype was defined as radiologically confirmed, symptomatic or incidental diagnosis of acute subsegmental or larger pulmonary embolism (PE), proximal or distal lower-extremity deep vein thrombosis (LE-DVT), proximal or distal upper-extremity DVT (UE-DVT), or bland splanchnic vein thrombosis (excluding tumor thrombus) within 1 year of cancer diagnosis. There was no uniform institutional guideline for outpatient thromboprophylaxis, although most patients received inpatient VTE prophylaxis. Two trained reviewers (WLC, CG) performed blinded chart abstraction of all radiology reports, clinical notes, and discharge summaries using keyword searches [(thromb) OR (embol) OR (filling defect) OR "DVT" OR "PE" OR "VTE"] in *Epic* within 1 year after cancer diagnosis. Each record was independently

FIGURE 1 Cohort construction and data sources. CPT, Current Procedural Terminology; HCPCS, Healthcare Common Procedure Coding System; ICD, International Classification of Diseases



assessed by a second reviewer (EMM, AKA, KMK, FJN, MDP), and discrepancies were resolved by a third clinician reviewer with expertise in cancer-associated thrombosis (AL). Data were collected and stored using the REDCap electronic data capture tools hosted at BCM.^{5,6}

2.3 | Data source for the study

After extracting ICD diagnosis codes, we applied sequential filters to ensure the codes would match only relevant and meaningful clinical encounters (Figure S1). All available ICD, Ninth Revision, Clinical Modification (ICD-9-CM) and ICD, Tenth Revision, Clinical Modification (ICD-10-CM) codes (as of 2021) for acute PE, acute LE-DVT, acute UE-DVT, nonspecific VTE, chronic VTE, and history of PE/DVT were selected to be filtered in later steps (Table S1).

CPT codes and Healthcare Common Procedure Coding System were extracted from chargemaster tables. Both inpatient and outpatient procedural codes were kept regardless of final billing or reimbursement status. Radiology procedures (including contrast-enhanced computed tomography or magnetic resonance imaging, ventilation/perfusion scans, Doppler ultrasound) for PE and DVT, inferior vena cava (IVC) filter placement, and thrombolysis/thrombectomy were defined using CPT codes (Table S2). Final impression reports of radiology studies were extracted, and those associated with relevant VTE-related CPT codes were kept for further analysis.

Medication details were extracted and extensively cleaned to remove duplicates, ordered but held, misspelled names, incorrect dosage, and frequency or route. Therapeutic anticoagulation at the time of VTE diagnosis was defined as the presence of an administered (inpatient/infusion center) or prescribed (outpatient/discharge) direct oral anticoagulant (including rivaroxaban, apixaban, edoxaban, and dabigatran) (any dose), oral vitamin K antagonist (any dose), subcutaneous enoxaparin (>1.3 mg/kg if daily or >0.8 mg/kg for twice-daily frequency), subcutaneous fondaparinux (5–10 mg), intravenous heparin (continuous drip only, excluding flushes or pushes), or subcutaneous heparin (>7500 mg) within 7 days before and 30 days after suspected VTE diagnosis.

2.4 | ICD and NLP algorithm derivation

The stepwise approach to algorithm discovery is shown in Figure 2. For the ICD algorithm, we reviewed 1000 patient charts with a positive ICD code for VTE after cancer diagnosis. We first examined the PPV for each of the included ICD-9-CM and ICD-10-CM codes especially for those with nonspecific descriptions such as “embolism and thrombosis of unspecified vein.” We then examined how the “carryover” effect from historical VTE events impacted miscoding. Third, we assessed various means to mitigate the “rule-out” effect such as using two or more outpatient visits >30 days apart and one outpatient visit with therapeutic anticoagulant within 7 days before and 30 days after VTE, similar to a previously study.⁷ Finally, we imposed other published rules such as IVC filter placement or death within 30 days according to another previous study.⁴ We reported acute VTE concordance as a binary outcome in lifetime after index date or within 90 days of the first documented ICD code.

For the NLP algorithm, we selectively reviewed 1300 radiology reports (350 positive and 950 negative scans for VTE) in the discovery set. We built a rule-based NLP pipeline using the CLAMP software (Melax Tech, Houston, TX, USA) that included built-in entity recognizer (VTE), assertion classifier, part-of-speech tagger, sentence detector, tokenizer, and Ruta rule engine.⁸ We then modified the dictionaries for VTE, negation, and preexisting tokens, along with appropriate semantic rules to optimize the performance of the pipeline. Specifically, a radiology report was predictive for VTE if it had a nonnegated and nonpreexisting VTE entity (either as a stand-alone term or a deep venous site plus disorder) in the same sentence (Table S3).

2.5 | Statistical analysis for algorithm validation

We performed an internal validation study using random sampling of the same overall population to account for the false-positive and false-negative cases, similar to previously published studies (Figure 3).⁹ Patients were split into two groups based on whether they had at least one acute VTE code after diagnosis (high vs low pretest probability for VTE). A total of 400 patients were then

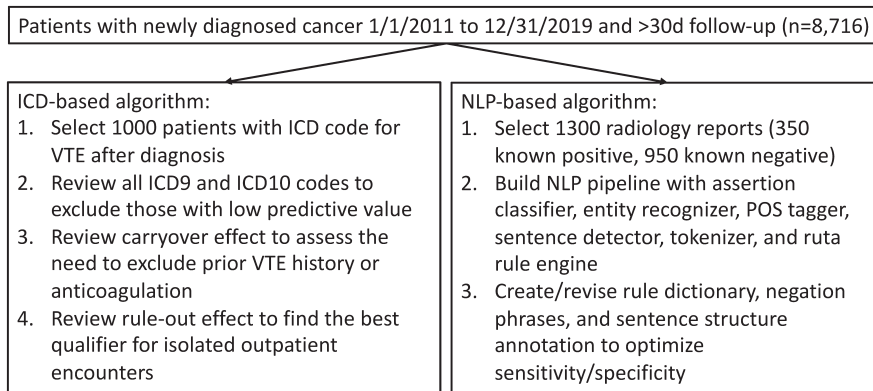


FIGURE 2 Study design for computable phenotype derivation. ICD, International Classification of Diseases; NLP, natural language processing; POS, part-of-speech; VTE, venous thromboembolism

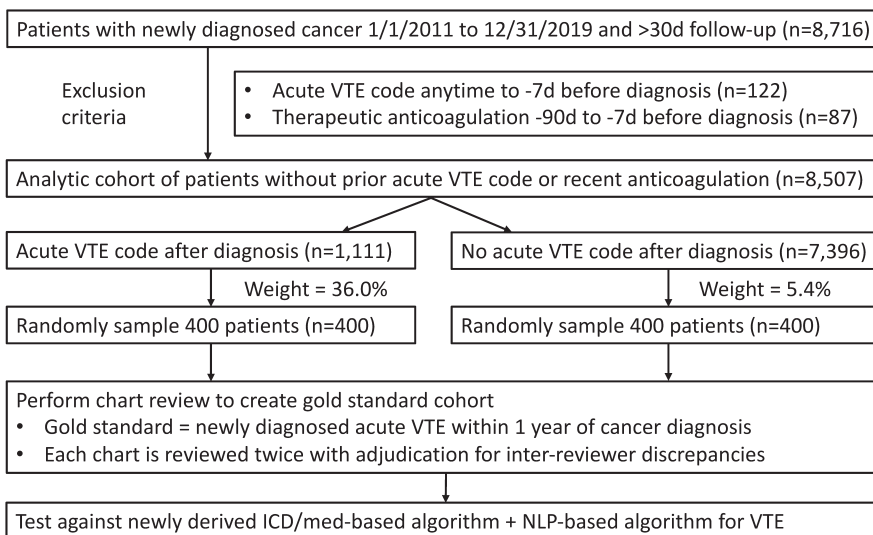


FIGURE 3 Study design for computable phenotype validation. ICD, International Classification of Diseases; NLP, natural language processing; VTE, venous thromboembolism

randomly sampled from each group for blinded chart review to determine VTE occurrence as described previously. Both ICD and NLP-based algorithms were applied and tested against the gold-standard chart review VTE outcomes within 1 year of cancer diagnosis. We specifically tested four different algorithms: previous ICD/CPT/medication algorithm from Sanfilippo et al,⁴ current ICD algorithm, current NLP algorithm, and a combined ICD/NLP algorithm.

Weighted sensitivity, specificity, PPV, and negative predictive value (NPV), were calculated using the inverse probability weighting (IPW) method, which is equivalent to the Begg and Greenes method to account for verification bias.¹⁰ The IPW used weights defined by the sampling rate within each stratum. Confidence intervals (CIs; 95%) were calculated via bootstrapping 1000 times and were presented as bias-corrected CIs adjusting for the sampling weights. Notably, the weights for each algorithm were different slightly due to differential sampling based on initial ICD-based selection and more restrictive algorithms after the initial sampling. Receiver operating characteristics analysis was implemented, adjusting for sample weighting to assess the area under the curve/concordance statistic (c statistic) for each algorithm. All analyses were performed using Stata 16.0 (StataCorp, College Station, TX, USA).

3 | RESULTS

3.1 | VTE computable phenotype algorithms derivation

A total of 8716 patients with newly diagnosed cancer over 9 years formed the population for the current study (Figure 1). Table 1 depicts the PPV for each category of the first ICD codes after the application of various exclusion filters among 1000 patients. The distribution of first documented ICD codes after cancer diagnosis were 32% (n = 324), 23% (n = 228), 10% (n = 98), 13% (n = 134), 17% (n = 173), 4% (n = 43) for acute PE, acute LE-DVT, acute UE-DVT, nonspecific VTE, historic VTE, and chronic VTE, respectively.

A detailed stepwise approach for optimizing the ICD algorithm is shown in Figure 2 and Table 1. Before any exclusion filters, 662 validated acute new VTE events were confirmed in 1000 patients with any VTE codes (PPV 66%) and 595 of them were within 90 days of the first ICD code (PPV 60%). Among them, chronic VTE codes had 60% PPV for acute VTE within 90 days but with very few events. Nonspecific and historic VTE codes had relatively low PPV for predicting acute VTE outcomes (PPV 31%-35%). Among the nonspecific codes, 451.2, 451.84, 451.89, 451.9, 453.1, 453.9, I80.8, I80.9, and I82.1 had low PPV (<50%; see Table S4 for details). After excluding these low-PPV codes

TABLE 1 Performance of the ICD-based algorithms after each exclusion filter in derivation data set

| ICD Category ^a | No exclusion ^b | | | Exclusion 1 ^c | | | Exclusion 2 ^d | | |
|---------------------------|---------------------------|----------------|----------------------|--------------------------|----------------|----------------------|--------------------------|----------|---------------|
| | Total | PPV VTE, n (%) | PPV \pm 90d, n (%) | Total | PPV VTE, n (%) | PPV \pm 90d, n (%) | Total | PPV VTE | PPV \pm 90d |
| Acute PE | 324 | 270 (83) | 252 (78) | 374 | 307 (82) | 279 (75) | 289 | 258 (89) | 241 (83) |
| Acute LE-DVT | 228 | 146 (64) | 139 (61) | 284 | 180 (63) | 167 (59) | 202 | 142 (70) | 134 (66) |
| Acute UE-DVT | 98 | 81 (83) | 78 (80) | 119 | 96 (81) | 88 (74) | 100 | 85 (85) | 78 (78) |
| Non-specific VTE | 134 | 59 (44) | 47 (35) | 78 | 53 (68) | 46 (59) | 62 | 42 (68) | 36 (58) |
| Historic VTE | 173 | 75 (43) | 53 (31) | | | | | | |
| Chronic VTE | 43 | 31 (72) | 26 (60) | | | | | | |
| Total | 1000 | 662 (66) | 595 (60) | 855 | 636 (74) | 580 (68) | 653 | 527 (81) | 489 (75) |

| ICD Category ^a | Exclusion #3 ^e | | | Exclusion #4 ^f | | |
|---------------------------|---------------------------|----------|---------------|---------------------------|----------|---------------|
| | Total | PPV VTE | PPV \pm 90d | Total | PPV VTE | PPV \pm 90d |
| Acute PE | 255 | 237 (93) | 223 (87) | 235 | 227 (97) | 217 (92) |
| Acute LE-DVT | 158 | 137 (87) | 132 (84) | 133 | 116 (87) | 113 (85) |
| Acute UE-DVT | 94 | 86 (91) | 79 (84) | 68 | 64 (94) | 63 (93) |
| Non-specific VTE | 40 | 36 (90) | 33 (83) | 27 | 22 (81) | 21 (78) |
| Historic VTE | | | | | | |
| Chronic VTE | | | | | | |
| Total | 547 | 496 (91) | 467 (85) | 463 | 429 (93) | 414 (89) |

Abbreviations: ICD, International Classification of Diseases; IVC, inferior vena cava; LE-DVT, lower-extremity deep vein thrombosis; PE, pulmonary embolism; PPV, positive predictive value; UE-DVT, upper-extremity deep vein thrombosis; VTE, venous thromboembolism.

^aSee Table S1 for detailed list of code included.

^bNo exclusion: use first ICD code after date of cancer diagnosis. Column "PPV VTE" indicates how many patients had acute VTE after cancer diagnosis regardless of timing; column "PPE \pm 90d" indicates how many patients had acute VTE within 90 days of first given ICD code.

^cExclusion 1 (wrong ICD code): use first ICD code after excluding chronic VTE codes, history VTE codes, and a subset of nonspecific VTE codes from consideration of "acute VTE" (see Table S4).

^dExclusion 2 (carryover effect): use first ICD code after excluding patients with known ICD codes for VTE (any time) or received therapeutic anticoagulation (up to 90 d) before date of cancer diagnosis + exclusion 1.

^eExclusion 3 (rule out effect): use first ICD code after excluding patients with outpatient encounter UNLESS having 2+ codes >30 d and <180 d apart or receiving therapeutic anticoagulation at time of encounter (-7 d to +30 d) + exclusion 1 + exclusion 2.

^fExclusion 4 (anticoagulation, IVC filter and death): use first ICD code that had anticoagulation, IVC filter placement, or death within 30 d regardless of inpatient or outpatient encounter +exclusion 1 + exclusion 2.

(exclusion filter 1), the remaining first ICD codes had 68% PPV for acute VTE within 90 days. As many of the false-positive events were related to "carry-over" effect from the historic VTE events mistakenly coded as acute VTE, we further excluded patients with acute VTE ICD codes anytime before cancer diagnosis or those receiving therapeutic anticoagulation within 1 month before cancer diagnosis (exclusion filter 2). This change improved the overall PPV of acute VTE within 90 days to 75%. Finally, we applied additional exclusion criteria to consider an ICD code from an outpatient encounter only if it had two or more codes >30 days and <180 days apart (the "rule-out" criterion) unless the patient was prescribed or received therapeutic anticoagulation within 30 days after the VTE ICD code (exclusion filter 3). With an improved PPV of acute VTE within 90 days to 85%, we chose this to be our final algorithm for validation testing. Notably, adding IVC filter or death in addition to anticoagulation within 30 days marginally improved the PPV (89%) but captured significantly fewer VTE events (exclusion filter 4).

Similar to the ICD algorithm creation, we optimized the NLP algorithm performance within the discovery set of 1300 radiology impression reports. Specifically, we locked the pipeline once the sensitivity was 97% (340/350 true positives) and PPV was 98% (340/346 predicted positives) for predicting acute VTE event on the radiology report (Table S5). Notably, this did not account for serial scans or missing scans, and the validation performance was expected to be less optimal.

3.2 | VTE computable phenotype algorithm validation via random sampling

After excluding those with acute VTE codes (n = 122) or receiving therapeutic anticoagulation (n = 87) before cancer diagnosis, 400 patients were randomly sampled from 1111 with acute VTE codes (sampling weight 36.0%), and 400 were sampled from 7396 without VTE ICD codes (sampling weight 5.4%) (Figure 3).

Table 2 presents the observed numbers for VTE by algorithm for the unweighted and weighted VTE populations, in addition to the weighted test characteristics. Among 800 randomly selected patients, 280 had a confirmed acute VTE that occurred during the first year after cancer diagnosis. The current ICD algorithm with two or more outpatient or anticoagulation criteria had a PPV of 94.5% and a sensitivity of 80.8% (c statistic, 0.90; 95% CI, 0.89-0.91). In comparison, the Sanfilippo algorithm with additional IVC filter CPT codes or death within 30 days (applied to all encounters, not only outpatient encounters) had the highest PPV (96.7%) but the lowest sensitivity (71.7%) (c statistic, 0.86; 95% CI, 0.84-0.87). On the contrary, the current NLP algorithm had the best sensitivity of 89.5% but the lowest PPV of 79.5% (c statistic, 0.93; 95% CI, 0.92-0.94).

The ICD/NLP combined algorithm using two-point cutoffs offered the best prediction, with a c statistic of 0.98 (95% CI, 0.97-0.98). Approximately 87% of the population was not predicted to

have VTE by both ICD and NLP (concordant negative), and the sensitivity was 96.1% and NPV was 99.5%. Approximately 8% of the population was predicted to have VTE by both ICD and NLP, and the PPV was 97.5% and the specificity was 99.8%.

4 | DISCUSSION

In the current study, we optimized and internally validated an acute VTE computable phenotype after a predetermined index date using a combination of structured (ICD/medication) and unstructured data (NLP/radiology) from a large *Epic* EHR database linked to the local cancer registry. We found that while ICD codes could identify VTE, full reliance without appropriate exclusion filters was associated with an unacceptably low PPV. In contrast, a systematically derived algorithm that combined both ICD/medication and NLP/radiology

TABLE 2 Performance of ICD and NLP-based algorithms in validation data set

| | Unweighted VTE (n = 800) | | Weighted VTE (n = 8507) ^e | | Weighted Sensitivity, % | Weighted Specificity, % | Weighted PPV, % | Weighted NPV, % | Weighted c statistic | |
|---|--------------------------|-----|--------------------------------------|-----|-------------------------|-------------------------|------------------|------------------|----------------------|--|
| | No | Yes | No | Yes | | | | | | |
| Previous ICD algorithm^a | | | | | | | | | | |
| Predicted no (0) | 512 | 51 | 7604 | 253 | 71.7 (68.9-74.6) | 99.7 (99.6-99.8) | 96.7 (95.2-97.7) | 96.8 (96.4-97.2) | 0.86 | |
| Predicted yes (1) | 8 | 229 | 22 | 641 | | | | | (0.84-0.87) | |
| Current ICD algorithm^b | | | | | | | | | | |
| Predicted no (0) | 505 | 22 | 7584 | 172 | 80.8 (78.1-83.1) | 99.4 (99.3-99.6) | 94.5 (92.9-96.2) | 97.8 (97.4-98.1) | 0.90 | |
| Predicted yes (1) | 15 | 258 | 42 | 722 | | | | | (0.89-0.91) | |
| Current NLP algorithm^c | | | | | | | | | | |
| Predicted no (0) | 497 | 28 | 7420 | 94 | 89.5 (87.1-91.2) | 97.3 (96.9-97.6) | 79.5 (77.1-81.9) | 98.7 (98.5-99.0) | 0.93 | |
| Predicted yes (1) | 23 | 252 | 206 | 800 | | | | | (0.92-0.94) | |
| ICD + NLP algorithm^d | | | | | | | | | | |
| Predicted no (0) | 488 | 7 | 7356 | 35 | 96.1 (94.7-97.2) | 97.0 (96.6-97.3) | 78.8 (76.2-81.2) | 99.5 (99.3-99.7) | 0.98 | |
| Predicted yes (1) | 26 | 36 | 218 | 203 | 74.3 (71.5-77.1) | 99.8 (99.6-99.9) | 97.5 (96.0-98.4) | 97.1 (96.7-97.5) | (0.97-0.98) | |
| Predicted yes (2) | 6 | 237 | 16 | 640 | | | | | | |

Note: Acute VTE ICD-9-CM and ICD-10-CM codes are listed in Table S1 (after excluding certain nonspecific, historic, and chronic codes). Therapeutic anticoagulation is defined as the presence of an administered (inpatient/infusion center) or prescribed (outpatient) direct oral anticoagulant (DOAC) (any dose), oral warfarin (any dose), subcutaneous enoxaparin (>1.3 mg/kg if daily or >0.8 mg/kg for twice-daily frequency), subcutaneous fondaparinux (5-10 mg), intravenous (IV) heparin (continuous drip only excluding flushes or pushes), or subcutaneous heparin (>7500 mg) within 7 days before and 30 days after suspected VTE diagnosis. IVC filter is defined by CPT codes 37191, 37620, 36005, and 36010. Relevant radiology reports are defined as any contrast scan or Doppler ultrasound with CPT codes listed in Table S2.

Abbreviations: ICD-9-CM, International Classification of Diseases, Ninth Revision, Clinical Modification; ICD-10-CM, International Classification of Diseases, Tenth Revision, Clinical Modification; NLP, natural language processing; NPV, negative predictive value; PPV, positive predictive value; VTE, venous thromboembolism.

^a Previous ICD algorithm (Sanfilippo et al,⁴ ICD-9-CM converted to ICD-10-CM): first of (any inpatient or outpatient acute VTE code) with (therapeutic anticoagulation or IVC filter or death -7 d to +30 d) within 365 d.

^b Current ICD algorithm: first of (any inpatient acute VTE code) or (2+ outpatient acute VTE code >30 d and <180 d) or (any outpatient acute VTE code with therapeutic anticoagulation -7 d to +30 d) within 365 d.

^c Current NLP algorithm: first relevant radiology impression predicted to be positive for VTE based on rule-based NLP prediction.

^d ICD + NLP algorithm: 1st of either ICD or NLP algorithm within 365d. Predicted no (0) indicates both algorithms did not identify VTE; predicted yes (1) indicates one of the two algorithms identified VTE; predicted yes (2) indicates both algorithms identified VTE.

^e Weighted events are slightly different due to decimal rounding.

Number ranges in parentheses refer to 95% Confidence Intervals (CI) for the weighted estimates.

provided the optimal PPV and sensitivity trade-off, although either the simplified ICD algorithm (higher PPV) or the NLP algorithm (higher sensitivity) alone was likely sufficient for most clinical studies. While our analysis was limited to the cancer population, the design concept applies to any closed population cohort with a clear index date. As *Epic* accounts for the EHR experience of nearly half of the hospitals in the United States, our computable phenotype algorithms are likely generalizable to many other large health care systems that have adequate capture of inpatient and outpatient encounters.¹¹

It is important to consider the relevant test characteristics of an algorithm for a phenotype with a low prevalence. While sensitivity, specificity, PPV, and NPV are often discussed as equally important metrics, sensitivity and PPV (or recall and precision in data science) often supply the most variable and important information on an algorithm's misclassification error. Assuming the error is nondifferential between different exposure risk groups, an algorithm with low PPV will lead to an attenuated risk ratio (important for etiologic association testing), while one with low sensitivity will lead to an attenuated absolute risk difference (important for incidence estimation and causal inference).¹² Therefore, an optimal phenotype prediction algorithm should strive for a combination of high PPV and sensitivity. Furthermore, we must use an appropriate sampling method when designing validation studies. Since it is impossible to know the true-positive and -negative VTE cases in a large data set, most studies (including ours) rely on random sampling from positive versus negative predicted strata. With this approach, it is easy to estimate the PPV. However, to estimate the true sensitivity in the overall population, we must overweigh every single false-negative event 20 times if the sampling fraction was 5%; otherwise, we would have incorrectly reported a sensitivity of 92% instead of 81% for our ICD-based algorithm. This approach was conducted in some previous studies but not others^{9,13} and also highlights the reason why most studies only reported PPV over sensitivity.³

The accuracy of using ICD codes to predict VTE events has always been a topic of debate. In a systematic review in 2012, Tamariz et al³ found that the PPV of relevant ICD-9-CM codes from administrative claims data varied between 65% and 95%, depending on the study population. Furthermore, most of the studies relied on validation of administrative codes at the time of discharge and did not account for the longitudinal nature of a patient's clinical course. More recent publications using EHR data confirmed that overreliance on administrative data could lead to significantly misclassified finding.¹⁴ In one example, the Cardiovascular Research Network Venous Thromboembolism study in 2017 reported PPV as low as 65% for inpatient and 31% for outpatients when assessed in a cohort setting.¹⁵ Sanfilippo et al⁴ published one of the best existing VTE algorithms using a longitudinal cancer cohort from the Veterans Affairs database. We tested this algorithm in our validation study and reached a similar conclusion (PPV 91% and sensitivity 72% in the original study, PPV 95% and sensitivity 72% in the current study). In contrast to this previous study, we found that the addition of IVC filter CPT codes or death within 30 days to all inpatient and

outpatient encounters did not drastically improve the performance of the algorithm. This could partially be explained by the decreased use of IVC filters over the past decade after the 2010 US Food and Drug Administration Advisory safety warning.¹⁶ Our revised and simplified ICD algorithm relied only on ICD coding selection (if one inpatient or two or more outpatient) and therapeutic anticoagulation (if one isolated outpatient) to achieve a similar PPV of 95% and an improved sensitivity of 81%. Notably, we used both prescribed (outpatient) and administered (inpatient) anticoagulant medications and carefully defined "therapeutic dose" based on the frequency of the drug administration.

We also explored the value of an independent NLP algorithm based on radiology reports. Comparing to the NLP protocol from the eMERGE Mayo Group phenotype published on PheKB,¹⁷ while we used similar concept of defining VTE using either "standalone" or "site" plus "disorder," the existing protocol was neither sensitive nor specific in our population of cancer patients. Our modified NLP protocol specifically used negation to mitigate the false positives related to "septic thrombi," "tumor thrombi," or "superficial thrombi." We also distinguished acute from potentially preexisting events such as "history of DVT" or "persistent clot." The performance characteristics of the NLP pipeline on individual radiology reports (discovery set) was expectedly better than its application in a longitudinal cohort setting (validation set) because patients often had multiple reports over time and might have missing reports from outside their hospital. Upon further review of the events captured by NLP but missed by VTE, many of them were nontumor splanchnic vein thromboses incidentally detected in gastrointestinal malignancies and not anticoagulated. The performance of our NLP algorithm (sensitivity 90% and PPV 80%) is similar in performance to other NLP based VTE algorithms that relied on more heterogeneous clinical notes and problem lists. For example, one VTE NLP algorithm from Vanderbilt University had a sensitivity of 95% and PPV of 85% if the patients also had a concurrently positive ICD code.¹⁸ We believe the current NLP radiology algorithm is simpler and more representative as we sampled patients with both positive and negative ICD codes for VTE.

Finally, the combined ICD/NLP algorithm benefited from unique elements from each protocol to correctly classify the majority of the population as having true VTE (8%) or no VTE (87%). If applied appropriately, this combined algorithm could obviate the needs to chart review 95% of the cohort. Among the remaining 5% of people with discordant predictions (ICD positive but NLP negative or vice versa), approximately half of them had real VTE events. The only accurate way to assess the true outcome for this group is clinical chart review—the real VTE events detected by NLP but missed by ICD were more likely to be intra-abdominal splanchnic vein thrombosis. The ones detected by ICD but missed by NLP were more likely to be radiology procedures done at outside emergency department without available radiology reports.

There are limitations associated with the study. First, the VTE computable phenotype in the current study was derived and validated in a cancer population with a relatively high VTE incidence of

≈10% at one year; the predictive values of our algorithms may differ if they were applied to a noncancer population. Furthermore, unique efforts were taken in the current study to differentiate between bland venous thrombi and tumor thrombi. Second, the generalizability from Epic Clarity/Caboodle database to a non-Epic database (ie, Cerner or Sunquest) has not been studied and we could only internally validate our algorithms; however, health care systems with a preexisting electronic data warehouse can likely implement our search strategies. Third, the algorithms here would only work in a health care system with integrated EHR for inpatient and outpatient care. As such, academic hospitals with various contracted private practice physician groups likely would not have the consolidated data source readily available to them. Fourth, any NLP algorithm will require additional modifications depending on how the radiology impression reports are extracted. We recommend fastidious data selection (keep appropriate contrast- or Doppler-enhanced scans only with appropriate CPT codes) and data cleaning (remove indication for the scan and line-by-line description to only keep the final attending impression or interpretation) to improve the generalizability of the NLP algorithm to different institutions. Finally, since we reviewed a high proportion of charts to determine the best strategy for PPV optimization, 200 patients overlapped in both the initial discovery and the random sampling validation cohort in the ICD-positive group; there was no overlap in the ICD-negative group. We believe the use of a random sampling technique should mitigate some of the selection bias.

In summary, our ICD medication algorithm alone had a high PPV of 95% for correctly predicting acute VTE after cancer diagnosis. This can be further improved by combining with an NLP radiology algorithm to reach a weighted c statistic of 0.98. We recommend VTE researchers to use the ICD-based algorithm first after appropriate patient exclusion, and supplement it with the NLP algorithm if the institution has the appropriate data infrastructure and clinical informatics capability.

AUTHOR CONTRIBUTIONS

AL: conceptualization, methodology, data analysis, writing, manuscript review; WLdC Jr: data curation, data analysis, writing, manuscript review; DG: data curation, data analysis, manuscript review; EMM: data curation, manuscript review; AKA: data curation, manuscript review; KMK: data curation, manuscript review; FJN: data curation, manuscript review; MDP: data curation, manuscript review; RB: data curation, data analysis, manuscript review; CG: data curation, data analysis, manuscript review; CDW: data curation, data analysis, manuscript review; NAZ: conceptualization, methodology, manuscript review; CIA: conceptualization, methodology, manuscript review.

ACKNOWLEDGMENTS

This work was partially funded by a Research Training Award for Cancer Prevention Post-Graduate Training Program in Integrative Epidemiology from the Cancer Prevention & Research Institute of Texas, grant number RP160097 (principal investigator [PI]: M. Spitz),

the Systems Epidemiology of Cancer Training Program, grant number RP210037 (PI: A. Thrift). AL, a CPRIT Scholar in Cancer Research, was supported by the Cancer Prevention and Research Institute of Texas (RR190104).

RELATIONSHIP DISCLOSURE

The authors declare no conflicts of interest relevant to the content of this article.

ORCID

Ang Li  <https://orcid.org/0000-0002-8455-2309>

Wilson L. da Costa Jr  <https://orcid.org/0000-0003-4460-4706>

Neil A. Zakai  <https://orcid.org/0000-0001-8824-4410>

TWITTER

Ang Li  @AngLi_MD

REFERENCES

- Richesson R, Wiley L, Gold S. NIH pragmatic trials collaboratory - rethinking clinical trials. <https://rethinkingclinicaltrials.org/chapters/conduct/electronic-health-records-based-phenotyping/electronic-health-records-based-phenotyping-introduction/#references>. Accessed May 20, 2022.
- Richesson RL, Hammond WE, Nahm M, et al. Electronic health records based phenotyping in next-generation clinical trials: a perspective from the NIH health care systems collaboratory. *J Am Med Informatics Assoc*. 2013;20(E2):e226–e231.
- Tamariz L, Harkins T, Nair V. A systematic review of validated methods for identifying venous thromboembolism using administrative and claims data. *Pharmacoepidemiol Drug Saf*. 2012;21(Suppl 1):154-162.
- Sanfilippo KM, Wang T-F, Gage BF, Liu W, Carson KR. Improving accuracy of International Classification of Diseases codes for venous thromboembolism in administrative data. *Thromb Res*. 2015;135(4):616-620.
- Harris PA, Taylor R, Minor BL, et al. The REDCap consortium: Building an international community of software platform partners. *J Biomed Inform*. 2018;2019(95):103208.
- Harris PA, Taylor R, Thielke R, et al. Research electronic data capture (REDCap)-a metadata-driven methodology and workflow process for providing translational research informatics support. *J Biomed Inform*. 2009;42(2):377-381.
- Lyman GH, Eckert L, Wang Y, Wang H, Cohen A. Venous thromboembolism risk in patients with cancer receiving chemotherapy: a real-world analysis. *Oncologist*. 2013;18(12):1321-1329.
- Soysal E, Wang J, Jiang M, et al. CLAMP - a toolkit for efficiently building customized clinical natural language processing pipelines. *J Am Med Informatics Assoc*. 2018;25(3):331-336.
- White RH, Garcia M, Sadeghi B, et al. Evaluation of the predictive value of ICD-9-CM coded administrative data for venous thromboembolism in the United States. *Thromb Res*. 2010;126(1):61-67.
- Begg CB, Greenes RA. Assessment of diagnostic tests when disease verification is subject to selection bias. *Biometrics*. 1983;39(1):207.
- Koppel R, Lehmann CU. Implications of an emerging EHR monoculture for hospitals and healthcare systems. *J Am Med Inform Assoc*. 2015;22(2):465-471.
- LaMorfe WW. *Misclassification of outcome*. Boston University School of Public Health; 2020.
- Sanfilippo KM, Wang T-F, Luo S, et al. Predictive ability of the khorana score for venous thromboembolism (VTE) in multiple myeloma (MM). *J Clin Oncol*. 2018;36(15_suppl):e18733.

14. Pellathy T, Saul M, Clermont G, et al. Accuracy of identifying hospital acquired venous thromboembolism by administrative coding: implications for big data and machine learning research. *Journal of Clinical Monitoring and Computing*. 2021. doi:<https://doi.org/10.1007/s10877-021-00664-6>
15. Fang MC, Fan D, Sung SH, et al. Validity of using inpatient and outpatient administrative codes to identify acute venous thromboembolism. *Med Care*. 2017;55(12):e137-e143.
16. Reddy S, Lakhter V, Zack CJ, et al. Association between contemporary trends in inferior vena cava filter placement and the 2010 US Food and Drug Administration Advisory. *JAMA Intern Med*. 2017;177(9):1373-1374.
17. Heit J, Pathak J, Denny JGH. PheKB - a knowledgebase for discovering phenotypes from electronic medical records. *Venous Thromboembolism (VTE)*. <https://phekb.org/phenotype/venous-thromboembolism-vte>. Accessed May 20, 2022.
18. McPeck Hinz ER, Bastarache L, Denny JC. A natural language processing algorithm to define a venous thromboembolism phenotype. *AMIA Annu Symp Proc*. 2013;2013(7):975-983.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Li A, da Costa WL Jr, 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. doi:[10.1002/rth2.12733](https://doi.org/10.1002/rth2.12733)