HHS Public Access

Author manuscript

Eur J Haematol. Author manuscript; available in PMC 2024 December 01.

Published in final edited form as:

Eur J Haematol. 2023 December ; 111(6): 951–962. doi:10.1111/ejh.14110.

Artificial intelligence in the prediction of venous thromboembolism: A systematic review and pooled analysis

Thita Chiasakul[1,2,3], **Barbara D. Lam**[1,2], **Megan McNichol**[4], **William Robertson**[5,6], **Rachel P. Rosovsky**[7], **Leslie Lake**[5], **Ioannis S. Vlachos**[8], **Alys Adamski**[9], **Nimia Reyes**[9], **Karon Abe**[9], **Jeffrey I. Zwicker**[1,2,10], **Rushad Patell**[1,2]

[1]Division of Hematology, Department of Medicine, Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, Massachusetts, USA

[2]Division of Hemostasis and Thrombosis, Department of Medicine, Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, Massachusetts, USA

[3]Division of Hematology, Faculty of Medicine, Department of Medicine, Center of Excellence in Translational Hematology, Chulalongkorn University and King Chulalongkorn Memorial Hospital, Bangkok, Thailand

[4]Division of Knowledge Services, Department of Information Services (M.M.), Beth Israel Deaconess Medical Center, Boston, Massachusetts, USA

[5]National Blood Clot Alliance, Philadelphia, Pennsylvania, USA

[6]Department of Emergency Healthcare, College of Health Professions, Weber State University, Ogden, Utah, USA

[7]Division of Hematology/Oncology, Department of Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts, USA

[8]Department of Pathology, Cancer Research Institute, Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, Massachusetts, USA

[9]Division of Blood Disorders, National Center on Birth Defects and Developmental Disabilities, Centers for Disease Control and Prevention, Atlanta, Georgia, USA

Correspondence Rushad Patell, Beth Israel Deaconess Medical Center, 330 Brookline Ave, Boston, MA 02215, USA. rpatell@bidmc.harvard.edu.

AUTHOR CONTRIBUTIONS

Thita Chiasakul screened and selected the studies, extracted data, performed statistical analyses, and wrote the manuscript. Barbara D. Lam screened and selected the studies, and critically revised and approved the final manuscript. Megan McNichol performed the literature search. William Robertson critically revised and approved the final manuscript. Rachel P. Rosovsky critically revised and approved the final manuscript. Leslie Lake critically revised and approved the final manuscript. Ioannis S. Vlachos critically revised and approved the final manuscript. Alys Adamski critically revised and approved the final manuscript. Nimia Reyes critically revised and approved the final manuscript. Karon Abe critically revised and approved the final manuscript. Jeffrey I. Zwicker conceived and designed the study, and critically revised and approved the final manuscript. Rushad Patell conceived and designed the study, screened and selected the studies, and critically revised and approved the final manuscript.

CONFLICT OF INTEREST STATEMENT

The remaining authors declare no competing financial interests.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

[10]Department of Medicine, Hematology Service, Memorial Sloan Kettering Cancer Center, New York City, New York, USA

Abstract

Background: Accurate diagnostic and prognostic predictions of venous thromboembolism (VTE) are crucial for VTE management. Artificial intelligence (AI) enables autonomous identification of the most predictive patterns from large complex data. Although evidence regarding its performance in VTE prediction is emerging, a comprehensive analysis of performance is lacking.

Aims: To systematically review the performance of AI in the diagnosis and prediction of VTE and compare it to clinical risk assessment models (RAMs) or logistic regression models.

Methods: A systematic literature search was performed using PubMed, MEDLINE, EMBASE, and Web of Science from inception to April 20, 2021. Search terms included "artificial intelligence" and "venous thromboembolism." Eligible criteria were original studies evaluating AI in the prediction of VTE in adults and reporting one of the following outcomes: sensitivity, specificity, positive predictive value, negative predictive value, or area under receiver operating curve (AUC). Risks of bias were assessed using the PROBAST tool. Unpaired *t*-test was performed to compare the mean AUC from AI versus conventional methods (RAMs or logistic regression models).

Results: A total of 20 studies were included. Number of participants ranged from 31 to 111 888. The AI-based models included artificial neural network (six studies), support vector machines (four studies), Bayesian methods (one study), super learner ensemble (one study), genetic programming (one study), unspecified machine learning models (two studies), and multiple machine learning models (five studies). Twelve studies (60%) had both training and testing cohorts. Among 14 studies (70%) where AUCs were reported, the mean AUC for AI versus conventional methods were 0.79 (95% CI: 0.74–0.85) versus 0.61 (95% CI: 0.54–0.68), respectively ($p < .001$). However, the good to excellent discriminative performance of AI methods is unlikely to be replicated when used in clinical practice, because most studies had high risk of bias due to missing data handling and outcome determination.

Conclusion: The use of AI appears to improve the accuracy of diagnostic and prognostic prediction of VTE over conventional risk models; however, there was a high risk of bias observed across studies. Future studies should focus on transparent reporting, external validation, and clinical application of these models.

Keywords

artificial intelligence; prediction modeling; venous thromboembolism

1 | INTRODUCTION

Venous thromboembolism (VTE) is a common vascular disease that is associated with significant morbidity and mortality.[1,2] In patients presenting with suspected VTE, accurate and timely diagnosis is a prerequisite for appropriate medical intervention to prevent debilitating outcomes. In addition, accurate prediction of future VTE can facilitate the risk–

benefit consideration and allow for selection of high-risk patients who are most likely to benefit from pharmacological thromboprophylaxis.

A number of clinical risk prediction models have been developed for the diagnosis and prognostic prediction of VTE in various settings.[3–6] The reported performance of these models vary among population, baseline risks of VTE, and predictors included in the models. Traditionally, clinical risk prediction models are derived using a regression-based analysis, such as logistic regression and Cox regression, which result in several shortcomings including the limitation to highly structured and curated predictor variables. Artificial intelligence (AI) and machine learning modeling approaches have become increasingly popular as alternatives for the development of prediction models. While these approaches provide theoretical advantages, including more computational flexibility and consistency,[7] they remain susceptible to bias and are often hard to clinically interpret which limits their application. Moreover, the predictive performance of these AI-based models has not been consistently reported to perform better than conventional models.

Evidence for AI or machine learning-based models in the diagnosis and prognostic prediction of VTE has been accumulating over the past few years. This study aims to systematically review the performance of AI or machine learning-based models in the diagnosis and prediction of VTE and compare their performance with conventional clinical risk assessment models (RAMs) or regression-based models.

2 | METHODS

The study protocol is registered on PROSPERO (CRD248869). Preferred Reporting Items for Systematic Reviews and Meta-analyses (PRISMA) guidelines were followed.

2.1. | Data sources and search strategies

PubMed, MEDLINE, EMBASE, and Web of Science from inception to April 20, 2021 were queried. The following search terms were used: (“artificial intelligence” OR “machine learning” OR “natural language processing”) AND (“venous thromboembolism” OR “deep vein thrombosis” OR “venous thrombosis” OR “pulmonary embolism”). Language restriction to English was applied. Result from the detailed searches are presented in the Supporting Information Methods.

2.2. | Study selection

The eligible studies were original prospective or retrospective studies evaluating the performance of AI-based models to diagnose or predict occurrence of VTE in adults, defined as age ≥ 18 years of age. Diagnostic predictions models predict the probability of VTE in the presenting population, whereas prognostic prediction models predict the probability of developing VTE in the future. Studies were required to report one of the following outcomes: sensitivity, specificity, positive predictive value, negative predictive value, or area under receiver operating curve (AUC). Non-original articles (such as reviews, commentaries, or guidelines) and duplicated studies were excluded. Studies assessing natural language processing (NLP) accuracy to detect VTE from radiological reports or medical records were not included.

2.3 | Data extraction

Two authors (Thita Chiasakul and Barbara D. Lam) independently extracted data from included studies in duplicate using a standardized evidence table based on CHARMS (critical appraisal and data extraction for systematic reviews of prediction modeling studies) checklist.[8] Discrepancies were resolved by consensus or a third reviewer (Rushad Patell) when necessary. The primary outcome was the diagnostic or prognostic performance of AI-based clinical prediction models in VTE. The following data were collected: author, year of publication, study design, study population, inclusion and exclusion criteria, number of participants, AI model, the outcome being predicted, internal and external validation method, discrimination and calibration performance measure, and percentage of missing data.

2.4 | Risk of bias assessment

Methodological quality assessment was performed independently by two authors (Thita Chiasakul and Barbara D Lam) using the PROBAST (Prediction model Risk of Bias Assessment Tool).[9,10] The tool consisted of four key domains: participants, predictors, outcomes, and analysis. Studies were categorized by their risk of bias as having low, high, or unclear risk of bias. Any differences in quality rating were resolved by consensus or adjudication by a third reviewer (Rushad Patell).

2.5 | Statistical analysis

Due to the substantial heterogeneity in terms of study population, types of data sources, models, and outcomes of interest that were observed among the included studies, meta-analyses for the point estimates for the overall model performance were not performed. Narrative summary and descriptive statistics were used to describe the characteristics of included studies and the model's performance measures. Unpaired *t*-test was performed to compare the mean AUC from AI versus conventional methods (RAMs or logistic regression models). In studies reporting multiple models, the models with highest reported AUC were selected for analysis.

3 | RESULTS

3.1 | Study identification

The study was reported in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) and TRIPOD guidelines. The PRISMA flow diagram is shown in Figure 1. A total of 745 unique records were retrieved from the literature search. After screening by title and abstract, 656 records were excluded. The remaining 89 references underwent full-text review, 20 of which met eligibility criteria and were included in the systematic review. Fourteen studies provided adequate data for a pooled analysis. Collectively, these 14 studies included 249 111 patients.

3.2 | Study and patient characteristics of included studies

The characteristics of included studies are summarized in Tables 1 and 2. Of the 20 included studies, 13 were of prognostic prediction models[11–23] and 7 were of diagnostic

prediction models.[24–30] Among the 13 prognostic prediction models studies, the outcome being predicted were first VTE (8 studies), post-operative VTE (4 studies), and recurrent VTE (1 study). Among the seven diagnostic prediction model studies, the conditions being predicted were pulmonary embolism (PE; four studies), deep vein thrombosis (DVT; one study), arterial and venous thromboembolism (one study), and portal vein thrombosis (one study). The publication years ranged from 2004 to 2021. Studies that were included were conducted in the United States (10 studies), Europe (6 studies), Asia (3 studies), and South America (1 study). All studies were retrospective, with only one study reporting results from prospective validation.[16] The sample size ranged from 31 to 111 888 patients. The study population included patients presenting with suspected VTE or first VTE, ambulatory cancer patients, hospitalized medical patients, patients with anti-phospholipid syndrome, and post-operative patients. Data sources were mostly medical records obtained from single or multiple institutions, whereas two studies utilized data from an administrative database.[14,23] Methods of VTE outcome assessment were described in seven studies (35%); two studies used ICD codes and five studies used standard imaging such as duplex ultrasound and computed tomography pulmonary angiography.

3.3 | Models of AI

The AI-based models included artificial neural network (six studies), support vector machines (four studies), super learner ensemble (one study), Bayesian methods (one study), genetic programming (one study), unspecified machine learning models (two studies), and multiple machine learning models (including random forest, gradient boosting decision tree, logistic regression, support vector machine, K-nearest neighbor, and Naive-Bayes; five studies). Twelve studies (60%) reported both training and testing cohorts, with the proportion of testing cohorts ranging from 10%–30%. Predictor variables included clinical and laboratory variables that varied among studies, ranging from 3 to 68 variables. Of the 20 studies included, internal validation was reported in 11 studies (55%) and external validation was performed in 2 studies (10%). Methods of internal validation included bootstrapping, cross-validation, and data splitting. Missing values were excluded in five studies, whereas two studies utilized the predictive value imputation method by replacing missing values with the average of the attribute observed in the training set. The remaining 13 studies (65%) did not include their approach to missing data.

3.4 | Performance measures

The model discrimination and calibration performance measures are summarized in Table 1. Discrimination measures, reported as AUCs or c-statistics, were described in 14 studies (70%). Among these studies, confidence intervals were reported in only three studies. The mean AUC for AI-based models compared to conventional models (including previously published clinical prediction models and logistic regression models) were 0.79 (95% CI: 0.74–0.85) vs. 0.61 (95% CI: 0.54–0.68), respectively ($p < .001$, Figure 2). Calibration performance measures investigated using calibration plots and the Hosmer–Lemeshow test were reported in only three studies.

3.5 | Risk of bias assessment

Summary of the risk of bias assessment are shown in Figure 3. Most studies had high or unclear risk of bias according to the PROBAST tool. The common sources of bias among the included studies were absence of detailed descriptions of participant inclusion and exclusion criteria, lack of reporting on the definitions, methods of assessment, and blinding procedures for model predictors and outcome ascertainment, and lack of report on missing data handling and details of model calibration.

4 | DISCUSSION

In this systematic review, we identified 20 studies that evaluated the performance of AI-based prediction models in the diagnostic and prognostic prediction of VTE. The use of AI appears to provide superior discrimination performance than the conventional regression-based models; however, there were a number of shortcomings identified.

The included studies were heterogenous; each model was aimed to predict VTE in different medical contexts, such as predicting first VTE in an outpatient or inpatient settings, predicting recurrence after the first VTE, or predicting VTE after different surgical procedures. Thus, we did not summarize and compare the model performance across studies. Moreover, many studies fell short in adequately describing the inclusion and exclusion criteria of the study population. These details are paramount to understand the clinical utility and appropriate application of prediction models.

Concerning issues in the modeling process were noticeable among the included studies. The currently available studies of AI-based models mostly have high risks of bias. Essential elements of the model development, validation, and evaluation method were often omitted from the reports, hindering the appraisal of their performance, applicability, and reproducibility. In some studies,[11,17,26,27] sample sizes were limited, which may have led to overfitting of models. The time span of prediction, defined as the period between predictor assessment and outcomes, was not described in most studies, which also limits the clinical applicability. Moreover, less than half of the studies reported the proportion of missing data and how they were handled. Missing data introduces bias in the model development and affects the validity of the model's predictive performance.[31]

Although most of the included studies reported good to excellent discriminative performance (AUC ranging from 0.7 to 0.9), this accomplishment is unlikely to be replicated when used in clinical practice, owing to the studies' high risk of bias. Of the 20 included studies, only 2 studies performed external validation,[16,24] which showed that the model's performance in the validation cohort was inferior to the derivation cohort. In addition, our review observed significant gaps in the reporting of model's calibration performance (presented as calibration plot or Hosmer–Lemeshow test), which evaluates the ability of the model to accurately estimate the risk of the outcome. Ideally, reliable prediction models should show strong agreement between the predicted outcomes and the observed outcomes.[32] A model can have excellent discrimination but poor calibration, over- or under-estimating the individual's risk of outcome.[33] The use of such a model would be misleading and can be detrimental in clinical practice. Performing independent external validation and a rigorous evaluation of

AI-based model performance is an important step before actual clinical implementation. One advantage of the AI-based prediction models is their ability to undergo transfer learning, a machine learning method whereby a pre-trained model can be adapted to different populations and data set, making it more accurate and cost-effective.[34]

Of the 20 included studies, 8 studies reported the comparative performance of AI-based models to conventional models in the same data set.[13,14,18,20,22,24,25,29] In all eight studies, the intra-study comparison showed superior discriminatory performance of AI-based models. We reported that the mean AUC of AI-based models was higher than that of conventional models. Although this finding was not intended to be interpreted as overall summarization of the performance of individual models, it demonstrates the potential for integration of AI in the development of risk prediction models in VTE. A recent systematic review and meta-analysis of AI approaches, including NLP, in the prediction and diagnosis of VTE reported pooled sensitivity of 0.87 and pooled specificity of 0.96 in the testing data set based on five studies.[35] The heterogeneity was very high in this study (I^2 ranging from 93.6% to 99.4%), which was expected when combining studies with marked variation in clinical settings and objectives. Our review excluded NLP studies, due to their differences in purpose and characteristics from the machine learning models. In another systematic review comparing the performance of machine learning to logistic regression in 71 studies across various clinical domains, the AUCs of machine learning models were higher than those of logistic regression only in studies with high risk of bias.[36]

Despite the endorsement of the TRIPOD checklist,[37] inadequate reporting of published risk prediction models have been observed in both AI- and non-AI-based clinical prediction models.[36,38,39] Similar to the studies included in our analysis, common elements that were often omitted were the description of participants, sample size justification, definition of predictors and outcomes, missing data handling, calibration performances, and external validation. AI approaches in medicine include the more auditable algorithms, which are typically more interpretable and rely heavily on human annotation to accurately label features and outputs, and the more “black box” models, such as neural networks, which can be highly computationally complex and differ significantly from statistical techniques. These models often rely on nonlinear relationships between predictors and outcomes.[40] In depth comparisons of these methodologies and justifications of the approaches are important details that were broadly lacking in studies included in this systematic review. This should be a focus in future studies for clinical applications of AI including VTE management. AI algorithms, particularly in supervised learning for prediction modeling, are vulnerable to biases of their human designers.[41] These biases can be challenging to identify or rectify, and can be compounded by biases present in the datasets or the algorithm itself. It is important to be aware of these potential biases, as the use of AI becomes increasingly common, in order to improve the efficiency and accessibility of healthcare delivery.[42]

Aside from the model’s performance, the clinical applicability of an AI-based prediction model is dependent upon many other important factors, such as the availability of the model’s algorithm and software, the ability for model updating, security assurance, and the integration to clinical workflow.[43] There is an ongoing debate regarding the capacity of AI-generated algorithms to be safeguarded under intellectual property laws. This controversy

could potentially hinder developers from sharing their models with the public.[44] Moreover, after implementation, the impact of AI-based prediction models on clinical outcomes and decision making need continued evaluation. Thrombotic disorders and anticoagulation are the most commonly assessed domains in studies assessing electronic health record integration and implementation of clinical prediction models, which has shown promising impact on clinical outcomes.[45] A retrospective study evaluating the influence of AI-based clinical decision support system reported a 19% reduction in the rate of hospital-acquired VTE after implementation. Unfortunately, the details of model development were not reported in this study.[46] Moving forward, such implementation data are crucial in order to assess the applicability and impact of AI-based clinical prediction models in clinical care.

Despite the limitations, our review serves to provide a comprehensive overview of current literature evaluating the use of AI in the prediction models in the area of VTE and highlights the apparent inadequacy in the reporting of current studies. It is foreseeable that there will be an exponential increase in the number of reports on AI-based VTE prediction models in the upcoming decade. Our review identifies the current methodological and reporting challenges and issues at this early stage, such that future studies can take caution to improve reporting transparency and appraisability, ultimately leading to improved overall quality of evidence in this area.

5 | CONCLUSION

The use of AI-based models for diagnosis and prediction of VTE are increasing and might potentially be an improvement compared to existing models. Adherence to the standard reporting guidelines for clinical prediction models can increase the quality of future studies evaluating AI-based prediction models in VTE. At this stage of evidence, it is premature to incorporate AI-based models in routine clinical practice. Implementation of these models require consideration of other important factors including ethical and legal compliance, transparency, integration to clinical workflows, and continuous evaluation. Future studies should focus on transparent reporting, external validation, and clinical application of these models.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

The authors acknowledge the support by the Centers for Disease Control and Prevention (CDC).

FUNDING INFORMATION

This study was supported by the Centers for Disease Control and Prevention (CDC), Atlanta, GA Cooperative Agreement #DD20-2002. The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention.

Center of Disease Control and Prevention (CDC)

Rachel P. Rosovsky reports institutional research support from BMS, Janssen; Advisory/Consultancy for Abbott, BMS, Dova, Inari, Janssen, Penumbra. Jeffrey I. Zwicker reports prior research funding from Incyte and

Quercegen; consultancy for Sanofi, CSL Behring, and Calyx. Rushad Patell reports research funding from National Blood Clot Alliance and American Society of Clinical Oncology.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

REFERENCES

1. Brahmandam A, Abougergi MS, Ochoa Chaar CI. National trends in hospitalizations for venous thromboembolism. J Vasc Surg Venous Lymphat Disord. 2017;5(5):621.e2–629.e2.
2. ISTH Steering Committee for World Thrombosis Day. Thrombosis: a major contributor to the global disease burden. J Thromb Haemost. 2014;12(10):1580–1590. [PubMed: 25302663]
3. Darzi AJ, Karam SG, Charide R, et al. Prognostic factors for VTE and bleeding in hospitalized medical patients: a systematic review and meta-analysis. Blood. 2020;135(20):1788–1810. [PubMed: 32092132]
4. Ensor J, Riley RD, Moore D, Snell KI, Bayliss S, Fitzmaurice D. Systematic review of prognostic models for recurrent venous thromboembolism (VTE) post-treatment of first unprovoked VTE. BMJ Open. 2016; 6(5):e011190.
5. Pandor A, Daru J, Hunt BJ, et al. Risk assessment models for venous thromboembolism in pregnancy and in the puerperium: a systematic review. BMJ Open. 2022;12(10):e065892.
6. Moik F, Ay C, Pabinger I. Risk prediction for cancer-associated thrombosis in ambulatory patients with cancer: past, present and future. Thromb Res. 2020;191(Suppl 1):S3–S11. [PubMed: 32736775]
7. Steyerberg EW, van der Ploeg T, Van Calster B. Risk prediction with machine learning and regression methods. Biom J. 2014;56(4): 601–606. [PubMed: 24615859]
8. Moons KGM, de Groot JAH, Bouwmeester W, et al. Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist. PLoS Med. 2014;11(10):e1001744.
9. Collins GS, Dhiman P, Andaur Navarro CL, et al. Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. BMJ Open. 2021;11(7):e048008.
10. Wolff RF, Moons KGM, Riley RD, et al. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. Ann Intern Med. 2019;170(1):51–58. [PubMed: 30596875]
11. Agharezaei L, Agharezaei Z, Nemati A, et al. The prediction of the risk level of pulmonary embolism and deep vein thrombosis through artificial neural network. Acta Inform Med. 2016;24(5):354–359. [PubMed: 28077893]
12. Ferroni P, Zanzotto FM, Scarpato N, Riondino S, Guadagni F, Roselli M. Validation of a machine learning approach for venous thromboembolism risk prediction in oncology. Dis Markers. 2017; 2017:8781379.
13. Ferroni P, Zanzotto FM, Scarpato N, et al. Risk assessment for venous thromboembolism in chemotherapy-treated ambulatory cancer patients. Med Decis Making. 2017;37(2):234–242. [PubMed: 27491558]
14. Gowd AK, Agarwalla A, Amin NH, et al. Construct validation of machine learning in the prediction of short-term postoperative complications following total shoulder arthroplasty. J Shoulder Elbow Surg. 2019;28(12):e410–e421. [PubMed: 31383411]
15. Hollon TC, Parikh A, Pandian B, et al. A machine learning approach to predict early outcomes after pituitary adenoma surgery. Neurosurg Focus. 2018;45(5):E8.
16. Kline JA, Novobilski AJ, Kabrhel C, Richman PB, Courtney DM. Derivation and validation of a Bayesian network to predict pretest probability of venous thromboembolism. Ann Emerg Med. 2005;45(3):282–290. [PubMed: 15726051]

17. Martins TD, Annichino-Bizzacchi JM, Romano AVC, Maciel FR. Artificial neural networks for prediction of recurrent venous thromboembolism. Int J Med Inform. 2020;141:104221.
18. Nafee T, Gibson CM, Travis R, et al. Machine learning to predict venous thrombosis in acutely ill medical patients. Res Pract Thromb Haemost. 2020;4(2):230–237. [PubMed: 32110753]
19. Sabra S, Mahmood Malik K, Alobaidi M. Prediction of venous thromboembolism using semantic and sentiment analyses of clinical narratives. Comput Biol Med. 2018;94:1–10. [PubMed: 29353160]
20. Shah AA, Devana SK, Lee C, Kianian R, van der Schaar M, SooHoo NF. Development of a novel, potentially universal machine learning algorithm for prediction of complications after Total hip arthroplasty. J Arthroplasty. 2021;36(5):1655.e1–1662.e1.
21. Xue B, Li D, Lu C, et al. Use of machine learning to develop and evaluate models using preoperative and intraoperative data to identify risks of postoperative complications. JAMA Netw Open. 2021;4(3):e212240. [PubMed: 33783520]
22. Yang Y, Wang X, Huang Y, Chen N, Shi J, Chen T. Ontology-based venous thromboembolism risk assessment model developing from medical records. BMC Med Inform Decis Mak. 2019;19(Suppl 4):151. [PubMed: 31391095]
23. James S, Suguness A, Hill A, Shatzel J. Novel algorithms to predict the occurrence of In-hospital venous thromboembolism: machine learning classifiers developed from the 2012 National Inpatient Sample. Chest. 2015;148(4):492A.
24. Banerjee I, Sofela M, Yang J, et al. Development and performance of the Pulmonary Embolism Result Forecast Model (PERFORM) for computed tomography clinical decision support. JAMA Netw Open. 2019; 2(8):e198719.
25. Biesheuvel CJ, Siccama I, Grobbee DE, Moons KG. Genetic programming outperformed multivariable logistic regression in diagnosing pulmonary embolism. J Clin Epidemiol. 2004;57(6):551–560. [PubMed: 15246123]
26. Kremers R, Zuily S, De Groot P, Hemker C, Wahl D, De Laat B. Development of a neural network to predict thrombosis in the anti-phospholipid syndrome with an accuracy of 87%. Blood. 2018; 132(Supplement 1):3796.
27. Liu K, Chen J, Zhang K, Wang S, Li X. A diagnostic prediction model of acute symptomatic portal vein thrombosis. Ann Vasc Surg. 2019; 61:394–399. [PubMed: 31352086]
28. Willan J, Katz H, Keeling D. The use of artificial neural network analysis can improve the risk-stratification of patients presenting with suspected deep vein thrombosis. Br J Haematol. 2019;185(2):289–296. [PubMed: 30727024]
29. Falsetti LG, Merelli E, Rucco M, et al. A data-driven clinical prediction rule for pulmonary embolism. Eur Heart J. 2013;34(suppl_1):P243.
30. Gay D, Elliott CG, Snow G. Derivation of an electronic prediction tool, ePE, for prediction of pulmonary embolism In the emergency department. B65 Pulmonary Embolic Disease. American Journal of Respiratory and Critical Care Medicine. 2013;187:A3313.
31. Zhang Z. Missing values in big data research: some basic skills. Ann Transl Med. 2015;3(21):323. [PubMed: 26734633]
32. Steyerberg EW, Vickers AJ, Cook NR, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology. 2010;21(1):128–138. [PubMed: 20010215]
33. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW, Topic Group ‘Evaluating diagnostic tests and prediction models’ of the STRATOS initiative. calibration: the Achilles heel of predictive analytics. BMC Med. 2019;17(1):230. [PubMed: 31842878]
34. Zhang X, Xue Y, Su X, et al. A transfer learning approach to correct the temporal performance drift of clinical prediction models: retrospective cohort study. JMIR Med Inform. 2022;10(11):e38053.
35. Wang Q, Yuan L, Ding X, Zhou Z. Prediction and diagnosis of venous thromboembolism using artificial Intelligence approaches: a systematic review and meta-analysis. Clin Appl Thromb Hemost. 2021; 27:10760296211021162.
36. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. J Clin Epidemiol. 2019;110:12–22. [PubMed: 30763612]

37. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD Statement. BMC Med. 2015;13(1):1. [PubMed: 25563062]

38. Groot OQ, Ogink PT, Lans A, et al. Machine learning prediction models in orthopedic surgery: a systematic review in transparent reporting. J Orthop Res. 2022;40(2):475–483. [PubMed: 33734466]

39. Sanfilippo KM, Wang TF, Carrier M, et al. Standardization of risk prediction model reporting in cancer-associated thrombosis: communication from the ISTH SSC subcommittee on hemostasis and malignancy. J Thromb Haemost. 2022;20(8):1920–1927. [PubMed: 35635332]

40. Sidey-Gibbons JAM, Sidey-Gibbons CJ. Machine learning in medicine: a practical introduction. BMC Med Res Methodol. 2019;19(1):64. [PubMed: 30890124]

41. Norori N, Hu Q, Aellen FM, Faraci FD, Tzovara A. Addressing bias in big data and AI for health care: a call for open science. Patterns (N Y). 2021;2(10):100347.

42. Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. Nat Med. 2022;28(1):31–38. [PubMed: 35058619]

43. Singh RP, Hom GL, Abramoff MD, Campbell JP, Chiang MF, Intelligence AAOTFoA. Current challenges and barriers to real-world artificial Intelligence adoption for the healthcare system, provider, and the patient. Transl Vis Sci Technol. 2020;9(2):45.

44. George A, Walsh T. Artificial intelligence is breaking patent law. Nature. 2022;605(7911):616–618. [PubMed: 35610374]

45. Lee TC, Shah NU, Haack A, Baxter SL. Clinical implementation of predictive models embedded within electronic health record systems: a systematic review. Informatics (MDPI). 2020;7(3):25. [PubMed: 33274178]

46. Zhou S, Ma X, Jiang S, et al. A retrospective study on the effectiveness of artificial Intelligence-based clinical decision support system (AI-CDSS) to improve the incidence of hospital-related venous thromboembolism (VTE). Ann Transl Med. 2021;9(6):491. [PubMed: 33850888]

Novelty Statements

What is the new aspect of your work?

This is a systematic review of literature evaluating the performance of artificial intelligence (AI)-based models in the diagnostic and prognostic prediction of venous thromboembolism (VTE).

What is the central finding of your work?

Although AI-based models may potentially have superior performance to conventional models, current studies have high risks of bias and inadequate reporting of methods.

What is (or could be) the specific clinical relevance of your work?

Before the widespread clinical implementation of AI-based models for the prediction of VTE risk, clinicians should be aware of the potential bias present in the current literature and their applicability.

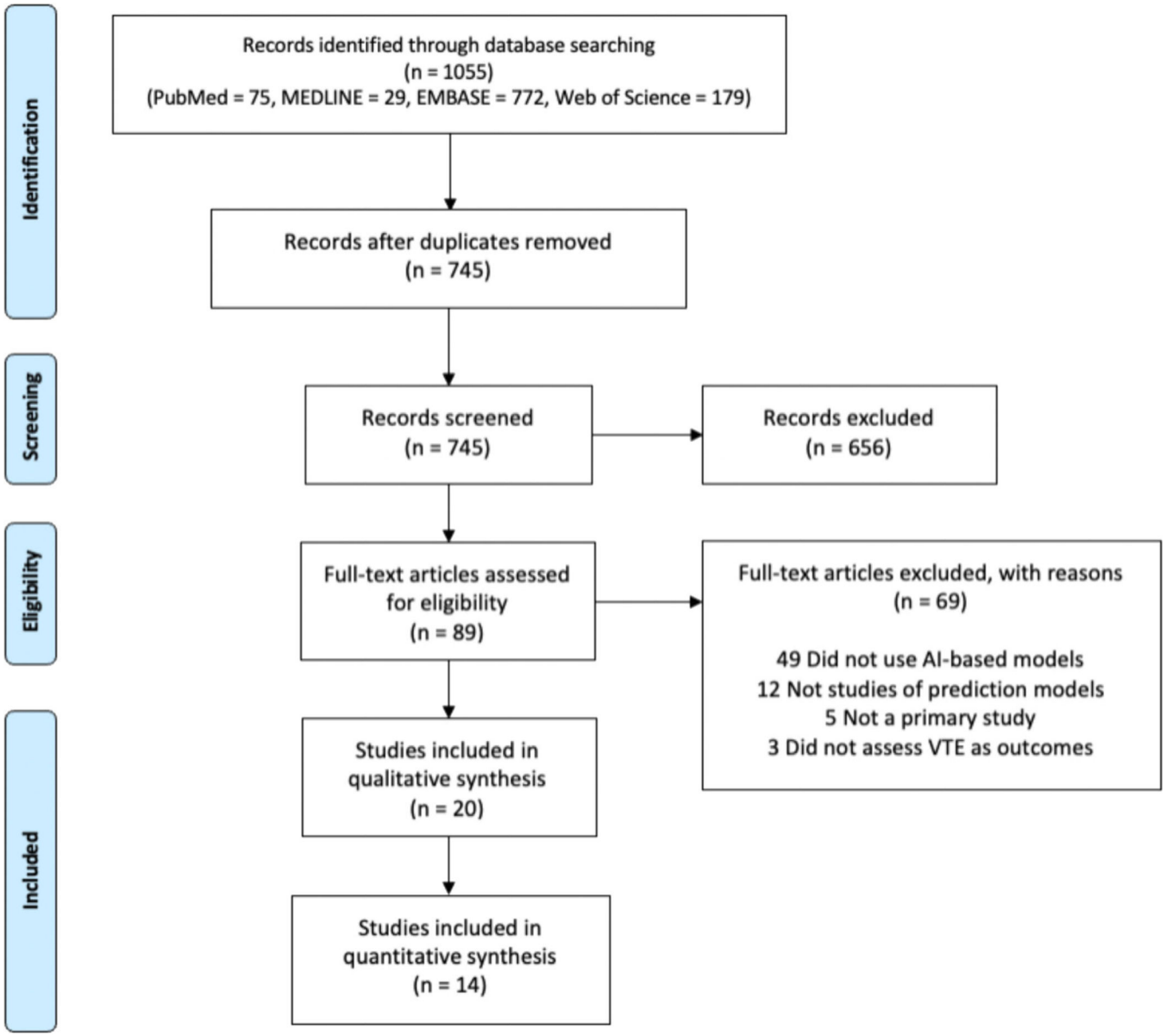

FIGURE 1.
PRISMA flow diagram.

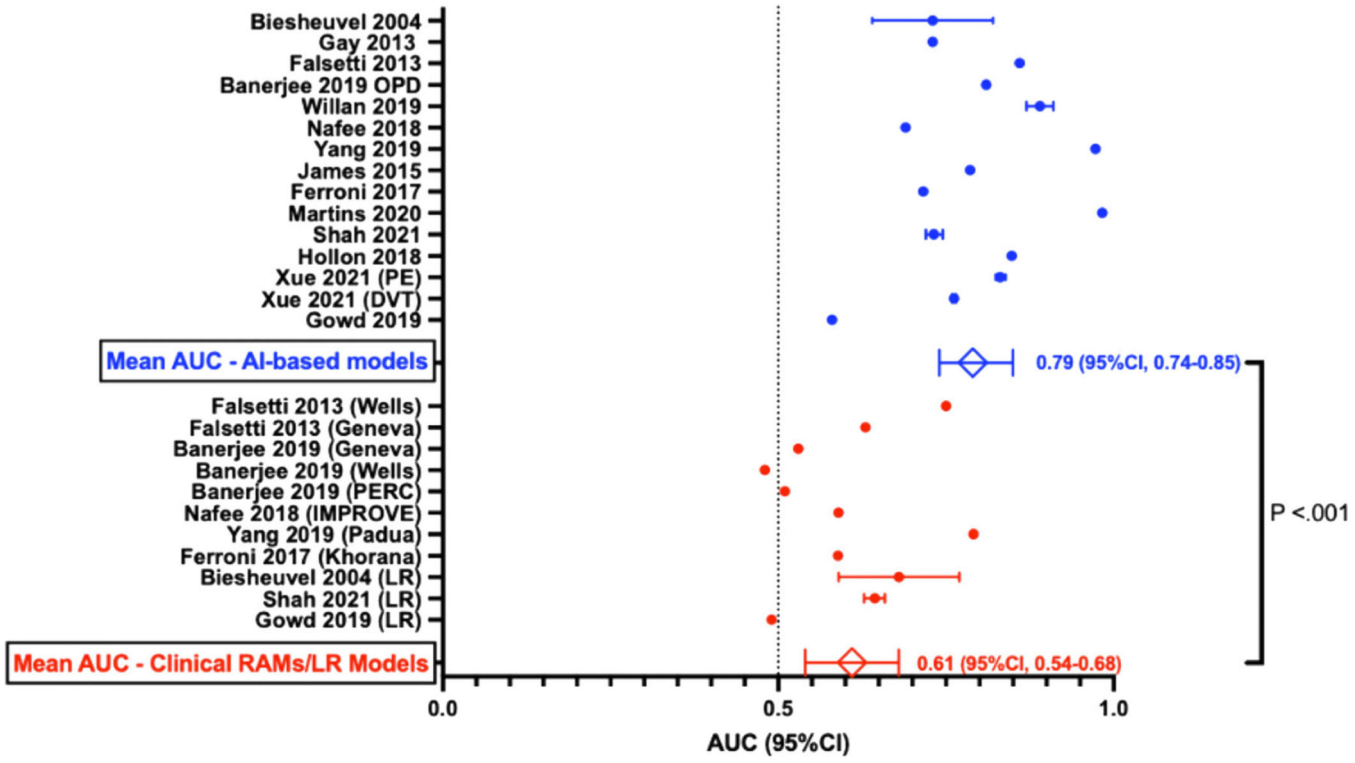

FIGURE 2.

Scatter plot comparing the area under receiver operating curve (AUC) in AI-based models compared to previous clinical risk assessment models (RAMs)/logistic regression models.

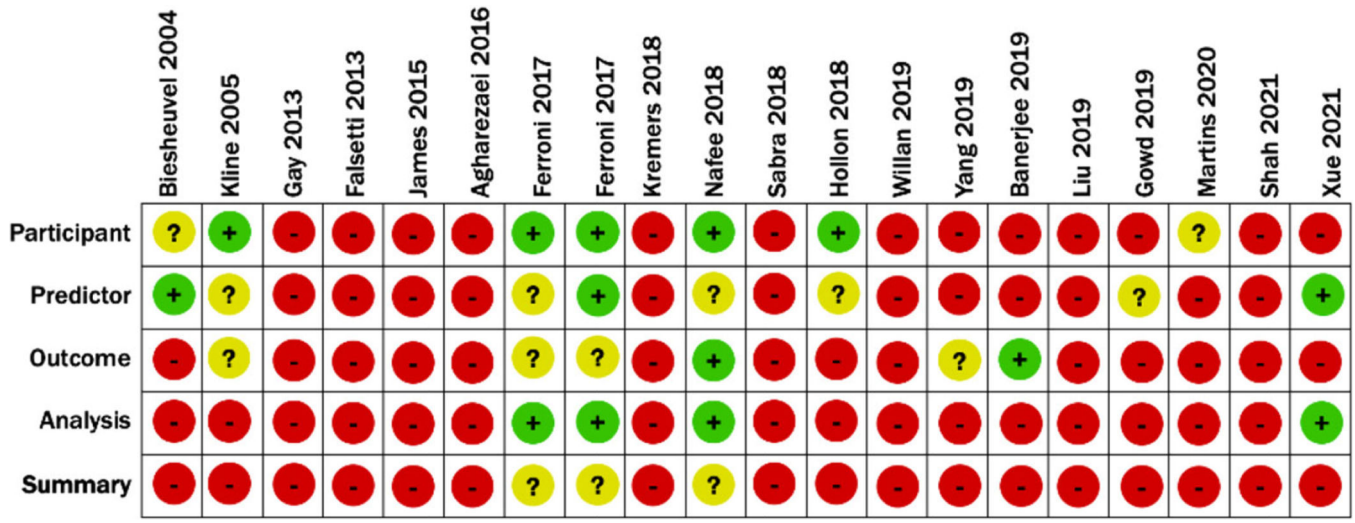

FIGURE 3.

Risk of bias assessment using PROBAST tool. +, low risk of bias;?, unclear risk of bias −, high risk of bias.

TABLE 1

Characteristics of included studies and model performance measures.

Study	Country	Population	Total, *N*	Outcome to be predicted	Outcome prevalence	Model performance measure	
						Discrimination	**Calibration**
Studies of diagnostic prediction model (*N* = 7)							
Biesheuvel, 2004	Netherlands	Adults with suspected PE	398	PE	43%	Genetic programming: AUC, 0.73 (95% Cl: 0.64–0.82) Logistic regression: AUC, 0.68 (95% Cl: 0.59–0.77)	Calibration plot (Hosmer-Lemeshow test $p > .50$)
Gay, 2013 (Abstract)	USA	Patients who underwent CT pulmonary angiography for suspected PE	3500	PE	NR	AUC: 0.73	NR
Falsetti, 2013 (Abstract)	Italy	Outpatients with suspected PE	755	PE	NR	ANN: AUC, 0.86 Wells score: AUC, 0.75 rGeneva score: AUC, 0.63	NR
Banerjee, 2019	USA	CTPA images from adult patients	3214	PE	15.8%–61.2%	Derivation cohort: PE neural: AUC 0.81 ElasticNet: AUC, 0.73 PERC score: AUC, 0.51 Wells score: AUC, 0.48 rGeneva score: AUC, 0.53 Validation cohort: PE neural: AUC, 0.81 ElasticNet: AUC, 0.74 PERC score: AUC, 0.60 Wells score: AUC, 0.51 rGeneva score: AUC, 0.47	NR
Willan, 2019	UK	Patients with suspected DVT	7080	DVT	11.6%	AUC, 0.89 (0.87–0.91)	NR
Kremers, 2018 (Abstract only)	Netherlands	Antiphospholipid syndrome	31	History of thrombosis (arterial, venous, small vessels	68%	Sensitivity: 86.7% Specificity: 63.0% Accuracy: 74.9%	N/A
Liu, 2019	China	Patients with PVT and matched controls	141	Acute symptomatic PVT	33%	Sensitivity: 0.92 (0.79–0.97) Specificity: 1.00 (95.1–1.00)	N/A
Studies of prognostic prediction models (N = 13)							
Nafee, 2018	USA	Hospitalized medical illness (APEX trials)	6459	VTE	6.3%	ML: c-statistics, 0.69 Sensitivity: 0.57 (0.43–0.75) Specificity: 0.72 (0.53–0.84) IMPROVE score: c-statistics, 0.59	Calibration plot (Hosmer-Lemeshow test $p > .06$ (ML), p = .44 (rML) versus IMPROVE score ($p < .001$)
Agharezaei, 2016	Iran	Hospitalized patients with PE/DVT	294	VTE	NR	Accuracy: 93.23%	NR
Yang, 2019	China	Internal medicine inpatients	3106	VTE	7.2%	ML: AUC 0.973 (±0.006) Sensitivity: 0.900 ± 0.037	NR

Study	Country	Population	Total, *N*	Outcome to be predicted	Outcome prevalence	Model performance measure	
						Discrimination	**Calibration**
						Specificity: 0.918 ± 0.012 Padua score: AUC, 0.791 ± 0.022	
Kline, 2005	USA	Emergency department patients with suspected VTE	4568	VTE	8%–ll%	Sensitivity: 0.95 (0.92–0.97) Specificity: 0.30 (0.08–0.32)	NR
Sabra, 2018	USA	NR	150	VTE	NR	Sensitivity: 0.857	NR
Ferroni, 2017	Italy	Ambulatory cancer patients	1179	VTE	8%	Sensitivity: 0.889	NR
James, 2015 (Abstract only)	USA	Oncology patients	1000	in-hospital VTE	NR	AUC 0.786 (95% Cl NR)	NR
Ferroni, 2017	Italy	Ambulatory cancer patients	608	VTE	7.1%	ML: AUC 0.716 Khorana score: AUC 0.589	NR
Martins, 2020	Brazil	Patients with first provoked/ unprovoked VTE	235	Recurrent VTE	21%	AUC up to 0.983	NR
Shah, 2021	USA	Hip arthroplasty patients	89 986	Post-operative complications (included PE)	NR	AutoPrognosis: AUC 0.732 (95% Cl: 0.720–0.745) Logistic regression: AUC, 0.644 (85% Cl: 0.628–0.659)	Calibration plot (similar to logistic regression)
Hollon, 2018	USA	Pituitary adenoma patients treated with surgical resection via an endoscopic endonasal approach	400	Post-operative complications (included DVT and PE)	NR	AUC, 0.83	NR
Xue, 2021	USA	Post-operative patients	111 888	Post-operative complications (included DVT and PE)	NR	DVT: AUC, 0.831 (95% Cl: 0.824–0.839) PE: AUC, 0.762 (95% Cl: 0.759–0.765)	NR
Gowd, 2019	USA	Patients who underwent total shoulder replacement procedures	17 119	Post-operative complications (included DVT and PE)	NR	Random forest: AUC DVT/PE 0.58 Logistic regression: AUC DVT/PE 0.49	NR

Abbreviation: AUC, area under the curve; CTPA, computed tomography pulmonary angiography; DVT, deep vein thrombosis; NR, not reported; PE, pulmonary embolism; VTE, venous thromboembolism.

TABLE 2

Characteristics of models used in each study.

Study	AI Models	Data input	Training cohort	Testing cohort	Missing data	Internal validation method	External validation
Studies of diagnostic prediction model (*N* = 7)							
Biesheuvel, 2004	Genetic programming	10 clinical variables	165 (67%)	133 (33%)	NR	Bootstrapping Split of data (random)	None
Gay, 2013 (Abstract)	Machine learning	27 clinical variables	Not described	Not described	NR	NR	None
Falsetti, 2013 (Abstract)	Artificial neural network	24 clinical, instrumental and laboratory variables	67%	33%	NR	Split of data (in supervised classification step)	None
Banerjee, 2019	Machine learning/artificial neural network	Temporal features within 1 year prior to PE encounter	3057 (40%)	340 (10%)	NR	Split of data (random); 10-fold cross validation	Yes (*N* = 240)
Willan, 2019	Artificial neural network	13 variables (Sex, age, D-dimer, and 10 components of Wells' score)	5270 (75%)	1810 (25%)	Excluded (38%)	Split of data (divided by date)	None
Kremers, 2018 (Abstract only)	Multiple neural networks	APL antibody profile (aCL and aB2GPI) with/without thrombin generation test	31	NR	NR	NR	None
Liu, 2019	LASSO-SVM	43 clinical data	141	NR	NR	10-fold cross-validation	None
Studies of prognostic prediction models (*N* = 13)							
Nafee, 2018	Super learner ensemble method (machine learning)	ML 68 baseline variables from APEX rML 16 variables thought to be risk factors of VTE	6459	–	Included only variables with <1% missing data	10-fold cross-validation	None
Agharezaei, 2016	Artificial neural network	31 clinical risk factors	235 (80%)	59 (20%)	Excluded (32%)	Split of data (random)	None
Yang, 2019	NLP and machine learning (random forest, gradient boosting decision tree, logistic regression, and support vector machine)	Ontology terms extracted from medical records	80%	20%	NR	Split of data (random)	
Kline, 2005	Bayesian network	25 clinical variables	3145	1423	NR	NA	Yes
Sabra, 2018	SESARF-SVM	Clinical narratives and list of known VTE risk factors	120	30	NR	8% in validation cohort	
Ferroni, 2017	Multiple kernel learning (machine learning)—SVM and RO	Clinical and laboratory attributes clustered into 9 groups	70%	30%	NR	Three-fold crossvalidation	
James, 2015 (Abstract only)	Machine learning	NR	80%	20%	NR	Split of data (random)	None
Ferroni, 2017	Multiple kernel learning (Machine learning)—SVM and RO	Clinical and laboratory attributes clustered into 9 groups	–	608	Imputation methods	Split of data (random)	None

Study	AI Models	Data input	Training cohort	Testing cohort	Missing data	Internal validation method	External validation
Martins, 2020	Artificial Neural Network	39 clinical factors	80%	20%	Excluded	Five-fold cross-validation	None
Shah, 2021	Ensemble machine learning: random forest, AdaBoost, gradient boosting machines, and XGBoost	Clinical variables	Not described	Not described	NR	Five-fold cross-validation	None
Hollon, 2018	Supervised machine learning: naïve Bayes, logistic regression with elastic net regularization, support vector machines with linear kernel, and random forest	26 patient characteristics	300 (75%)	100 (25%)	NR	Split of data (random) 10-fold cross-validation	None
Xue, 2021	Machine learning: support vector machine, logistic regression, random forest, gradient boosting tree (GBT), and deep neural network (DNN)	Preoperative and intraoperative variables	Not described	Not described	Imputation methods	Five-fold cross validation	None
Gowd, 2019	Machine learning: logistic regression, K-nearest neighbor, random forest, Naive-Bayes, decision tree, and gradient boosting trees.	22 clinical features	13 697 (80%)	3422 (20%)	Excluded	Split of data (random)	None

Abbreviation: LASSO-SVM, least absolute shrinkage and selection operator—support vector machine; NLP, natural language processing; NR, not reported; SESARF, Semantic Extraction and Sentiment Assessment of Risk Factors.