Published in final edited form as:
Thromb Haemost. 2023 June ; 123(6): 649–662. doi:10.1055/a-2039-3222.

Developing Validated Tools to Identify Pulmonary Embolism in Electronic Databases: Rationale and Design of the PE-EHR+ Study

Behnood Bikdeli, MD, MS[1,2,3,4], **Ying-Chih Lo, MD**[5], **Candrika D. Khairani, MD, MMSc**[2], **Antoine Bejjani, MD**[2], **David Jimenez, MD, PhD**[6], **Stefano Barco, MD, PhD**[7,8], **Shiwani Mahajan, MD, MHS**[3,9], **César Caraballo, MD**[3], **Eric A. Secemsky, MD, MSc**[10,11,12], **Frederikus A. Klok, MD, PhD**[13], **Andetta R. Hunsaker, MD**[14], **Ayaz Aghayev, MD**[14], **Alfonso Muriel, PhD**[15], **Yun Wang, PhD**[3,10], **Mohamad A. Hussain, MD, PhD**[16,17], **Abena Appah-Sampong, MD**[18], **Yuan Lu, ScD**[3], **Zhenqiu Lin, PhD**[3], **Sanjay Aneja, MD**[19], **Rohan Khera, MD, MS**[3,20], **Samuel Z. Goldhaber, MD**[1,2], **Li Zhou, MD, PhD**[5], **Manuel Monreal, MD, PhD**[21], **Harlan M. Krumholz, MD, SM**[3,20,22], **Gregory Piazza, MD, MS**[1,2]

[1.]Cardiovascular Medicine Division, Brigham and Women’s Hospital, Harvard Medical School, Boston, MA, USA.

[2.]Thrombosis Research Group, Brigham and Women’s Hospital, Harvard Medical School, Boston, MA, USA.

[3.]YNHH/ Yale Center for Outcomes Research and Evaluation (CORE), New Haven, CT, USA.

[4.]Cardiovascular Research Foundation (CRF), New York, NY, USA.

[5.]Division of General Internal Medicine and Primary Care, Department of Medicine, Brigham and Women’s Hospital, Harvard Medical School, Boston, MA, USA.

[6.]Respiratory Department, Hospital Ramón y Cajal and Medicine Department, Universidad de Alcalá (Instituto de Ramón y Cajal de Investigación Sanitaria), Centro de Investigación Biomédica en Red de Enfermedades Respiratorias, Madrid, Spain.

[7.]Department of Angiology, University Hospital Zurich, Zurich, Switzerland.

[8.]Center for Thrombosis and Hemostasis, Johannes Gutenberg University Mainz, Mainz, Germany.

[9.]Department of Internal Medicine, Yale School of Medicine, New Haven, CT, USA.

[10.]Richard A. and Susan F. Smith Center for Outcomes Research in Cardiology, Department of Medicine, Beth Israel Deaconess Medical Center, Boston, MA, USA

[11.]Harvard Medical School, Boston, MA, USA

[12.]Division of Cardiology, Department of Medicine, Beth Israel Deaconess Medical Center, Boston, MA, USA.

Corresponding author information: Dr. Bikdeli, Cardiovascular Medicine Division, Brigham and Women’s Hospital, 75 Francis Street, Boston, MA 02115; bbikdeli@bwh.harvard.edu, Behnood.bikdeli@yale.edu Twitter handle: @bbikdeli.

[13] Department of Medicine - Thrombosis & Hemostasis, Leiden University Medical Centre, Leiden, The Netherlands.

[14] Department of Radiology, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA.

[15] Clinical Biostatistics Unit. Hospital Universitario Ramón y Cajal. IRYCIS. CIBERESP: Universidad de Alcalá. Madrid,Spain.

[16] Division of Vascular and Endovascular Surgery, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA.

[17] Centre for Surgery and Public Health, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA.

[18] Department of Surgery, Brigham and Women's Hospital, Boston, MA, USA.

[19] Department of Therapeutic Radiology, Yale School of Medicine, New Haven, CT, USA.

[20] Section of Cardiovascular Medicine, Department of Internal Medicine, Yale School of Medicine, New Haven, CT, USA.

[21] Cátedra de Enfermedad Tromboembólica, Universidad Católica de Murcia, Murcia, Spain.

[22] Department of Health Policy and Management, Yale School of Public Health, New Haven, CT, USA.

Abstract

Background: Contemporary pulmonary embolism (PE) research, in many cases, relies on data from electronic health records (EHRs) and administrative databases that use International Classification of Diseases (ICD) codes. Natural language processing (NLP) tools can be used for automated chart review and patient identification. However, there remains uncertainty with the validity of ICD-10 codes or NLP algorithms for patient identification.

Methods: The PE-EHR+ study has been designed to validate ICD-10 codes as Principal Discharge Diagnosis, or Secondary Discharge Diagnoses, as well as NLP tools set out in prior studies to identify patients with PE within EHRs. Manual chart review by two independent abstractors by predefined criteria will be the reference standard. Sensitivity, specificity, and positive and negative predictive values will be determined. We will assess the discriminatory function of code subgroups for intermediate– and high-risk PE. In addition, accuracy of NLP algorithms to identify PE from radiology reports will be assessed.

Results: A total of 1,734 patients from the Mass General Brigham health system have been identified. These include 578 with ICD-10 Principal Discharge Diagnosis codes for PE, 578 with codes in the secondary position, and 578 without PE codes during the index hospitalization. Patients within each group were selected randomly from the entire pool of patients at the Mass General Brigham health system. A smaller subset of patients will also be identified from the Yale-New Haven Health System. Data validation and analyses will be forthcoming.

Conclusions: The PE-EHR+ study will help validate efficient tools for identification of patients with PE in EHRs, improving the reliability of efficient observational studies or randomized trials of patients with PE using electronic databases.

INTRODUCTION

Annually, ~1,000,000 new cases of fatal or non-fatal pulmonary embolism (PE) occur in the United States and Europe.[1–6] Traditional cohort studies and registries continue to inform the epidemiology, prognosis, and outcomes of PE.[7–15] In turn, randomized controlled trials (RCTs) have informed the safety and efficacy of interventions, such as type and dose of anticoagulation and the utility of advanced therapies.[7–9] However, many questions about PE epidemiology and comparative effectiveness of health interventions remain unanswered. Despite the merits of traditional cohort studies and RCTs for informing PE epidemiology and effectiveness of PE treatment options, individual patient screening and enrollment with traditional methods are resource intensive. Prospective enrollment at large scales such as a national level is also burdensome and often unfeasible. Therefore, more efficient ways are needed to identify patients with PE.

Electronic databases such as electronic health records (EHRs) or large administrative databases are advantageous for patient selection in retrospective studies. EHRs are also helpful for case selection in prospective observational studies, or for case selection in RCTs, as they can be screened fairly quickly. Querying the EHRs is more efficient than prospective manual screening of clinical practices.

The most common way to identify patients with PE through electronic databases is by using the International Classification of Diseases (ICD) codes. In recent years, ICD codes were revised to 10th modification (ICD-10). These codes make it possible for investigators to query individual hospitals or health system records, or to analyze large insurance databases, such as assessment of regional or national practice patterns, or trends in PE incidence and outcomes.[16–20] The American Heart Association (AHA) uses the codes for the annual *Heart Disease and Stroke Statistics*.[1, 21] The Agency for Healthcare Research and Quality (AHRQ) uses the PE ICD-10 codes to track perioperative quality of care.[21] Observational comparative effectiveness studies have used these codes to share routine practice perspectives complementing RCT results and providing insights in contexts in which an RCT is unfeasible.[22, 23] Recently, ICD codes have had novel uses such as patient screening and successful inclusion in pragmatic RCTs for cardiovascular diseases.[24]

Natural language processing (NLP), a branch of artificial intelligence, uses computers to transform unstructured data into analyzable variables.[25–31] NLP has received growing attention in biomedical research.[32] NLP is attractive for identification of patients with PE since it can potentially use various sections of the medical records including imaging reports for computed tomography pulmonary angiography (CTPA) or ventilation-perfusion imaging to confirm the diagnosis of PE, or even to automate additional features for screening or risk stratification.

However, there are important knowledge gaps related to the optimal approach for case selection of patients with PE. The existing studies using ICD-10 (Table 1 with codes, Table 2 with studies)[33–40] or NLP (Table 3)[27–31 38, 39] have had limitations including small number or being from a single center, lack of sharing sufficient details including about the location of the codes (in the principal discharge diagnosis position versus secondary discharge diagnosis position), or limited cross-validation. The PE-EHR+ study has been designed to address these gaps in knowledge and to validated efficient tools for identification of patients with PE in electronic databases.

METHODS

General design features and data sources

The PE-EHR+ study has three distinct and complementary goals: 1. To validate ICD-10 codes, including the location and subtype of codes for selection of patients with PE through EHRs; 2. To validate an efficient NLP algorithm for selection of patients with PE in EHRs that have electronic versions of the imaging reports available; and 3. As a practical application of the codes, we will use the ICD-10 codes to report the trends in PE hospitalization and outcomes via validated ICD-10 codes in a national database of patients with PE in the United States (Figure 1).

For the first aim, we will use data from the Mass General Brigham (MGB) Health System, in Massachusetts, USA. MGB includes several community hospitals and 2 large referral hospitals. It has been also pre-specified to screen and explore an additional subset of charts from another large health system from the U.S. (the Yale-New Haven Health System). The Institutional Review Board at Brigham and Women’s Hospital (BWH) reviewed the study protocol and approved it, waiving the need for informed consent (IRB #2022P001226). For chart review from other sites, related Institutional Review Board approval will be obtained. The study will be performed at the Thrombosis Research Group at BWH, in close collaboration with the Medical Text Extraction, Reasoning, and Mapping System (MTERMS) laboratory at BWH, and the Yale-New Haven Hospital/ Yale Center for Outcomes Research & Evaluation (CORE).

The initial study protocol was used as a platform for generation of the list for patient identification by two authors (YCL and BB). We selected the patient cohort from Enterprise Data Warehouse of MGB by using the following criteria: (1) patient age equal to or greater than 18 years (2) inpatient encounter (hospitalization) with diagnosis date between January 1, 2016 and December 31, 2021. In the process of patient selection (see below) in addition to obtaining data related to presence or absence and position of the ICD-10 discharge codes for PE, we collected information such as age, sex, admission diagnosis, admission date and discharge date for further reviewing purpose.

Study Samples

Three distinct groups of patients will be identified: A. Patients with ICD-10 Principal Discharge Diagnosis (primary codes) for PE, B. Patients with Secondary Discharge Diagnosis for PE (but no PE codes in the primary position), and C. Patients in whom no ICD-10 PE codes were mentioned during the index hospitalization event, either in the primary or in the secondary positions. A list of ICD-10 PE codes and their definitions is summarized in Table 1. Table S1 summarizes the search query for identification of prior studies.

ICD-10 codes were introduced into practice in the U.S. since the fourth quarter of 2015. Considering a potential learning curve in the health systems, we set the period for inclusion of patients and their hospitalization events from January 1, 2016 through December 31, 2021. If a given patient had multiple hospitalizations with similar patterns of codes in the study period (e.g., multiple hospitalizations with secondary discharge diagnosis of PE), only one hospitalization was selected randomly.

As an exploratory goal, if resources allow, we will also explore the accuracy of ICD-10 codes for chronic thromboembolic pulmonary hypertension (CTEPH) (I27.24).

Exposure variable and data extraction for the ICD-10 code analysis

The main exposure variable is the presence of ICD-10 codes for PE in the primary position, secondary position, or none at all in the discharge records in the ICD-10 code analysis.

The reference standard for identification of PE will be chart review by two trained independent clinician abstractors using standardized definitions (Table 4). The data abstraction form will be created and piloted in five charts per group. Once the form is finalized, the study protocol will be made available to abstractors. The abstractors will then review the patient charts, including imaging studies, discharge summaries, and other records, to verify the diagnosis of PE. For review of each individual chart, the abstractors will have full access to electronic medical records, but not the designated ICD-10 codes in the research database, to provide unbiased assessment of each chart. Discrepancies between the two abstractors’ findings will be discussed and, if unresolved, will be decided by input from the Principal Investigator. In the unlikely event that PE ascertainment is not feasible for a given chart, that chart will be excluded (see statistical analysis).

Exposure variable and data extraction for the NLP analyses

The main exposure variable in the NLP analysis will be the presence of PE based on NLP automated review of radiology reports. The reference standard for identification of PE will be chart review by trained clinician abstracts, as summarized above.

EHRs provide large amounts of data for research. While data elements such as laboratory tests are structured, medical notes or imaging reports are created as free text without pre-defined structured data elements.[26, 41–43] Natural language, such as words in medical charts, are not typically “*coded*” or conducive to computations for case selection or statistical analyses in research studies. The resource-intense nature of manual chart review to abstract data from free-text fields precludes timely or large-scale analyses.

NLP re-encodes free-text notes (natural language) into structured format that facilitate data extraction and analysis. Briefly, EHR-based NLP techniques can be grouped into three categories: 1) Keyword searches or rule-based systems; 2) supervised learning systems; and

3) unsupervised learning systems. The development of a successful NLP algorithm entails multiple steps including tokenization, word stemming, lemmatization, and others (Table 5).[25] NLP can handle synonyms, acronyms, and typos that are added in the system (e.g., *embolsim* instead of *embolism*). Once the algorithm is derived (training set) and validated (testing set), with satisfiable performance, it can conduct the disease identification task automatically.

Outcome variables—The main outcomes will be the sensitivity, specificity, positive and negative predictive values of the ICD-10 codes for determining PE compared with medical chart review. These will be based on standard epidemiological definitions. In addition, we will determine the accuracy of these codes (defined as true positive plus true negative, divided by the combination of true positive, true negative, false positive, and false negative) (Table 6). Outcomes for the NLP analyses will be similar.

Statistical analysis

With respect to sample size estimates, we will select an equal number of patients with and without ICD-10 codes for PE to facilitate the assessment of both sensitivity and specificity of the codes for PE. With a two-sided alpha of 0.05 and confidence interval width of 10%, a sample of 550 per group (550 with ICD-10 codes and 550 without) provides 80% power to detect a positive predictive value of 80% for the PE-related ICD-10 codes compared with manual chart review. To assess patients who had a secondary discharge diagnosis ICD-10 PE codes, a separate set of 550 charts will be selected. Assuming a need to exclude 5% of the charts, 578 charts will be planned for review (total of 1,734 charts). Once the review of these charts is completed, to approximate the true incidence of PE, weighting will be applied to the completed database.

The total number of hospitalized patients with ICD-10 Principal Discharge Diagnosis of PE in the MGB in the aforementioned period (January 1, 2016 through December 31, 2021) is 4,878. The number of patients hospitalized with ICD-10 Secondary Discharge Diagnosis of PE is 3,224; whereas 373,540 adult patients did not have any codes for PE during their hospitalization. These are relatively similar to estimates from prior studies.[18, 44, 45] To be able to provide accurate estimates for not only sensitivity and specificity, but also other measures of test performance which may depend on prevalence of the studied condition, we will be weighing the results of the three 550-patient groups of patients proportionate to their actual size, before measures of test performance are calculated for ICD-10 codes in the primary discharge position, or secondary discharge position. A similar approach will be pursued to determine the measures of test performance for NLP compared with manual chart review.

Categorical variables will be reported with frequency counts and percentages. Test characteristics will be reported with their respective 95% confidence interval estimates. Weighting will not affect the sample size estimate for specificity.

Sensitivity analysis and subgroup analyses—We will conduct exploratory analyses in which a combination of thrombosis-related diagnostic (e.g., computed tomography pulmonary angiography) or therapeutic procedure codes (e.g., fibrinolytic therapy or vena cava filter placement, Table S2), or present-on-admission condes, will be added to the ICD-10 discharge codes to assess whether they improve the accuracy for patient identification compared with the ICD-10 codes alone.

Further, we will conduct analyses to assess the validity of specific subgroups of PE codes. For example, some PE codes indicate hemodynamic consequences (e.g., I26.0: pulmonary embolism with acute cor pulmonale). As the availability of subgroup-specific samples allow, the validity of the code subsets for classifying patient status will be compared against manual medical chart review with reference to definitions from the international clinical guidelines.[7, 8] Consistency of the results across the participating hospitals will be assessed. Consistency of the codes' accuracy will be also checked for patients included before versus after the COVID-19 pandemic.[46–49] In addition, if the resources allow, we may check the accuracy of the codes in the subgroup of patients with active cancer (diagnosed within prior 5 years and on treatment, palliative care, or close surveillance) and will investigate the trends in accuracy of codes over time.

In addition, the diagnosis of subsegmental PE has been a subject of intense debate.[50] We have pre-specified to validate the reports of subsegmental PE by independent verification of the diagnosis by two independent certified radiologists among 50–100 patients.

Practical implementation of ICD-10 codes

Finally, as a practical part of the PE-EHR study, the validated ICD-10 codes will be used to identify patients with PE in a 100% sample of patients in the Medicare Fee-For-Service database to report the trends in PE hospitalizations and mortality rates. Such analyses will be complemented by trends analyses from the Registro Informatizado de Pacientes con Enfermedad TromboEmbólica (RIETE) registry.[14]

RESULTS

As of July 11, 2022, a total of 1,734 patients from the hospitals in the Mass General Brigham health system have been identified. Of 1,734 patients, 578 had an ICD-10 Principal Discharge Diagnosis codes for PE, 578 patients had ICD-10 Secondary Discharge Diagnosis codes for PE, and 578 did not have any codes for PE codes during the index hospitalization event. Manual validation of the charts is ongoing. Analyses for the accuracy of the codes and analyses with NLP will be forthcoming in subsequent years. The process of

DISCUSSION

The PE-EHR+ study provides a unique opportunity to validate the tools for efficient identification of patients with PE via EHRs using ICD-10 codes and NLP algorithms (Figure 2). With respect to ICD-10 code validation, PE-EHR+ has several strengths compared with the existing investigations and will complement their findings.[33–40] Unlike several other studies, PE-EHR+ has a pre-specified power calculation. In addition, discharge records will be reviewed from both community hospitals and large referral hospitals with a diverse patient population. Further, we will separately assess the accuracy of the codes in the Principal Discharge Diagnosis versus Secondary Discharge Diagnosis positions. From one end, it is conceivable that PE codes in the Principal Discharge Diagnosis position have a higher specificity and positive predictive value for patient identification. In contrast, Principal Discharge Diagnosis codes may underestimate the PE burden, since PE events in some situations may be a complication of the hospitalization but not severe enough to warrant designation as the Principal Discharge Diagnosis. Coders who focus only on discharge summaries may miss radiology reports that would identify PE diagnoses.[51] PE codes placed as Secondary Discharge Diagnosis may be more sensitive but are prone to false positive findings. This is because PE may be coded in secondary discharge positions in patients with prior events that were relevant for the clinical care delivered in the index hospitalization, but were not acute events that occurred in that index hospitalization. An important strength of the PE-EHR+ study is that it includes not only hospitalization records for patients with claims codes for PE, but also hospitalization records for patients without PE claims codes. This gives the opportunity to ascertain the specificity and positive predictive value of the codes, but also the possibility of false negative results, and sensitivity of the codes. The pre-defined weighting criteria will be helpful in this process, as well. With respect to NLP algorithms for identification of PE[27–31], the PE-EHR+ study has the opportunity to validate those results in a large database of patients from diverse hospital settings and may modify the existing algorithms, as needed.

Pre-specified plan to validate the subgroups of the codes that may capture higher-risk is also of particular interest. Many questions about the epidemiology and durable outcomes for contemporary patients with intermediate-risk PE and high-risk PE remain unanswered. If the ICD-10 codes or NLP are proven to be efficient and reliable for patient screening, they may facilitate patient selection in future epidemiological or comparative effectiveness studies. Similarly, the ancillary goal to assess the accuracy of the codes against the original reports for sub-segmental PE, and to also validate the original diagnosis of subsegmental PE by review of images by two independent radiologists, will provide important novel data.

The components of the project related to validation of ICD-10 codes and NLP algorithms are meant to complement but not supplant each other. For example, some data sources (such as national administrative data) do not include radiology reports or medical notes, and as such, NLP will not be feasible in those data sources. In turn, in EHRs, use of NLP might be advantageous or even further, in databases that have access to both NLP and ICD-10 codes, a hybrid approach that incorporates both ICD-10 codes and NLP might yield the highest accuracy.

We did not pre-specify a particular threshold to consider a high enough accuracy (defined as combination of true positives and true negatives divided by all observations). Although an ideal test has both high sensitivity and positive predictive value (and therefore accuracy), it is possible that no single permutation of codes is able to achieve both goals, but that different combinations of codes would be required for maximizing sensitivity vs PPV.

The limitations of the PE-EHR+ study should be kept in mind for appropriate context and interpretation. First, this study will be focusing on PE. The available resource will not lend support to expand to other thrombotic conditions. As such, efficient and reliable tools will be similarly needed for identification of patients with deep vein thrombosis, or arterial thrombotic events such as acute myocardial infarction, ischemic stroke, and acute limb ischemia. Second, the reference standard for verification of PE in this study is review of medical records for presence of PE in the chart, but not independent re-assessment of the testing modalities that led to the diagnosis of the PE events in every case. Considering that the study is based on existing chart records, this can potentially be associated with certain limitations. However, prospective enrollment of such a large sample would require several years and enormous resources. In most cases with initial radiologist confirmation of PE in larger branches or the main pulmonary arteries, a false positive diagnosis is very unlikely.[52, 53] Subsegmental PEs may be an area of potential concern. To mitigate that, we have made *a priori* plans to do independent validation of the diagnosis for 50–100 patients with sub-segmental PE according to the imaging reports. Third, we should acknowledge that the original phase of the PE-EHR+ study will only include data from several centers in the United States. While the overall structure of the PE codes are similar around the world, minor differences with respect to granular subgroups of codes may exist. With several international investigators in the Steering Committee of the PE-EHR+, we envision to test the optimized algorithms identified through PE-EHR+ in future studies of non-US data sources to ascertain the consistency of the findings. Fourth, implementation of NLP algorithms for chart screening and automated abstraction is a complex resource-intensive undertaking. Therefore, the main focus will be on radiology reports, which are more structured and desirable for NLP. Further, we will perform external validation of the existing NLP algorithms used in studies for thrombotic diseases.[27–31] If their accuracy is suboptimal, modifications will be planned to optimize them. The teams at MTERMS and CORE have ample expertise to provide guidance for accomplishment of the project goals related to NLP. Finally, COVID-19 is associated with excess risk of venous thromboembolism[46–48] and may potentially impact PE presentation or how the codes were used, even among non-COVID-19 patients.[49] Therefore, we will do a sensitivity analysis for the codes, restricting the results to the pre-pandemic period.

In conclusion, the PE-EHR study will help validate efficient tools for identification of patients with PE in EHRs. These include ICD-10 codes in the Principal Discharge Diagnosis or Secondary Discharge Diagnosis positions, and NLP algorithms based on assessment of imaging reports. These validated tools will facilitate the timely use of EHRs for case selection for observational studies or randomized trials of patients with PE.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements:

Figure 2 was created using BioRender.com.

Funding:

Dr. Bikdeli is supported by a Career Development Award from the American Heart Association and VIVA Physicians (#938814) for the PE-EHR+ study. Outside the submitted work, Dr. Bikdeli is supported by the Scott Schoen and Nancy Adams IGNITE Award, as well as the Mary Ann Tynan Research Scientist award from the Mary Horrigan Connors Center for Women's Health and Gender Biology at Brigham and Women's Hospital, and the

Heart and Vascular Center Junior Faculty Award from Brigham and Women's Hospital. Dr. Hussain is funded by a Heart and Vascular Center Junior Faculty Award from Brigham and Women's Hospital.

Disclosures:

Dr. Bikdeli reports that he is a consulting expert, on behalf of the plaintiff, for litigation related to two specific brand models of IVC filters. Dr. Klok has received research support from Bayer, Bristol-Myers Squibb, MSD, BSCI, Leo Pharma, Actelion, The Netherlands Organisation for Health Research and Development, The Dutch Thrombosis Association, The Dutch Heart Foundation and the Horizon Europe Program, unrelated to the current work and paid to his institution. Dr. Barco has received research support from Boston Scientific, Bard, Novartis, Bayer, and Concept Medical; and consulting fees from Bayer, Concept Medical, Boston Scientific, and INARI. Dr. Khera reports grant support from the National Heart, Lung, and Blood Institute of the US National Institutes of Health, and the Doris Duke Charitable Foundation, and grant support, through Yale University, from Bristol-Myers Squibb, and has served on the Bristol-Myers Squibb digital health advisory board, all outside of the submitted work. Dr Goldhaber has received research support from Bayer, Bristol Myers Squibb, Boston Scientific BTG EKOS, Janssen, National Heart, Lung, and Blood Institute, and Pfizer; and has received consulting fees from Agile, Bayer, and Pfizer, outside the submitted work. Dr Krumholz received expenses and/or personal fees from UnitedHealth, Element Science, Aetna, Reality Labs, Tesseract/4Catalyst, the Siegfried and Jensen Law Firm, Arnold and Porter Law Firm, Martin/Baughman Law Firm, and F-Prime; being a cofounder of Refactor Health and HugoHealth; and contracts through Yale New Haven Hospital from the Centers for Medicare & Medicaid Services and through Yale University from Johnson & Johnson. Dr. Piazza has received research support from Bristol-Myers Squibb/Pfizer Alliance, Bayer, Janssen, Alexion, Amgen and Boston Scientific Corporation, and consulting fees from Bristol-Myers Squibb/Pfizer Alliance, Boston Scientific Corporation, Janssen, Namsa, Prairie Education and Research Cooperative, Boston Clinical Research Institute, and Amgen.

REFERENCES:

1. Virani SS, Alonso A, Aparicio HJ, Benjamin EJ, Bittencourt MS, Callaway CW, Carson AP, Chamberlain AM, Cheng S, Delling FN, Elkind MSV, Evenson KR, Ferguson JF, Gupta DK, Khan SS, Kissela BM, Knutson KL, Lee CD, Lewis TT, Liu J, Loop MS, Lutsey PL, Ma J, Mackey J, Martin SS, Matchar DB, Mussolino ME, Navaneethan SD, Perak AM, Roth GA, Samad Z, Satou GM, Schroeder EB, Shah SH, Shay CM, Stokes A, VanWagner LB, Wang NY,Tsao CW. Heart Disease and Stroke Statistics-2021 Update: A Report From the American Heart Association. Circulation. 2021;143:e254–e743. [PubMed: 33501848]
2. Cohen AT, Agnelli G, Anderson FA, Arcelus JI, Bergqvist D, Brecht JG, Greer IA, Heit JA, Hutchinson JL, Kakkar AK, Mottier D, Oger E, Samama MM, Spannagl M, Europe VTEIAGi. Venous thromboembolism (VTE) in Europe. The number of VTE events and associated morbidity and mortality. Thromb Haemost. 2007;98:756–64. [PubMed: 17938798]
3. Heit JA, Cohen AT, Anderson FJ. Estimated annual number of incident and recurrent, non-fatal and fatal venous thromboembolism (VTE) events in the US. Blood. 2005;106:1.
4. Bikdeli B, Bikdeli B. Updates on Advanced Therapies for Acute Pulmonary Embolism. Int J Cardiovasc Pract; 2016; 1: 47–50.
5. Barco S, Mahmoudpour SH, Valerio L, Klok FA, Münzel T, Middeldorp S, Ageno W, Cohen AT, Hunt BJ,Konstantinides SV. Trends in mortality related to pulmonary embolism in the European Region, 2000–15: analysis of vital registration data from the WHO Mortality Database. Lancet Respir Med. 2020;8:277–87. [PubMed: 31615719]
6. Barco S, Valerio L, Ageno W, Cohen AT, Goldhaber SZ, Hunt BJ, Iorio A, Jimenez D, Klok FA, Kucher N, Mahmoudpour SH, Middeldorp S, Münzel T, Tagalakis V, Wendelboe AM,Konstantinides SV. Age-sex specific pulmonary embolism-related mortality in the USA and Canada, 2000–18: an analysis of the WHO Mortality Database and of the CDC Multiple Cause of Death database. Lancet Respir Med. 2021;9:33–42. [PubMed: 33058771]
7. Konstantinides SV, Meyer G, Becattini C, Bueno H, Geersing GJ, Harjola VP, Huisman MV, Humbert M, Jennings CS, Jiménez D, Kucher N, Lang IM, Lankeit M, Lorusso R, Mazzolai L, Meneveau N, F NÁ, Prandoni P, Pruszczyk P, Righini M, Torbicki A, Van Belle E,Zamorano JL. 2019 ESC Guidelines for the diagnosis and management of acute pulmonary embolism developed in collaboration with the European Respiratory Society (ERS). Eur Heart J. 2020;41:543–603. [PubMed: 31504429]
8. Giri J, Sista AK, Weinberg I, Kearon C, Kumbhani DJ, Desai ND, Piazza G, Gladwin MT, Chatterjee S, Kobayashi T, Kabrhel C,Barnes GD. Interventional Therapies for Acute Pulmonary

Embolism: Current Status and Principles for the Development of Novel Evidence: A Scientific Statement From the American Heart Association. Circulation. 2019;140:e774–e801. [PubMed: 31585051]

9. Ortel TL, Neumann I, Ageno W, Beyth R, Clark NP, Cuker A, Hutten BA, Jaff MR, Manja V, Schulman S, Thurston C, Vedantham S, Verhamme P, Witt DM, I DF, Izcovich A, Nieuwlaat R, Ross S, H JS, Wiercioch W, Zhang Y,Zhang Y. American Society of Hematology 2020 guidelines for management of venous thromboembolism: treatment of deep vein thrombosis and pulmonary embolism. Blood Adv. 2020;4:4693–738. [PubMed: 33007077]

10. Aujesky D, Long JA, Fine MJ,Ibrahim SA. African American race was associated with an increased risk of complications following venous thromboembolism. J Clin Epidemiol. 2007;60:410–6. [PubMed: 17346616]

11. Baglin T, Bauer K, Douketis J, Buller H, Srivastava A,Johnson G. Duration of anticoagulant therapy after a first episode of an unprovoked pulmonary embolus or deep vein thrombosis: guidance from the SSC of the ISTH. J Thromb Haemost. 2012;10:698–702. [PubMed: 22332937]

12. Barnes GD, Muzikansky A, Cameron S, Giri J, Heresi GA, Jaber W, Wood T, Todoran TM, Courtney DM, Tapson V,Kabrhel C. Comparison of 4 Acute Pulmonary Embolism Mortality Risk Scores in Patients Evaluated by Pulmonary Embolism Response Teams. JAMA Netw Open. 2020;3:e2010779. [PubMed: 32845326]

13. Cushman M, Barnes GD, Creager MA, Diaz JA, Henke PK, Machlus KR, Nieman MT,Wolberg AS. Venous Thromboembolism Research Priorities: A Scientific Statement From the American Heart Association and the International Society on Thrombosis and Haemostasis. Circulation. 2020;142:e85–e94. [PubMed: 32776842]

14. Bikdeli B, Jimenez D, Hawkins M, Ortiz S, Prandoni P, Brenner B, Decousus H, Masoudi FA, Trujillo-Santos J, Krumholz HM, Monreal M,Investigators R. Rationale, Design and Methodology of the Computerized Registry of Patients with Venous Thromboembolism (RIETE). Thromb Haemost. 2018;118:214–24. [PubMed: 29304541]

15. Weitz JI, Haas S, Ageno W, Angchaisuksiri P, Bounameaux H, Nielsen JD, Goldhaber SZ, Goto S, Kayani G, Mantovani L, Prandoni P, Schellong S, Turpie AG,Kakkar AK. Global Anticoagulant Registry in the Field - Venous Thromboembolism (GARFIELD-VTE). Rationale and design. Thromb Haemost. 2016;116:1172–79. [PubMed: 27656711]

16. Wiener RS, Schwartz LM,Woloshin S. Time trends in pulmonary embolism in the United States: evidence of overdiagnosis. Arch Intern Med. 2011;171:831–7. [PubMed: 21555660]

17. Stein PD, Beemath A,Olson RE. Trends in the incidence of pulmonary embolism and deep venous thrombosis in hospitalized patients. Am J Cardiol. 2005;95:1525–6. [PubMed: 15950590]

18. Bikdeli B, Wang Y, Jimenez D, Parikh SA, Monreal M, Goldhaber SZ,Krumholz HM. Pulmonary Embolism Hospitalization, Readmission, and Mortality Rates in US Older Adults, 1999–2015. JAMA. 2019;322:574–76. [PubMed: 31408124]

19. Barco S, Valerio L, Gallo A, Turatti G, Mahmoudpour SH, Ageno W, Castellucci LA, Cesarman-Maus G, Ddungu H, De Paula EV, Dumantepe M, Goldhaber SZ, Guillermo Esposito MC, Klok FA, Kucher N, McLintock C, F NÁ, Simioni P, Spirk D, Spyropoulos AC, Urano T, Zhai ZG, Hunt BJ,Konstantinides SV. Global reporting of pulmonary embolism-related deaths in the World Health Organization mortality database: Vital registration data from 123 countries. Res Pract Thromb Haemost. 2021;5:e12520. [PubMed: 34263098]

20. Lehnert P, Lange T, Møller CH, Olsen PS,Carlsen J. Acute Pulmonary Embolism in a National Danish Cohort: Increasing Incidence and Decreasing Mortality. Thromb Haemost. 2018;118:539–46. [PubMed: 29536465]

21. Patient Safety Indicator 12 (PSI 12) Perioperative Pulmonary Embolism or Deep Vein Thrombosis Rate. Agency for Healthcare Research and Quality, 2016. Accessible at: Agency for Healthcare Research and Quality. Available at: https://www.qualityindicators.ahrq.gov/Downloads/Modules/PSI/V60-ICD10/TechSpecs/PSI_12_Perioperative_Pulmonary_Embolism_or_Deep_Vein_Thrombosis_Rate.pdf Date last accessed: OCtober 25, 2021.

22. Spyropoulos AC, Ashton V, Chen YW, Wu B,Peterson ED. Rivaroxaban versus warfarin treatment among morbidly obese patients with venous thromboembolism: Comparative effectiveness, safety, and costs. Thromb Res. 2019;182:159–66. [PubMed: 31493618]

23. Guo JD, Hlavacek P, Rosenblatt L, Keshishian A, Russ C, Mardekian J, Ferri M, Poretta T, Yuce H,McBane R. Safety and effectiveness of apixaban compared with warfarin among clinically-relevant subgroups of venous thromboembolism patients in the United States Medicare population. Thromb Res. 2020;198:163–70. [PubMed: 33348190]

24. Marquis-Gravel G, Roe MT, Robertson HR, Harrington RA, Pencina MJ, Berdan LG, Hammill BG, Faulkner M, Muñoz D, Fonarow GC, Nallamothu BK, Fintel DJ, Ford DE, Zhou L, Daugherty SE, Nauman E, Kraschnewski J, Ahmad FS, Benziger CP, Haynes K, Merritt JG, Metkus T, Kripalani S, Gupta K, Shah RC, McClay JC, Re RN, Geary C, Lampert BC, Bradley SM, Jain SK, Seifein H, Whittle J, Roger VL, Effron MB, Alvarado G, Goldberg YH, VanWormer JL, Girotra S, Farrehi P, McTigue KM, Rothman R, Hernandez AF,Jones WS. Rationale and Design of the Aspirin Dosing-A Patient-Centric Trial Assessing Benefits and Long-term Effectiveness (ADAPTABLE) Trial. JAMA Cardiol. 2020;5:598–607. [PubMed: 32186653]

25. Chen PH. Essential Elements of Natural Language Processing: What the Radiologist Should Know. Acad Radiol. 2020;27:6–12. [PubMed: 31537505]

26. Wu S, Roberts K, Datta S, Du J, Ji Z, Si Y, Soni S, Wang Q, Wei Q, Xiang Y, Zhao B,Xu H. Deep learning in clinical natural language processing: a methodical review. J Am Med Inform Assoc. 2020;27:457–70. [PubMed: 31794016]

27. Pham AD, Névéol A, Lavergne T, Yasunaga D, Clément O, Meyer G, Morello R,Burgun A. Natural language processing of radiology reports for the detection of thromboembolic diseases and clinically relevant incidental findings. BMC Bioinformatics. 2014;15:266. [PubMed: 25099227]

28. Raja AS, Ip IK, Prevedello LM, Sodickson AD, Farkas C, Zane RD, Hanson R, Goldhaber SZ, Gill RR,Khorasani R. Effect of computerized clinical decision support on the use and yield of CT pulmonary angiography in the emergency department. Radiology. 2012;262:468–74. [PubMed: 22187633]

29. Tian Z, Sun S, Eguale T,Rochefort CM. Automated Extraction of VTE Events From Narrative Radiology Reports in Electronic Health Records: A Validation Study. Med Care. 2017;55:e73–e80. [PubMed: 25924079]

30. Selby LV, Narain WR, Russo A, Strong VE,Stetson P. Autonomous detection, grading, and reporting of postoperative complications using natural language processing. Surgery. 2018;164:1300–05. [PubMed: 30056994]

31. Chen MC, Ball RL, Yang L, Moradzadeh N, Chapman BE, Larson DB, Langlotz CP, Amrhein TJ,Lungren MP. Deep Learning to Classify Radiology Free-Text Reports. Radiology. 2018;286:845–52. [PubMed: 29135365]

32. Ascent of machine learning in medicine. Nat Mater. 2019;18:407. [PubMed: 31000807]

33. Burles K, Innes G, Senior K, Lang E,McRae A. Limitations of pulmonary embolism ICD-10 codes in emergency department administrative data: let the buyer beware. BMC Med Res Methodol. 2017;17:89. [PubMed: 28595574]

34. Casez P, Labarere J, Sevestre MA, Haddouche M, Courtois X, Mercier S, Lewandowski E, Fauconnier J, Francois P,Bosson JL. ICD-10 hospital discharge diagnosis codes were sensitive for identifying pulmonary embolism but not deep vein thrombosis. J Clin Epidemiol. 2010;63:790–7. [PubMed: 19959332]

35. Alotaibi GS, Wu C, Senthilselvan A,McMurtry MS. The validity of ICD codes coupled with imaging procedure codes for identifying acute venous thromboembolism using administrative data. Vasc Med. 2015;20:364–8. [PubMed: 25834115]

36. Lawrence K, Joos C, Jones AE, Johnson SA,Witt DM. Assessing the accuracy of ICD-10 codes for identifying acute thromboembolic events among patients receiving anticoagulation therapy. J Thromb Thrombolysis. 2019;48:181–86. [PubMed: 31124033]

37. Prat M, Derumeaux H, Sailler L, Lapeyre-Mestre M,Moulis G. Positive predictive values of peripheral arterial and venous thrombosis codes in French hospital database. Fundam Clin Pharmacol. 2018;32:108–13. [PubMed: 29055145]

38. Johnson SA, Signor EA, Lappe KL, Shi J, Jenkins SL, Wikstrom SW, Kroencke RD, Hallowell D, Jones AE,Witt DM. A comparison of natural language processing to ICD-10 codes for identification and characterization of pulmonary embolism. Thromb Res. 2021;203:190–95. [PubMed: 34044246]

39. Verma AA, Masoom H, Pou-Prom C, Shin S, Guerzhoy M, Fralick M, Mamdani M,Razak F. Developing and validating natural language processing algorithms for radiology reports compared to ICD-10 codes for identifying venous thromboembolism in hospitalized medical patients. Thromb Res. 2022;209:51–58. [PubMed: 34871982]

40. Andersson T, Isaksson A, Khalil H, Lapidus L, Carlberg B,Söderberg S. Validation of the Swedish National Inpatient Register for the diagnosis of pulmonary embolism in 2005. Pulm Circ. 2022;12:e12037. [PubMed: 35506065]

41. Kreimeyer K, Foster M, Pandey A, Arya N, Halford G, Jones SF, Forshee R, Walderhaug M,Botsis T. Natural language processing systems for capturing and standardizing unstructured clinical information: A systematic review. J Biomed Inform. 2017;73:14–29. [PubMed: 28729030]

42. Wong A, Plasek JM, Montecalvo SP,Zhou L. Natural Language Processing and Its Implications for the Future of Medication Safety: A Narrative Review of Recent Advances and Challenges. Pharmacotherapy. 2018;38:822–41. [PubMed: 29884988]

43. Zeng Z, Deng Y, Li X, Naumann T,Luo Y. Natural Language Processing for EHR-Based Computational Phenotyping. IEEE/ACM Trans Comput Biol Bioinform. 2019;16:139–53. [PubMed: 29994486]

44. Virani SS, Alonso A, Benjamin EJ, Bittencourt MS, Callaway CW, Carson AP, Chamberlain AM, Chang AR, Cheng S, Delling FN, Djousse L, Elkind MSV, Ferguson JF, Fornage M, Khan SS, Kissela BM, Knutson KL, Kwan TW, Lackland DT, Lewis TT, Lichtman JH, Longenecker CT, Loop MS, Lutsey PL, Martin SS, Matsushita K, Moran AE, Mussolino ME, Perak AM, Rosamond WD, Roth GA, Sampson UKA, Satou GM, Schroeder EB, Shah SH, Shay CM, Spartano NL, Stokes A, Tirschwell DL, VanWagner LB,Tsao CW. Heart Disease and Stroke Statistics-2020 Update: A Report From the American Heart Association. Circulation. 2020;141:e139–e596. [PubMed: 31992061]

45. Minges KE, Bikdeli B, Wang Y, Attaran RR,Krumholz HM. National and Regional Trends in Deep Vein Thrombosis Hospitalization Rates, Discharge Disposition, and Outcomes for Medicare Beneficiaries (), (). Am J Med. 2018.

46. Klok FA, Kruip M, van der Meer NJM, Arbous MS, Gommers D, Kant KM, Kaptein FHJ, van Paassen J, Stals MAM, Huisman MV,Endeman H. Incidence of thrombotic complications in critically ill ICU patients with COVID-19. Thromb Res. 2020;191:145–47. [PubMed: 32291094]

47. Bikdeli B, Madhavan MV, Jimenez D, Chuich T, Dreyfus I, Driggin E, Nigoghossian C, Ageno W, Madjid M, Guo Y, Tang LV, Hu Y, Giri J, Cushman M, Quéré I, Dimakakos EP, Gibson CM, Lippi G, Favaloro EJ, Fareed J, Caprini JA, Tafur AJ, Burton JR, Francese DP, Wang EY, Falanga A, McLintock C, Hunt BJ, Spyropoulos AC, Barnes GD, Eikelboom JW, Weinberg I, Schulman S, Carrier M, Piazza G, Beckman JA, Steg PG, Stone GW, Rosenkranz S, Goldhaber SZ, Parikh SA, Monreal M, Krumholz HM, Konstantinides SV, Weitz JI,Lip GYH. COVID-19 and Thrombotic or Thromboembolic Disease: Implications for Prevention, Antithrombotic Therapy, and Follow-Up: JACC State-of-the-Art Review. J Am Coll Cardiol. 2020;75:2950–73. [PubMed: 32311448]

48. Bikdeli B Anticoagulation in COVID-19: Randomized trials should set the balance between excitement and evidence. Thromb Res. 2020;196:638–40. [PubMed: 33066998]

49. Nopp S, Janata-Schwatczek K, Prosch H, Shulym I, Königsbrügge O, Pabinger I,Ay C. Pulmonary embolism during the COVID-19 pandemic: decline in diagnostic procedures and incidence at a University Hospital.

50. Bikdeli B, Carrier M,Bates SM. Subsegmental pulmonary embolism: May not be a killer but indicates significant risk. Thromb Res. 2020;185:180–82. [PubMed: 31796210]

51. Baumgartner C, Go AS, Fan D, Sung SH, Witt DM, Schmelzer JR, Williams MS, Yale SH, VanWormer JJ,Fang MC. Administrative codes inaccurately identify recurrent venous thromboembolism: The CVRN VTE study. Thromb Res. 2020;189:112–18. [PubMed: 32199174]

52. Hutchinson BD, Navin P, Marom EM, Truong MT,Bruzzi JF. Overdiagnosis of Pulmonary Embolism by Pulmonary CT Angiography. AJR Am J Roentgenol. 2015;205:271–7. [PubMed: 26204274]

53. Miller WT Jr., Marinari LA, Barbosa E Jr., Litt HI, Schmitt JE, Mahne A, Lee V, Akers SR. Small pulmonary artery defects are not reliable indicators of pulmonary embolism. Ann Am Thorac Soc. 2015;12:1022–9. [PubMed: 25961445]

54. Tapson VF, Platt DM, Xia F, Teal SA, de la Orden M, Divers CH, Satler CA, Joish VN,Channick RN. Monitoring for Pulmonary Hypertension Following Pulmonary Embolism: The INFORM Study. Am J Med. 2016;129:978–85 e2. [PubMed: 27046247]

55. Vinson DR, Drenten CE, Huang J, Morley JE, Anderson ML, Reed ME, Nishijima DK, Liu V, Kaiser Permanente Clinical Research on Emergency S,Treatment N. Impact of relative contraindications to home management in emergency department patients with low-risk pulmonary embolism. Ann Am Thorac Soc. 2015;12:666–73. [PubMed: 25695933]

56. Jung RG, Simard T, Hibbert B, Harris AH, Hohmann SF, Giri JS, Bashir R,Alkhouli M. Association of annual volume and in-hospital outcomes of catheter-directed thrombolysis for pulmonary embolism. Catheter Cardiovasc Interv. 2022;99:440–46. [PubMed: 35083846]

57. Elbadawi A, Mahtta D, Elgendy IY, Saad M, Krittanawong C, Hira RS, Omer M, Ogunbayo GO, Garratt K, Rao SV,Jneid H. Trends and Outcomes of Fibrinolytic Therapy for STEMI: Insights and Reflections in the COVID-19 Era. JACC Cardiovasc Interv. 2020;13:2312–14. [PubMed: 33032721]

58. Otite FO, Saini V, Sur NB, Patel S, Sharma R, Akano EO, Anikpezie N, Albright K, Schmidt E, Hoffman H, Gould G, Khandelwal P, Latorre JG, Malik AM, Sacco RL,Chaturvedi S. Ten-Year Trend in Age, Sex, and Racial Disparity in tPA (Alteplase) and Thrombectomy Use Following Stroke in the United States. Stroke. 2021;52:2562–70. [PubMed: 34078107]

59. Guez D, Hansberry DR, Eschelman DJ, Gonsalves CF, Parker L, Rao VM,Levin DC. Inferior Vena Cava Filter Placement and Retrieval Rates among Radiologists and Nonradiologists. J Vasc Interv Radiol. 2018;29:482–85. [PubMed: 29305114]

60. Gayou EL, Makary MS, Hughes DR, Hemingway J, Elliott ED, Spain JW,Prevedello LM. Nationwide Trends in Use of Catheter-Directed Therapy for Treatment of Pulmonary Embolism in Medicare Beneficiaries from 2004 to 2016. J Vasc Interv Radiol. 2019;30:801–06. [PubMed: 31040058]

61. Pasrija C, Kronfli A, Rouse M, Raithel M, Bittle GJ, Pousatis S, Ghoreishi M, Gammie JS, Griffith BP, Sanchez PG,Kon ZN. Outcomes after surgical pulmonary embolectomy for acute submassive and massive pulmonary embolism: A single-center experience. J Thorac Cardiovasc Surg. 2018;155:1095–106 e2. [PubMed: 29452460]

62. 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 1:154–62. [PubMed: 22262602]

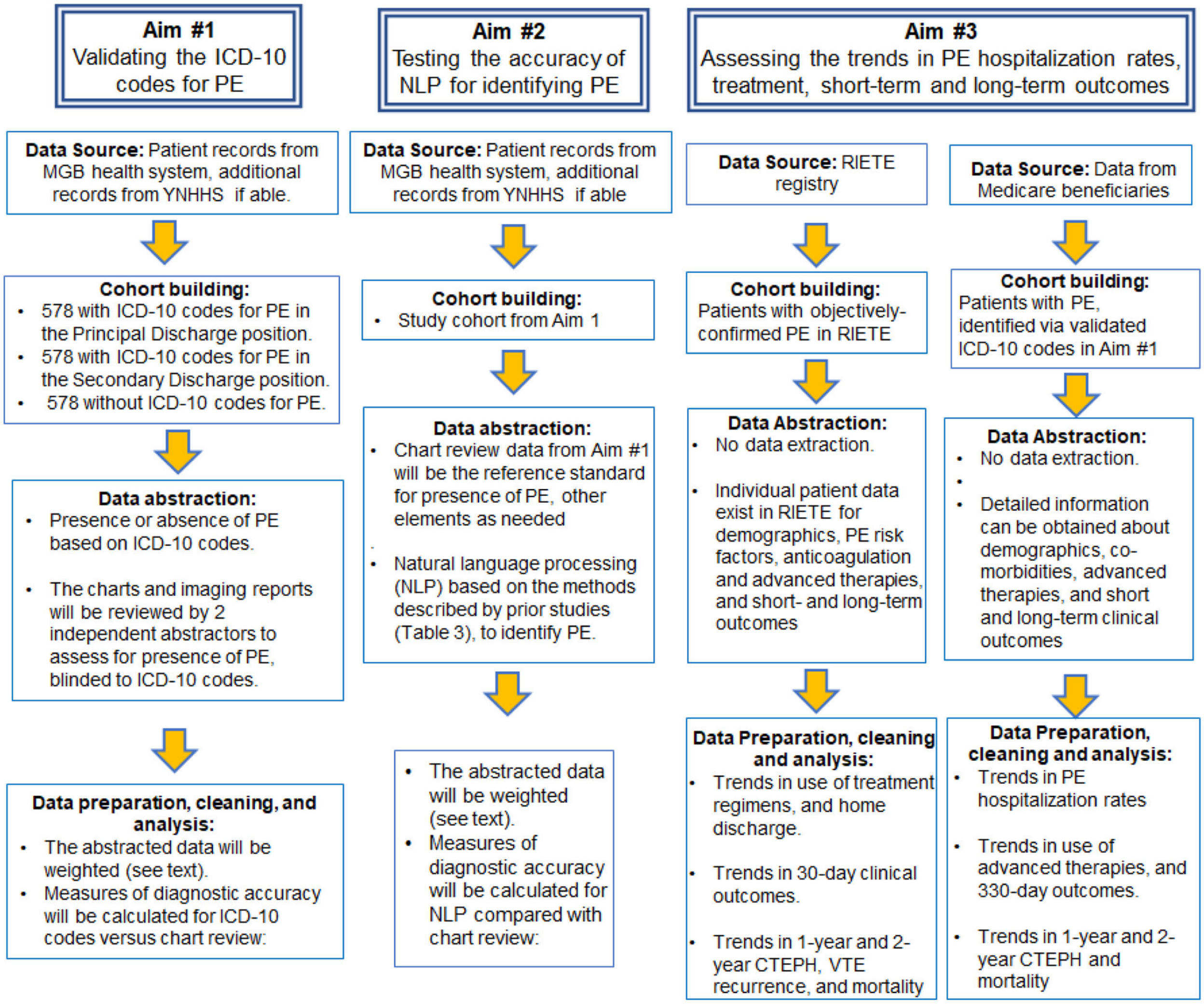
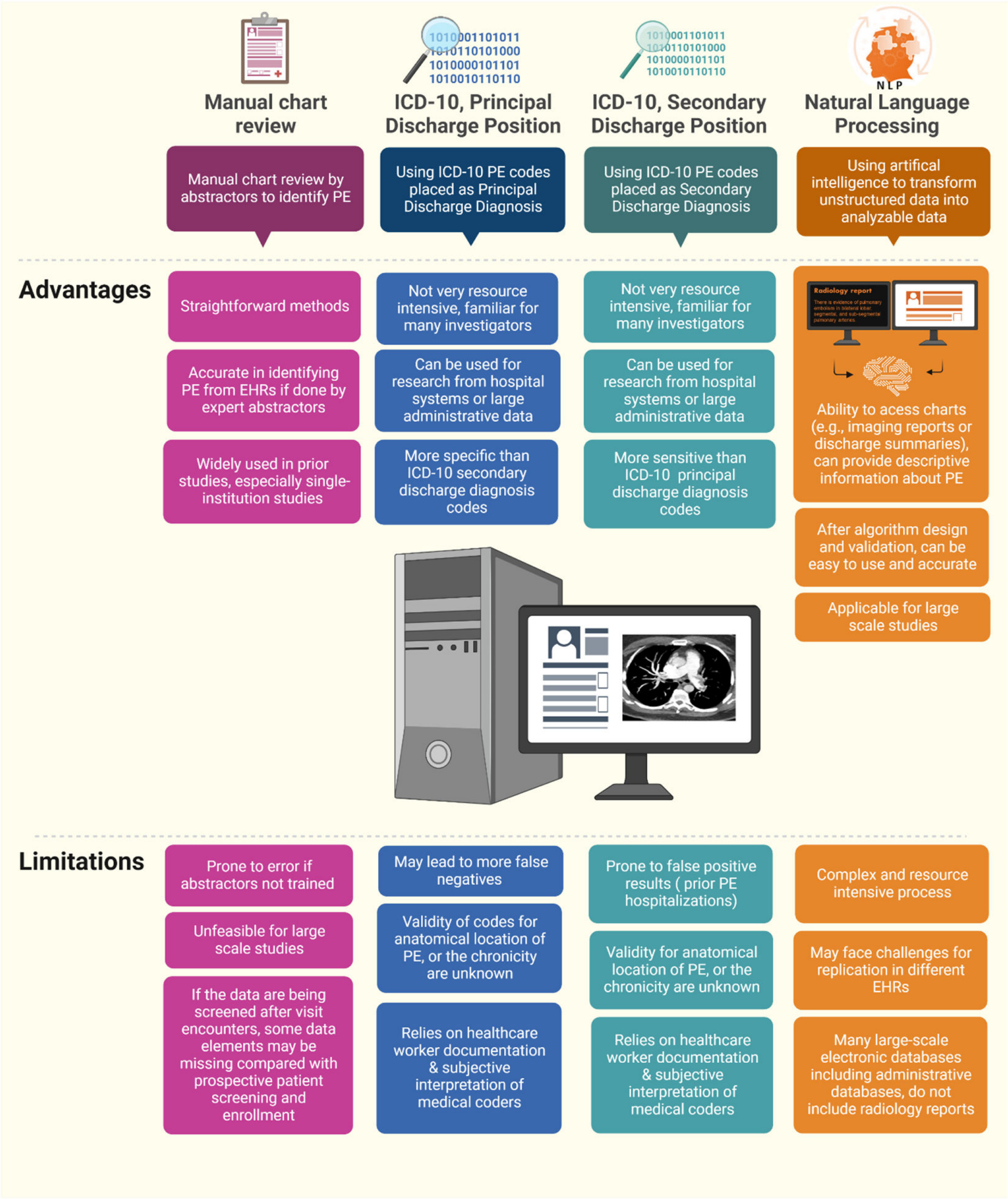

Aim #1
Validating the ICD-10 codes for PE
Aim #2
Testing the accuracy of NLP for identifying PE
Aim #3
Assessing the trends in PE hospitalization rates, treatment, short-term and long-term outcomes
Data Source: Patient records from MGB health system, additional records from YNHHS if able.
Cohort building:
• 578 with ICD-10 codes for PE in the Principal Discharge position.
• 578 with ICD-10 codes for PE in the Secondary Discharge position.
• 578 without ICD-10 codes for PE.
Data abstraction:
• Presence or absence of PE based on ICD-10 codes.
• The charts and imaging reports will be reviewed by 2 independent abstractors to assess for presence of PE, blinded to ICD-10 codes.
Data preparation, cleaning, and analysis:
• The abstracted data will be weighted (see text).
• Measures of diagnostic accuracy will be calculated for ICD-10 codes versus chart review:
Data Source: Patient records from MGB health system, additional records from YNHHS if able
Cohort building:
• Study cohort from Aim 1
Data abstraction:
• Chart review data from Aim #1 will be the reference standard for presence of PE, other elements as needed
• Natural language processing (NLP) based on the methods described by prior studies (Table 3), to identify PE.
• The abstracted data will be weighted (see text).
• Measures of diagnostic accuracy will be calculated for NLP compared with chart review:
Data Source: RIETE registry
Cohort building:
Patients with objectively-confirmed PE in RIETE
Data Abstraction:
• No data extraction.
• Individual patient data exist in RIETE for demographics, PE risk factors, anticoagulation and advanced therapies, and short- and long-term outcomes
Data Preparation, cleaning and analysis:
• Trends in use of treatment regimens, and home discharge.
• Trends in 30-day clinical outcomes.
• Trends in 1-year and 2-year CTEPH, VTE recurrence, and mortality
Data Source: Data from Medicare beneficiaries
Cohort building:
Patients with PE, identified via validated ICD-10 codes in Aim #1
Data Abstraction:
• No data extraction.
•
• Detailed information can be obtained about demographics, co-morbidities, advanced therapies, and short and long-term clinical outcomes
Data Preparation, cleaning and analysis:
• Trends in PE hospitalization rates
• Trends in use of advanced therapies, and 330-day outcomes.
• Trends in 1-year and 2-year CTEPH and mortality

Figure 1.

Graphical Summary of the Goals of the PE-EHR+ Study



	Manual chart review	ICD-10, Principal Discharge Position	ICD-10, Secondary Discharge Position	Natural Language Processing
	Manual chart review by abstractors to identify PE	Using ICD-10 PE codes placed as Principal Discharge Diagnosis	Using ICD-10 PE codes placed as Secondary Discharge Diagnosis	Using artifical intelligence to transform unstructured data into analyzable data
Advantages	Straightforward methods	Not very resource intensive, familiar for many investigators	Not very resource intensive, familiar for many investigators	Ability to acess charts (e.g., imaging reports or discharge summaries), can provide descriptive information about PE
	Accurate in identifying PE from EHRs if done by expert abstractors	Can be used for research from hospital systems or large administrative data	Can be used for research from hospital systems or large administrative data	After algorithm design and validation, can be easy to use and accurate
	Widely used in prior studies, especially single-institution studies	More specific than ICD-10 secondary discharge diagnosis codes	More sensitive than ICD-10 principal discharge diagnosis codes	Applicable for large scale studies
Limitations	Prone to error if abstractors not trained	May lead to more false negatives	Prone to false positive results (prior PE hospitalizations)	Complex and resource intensive process
	Unfeasible for large scale studies	Validity of codes for anatomical location of PE, or the chronicity are unknown	Validity for anatomical location of PE, or the chronicity are unknown	May face challenges for replication in different EHRs
	If the data are being screened after visit encounters, some data elements may be missing compared with prospective patient screening and enrollment	Relies on healthcare worker documentation & subjective interpretation of medical coders	Relies on healthcare worker documentation & subjective interpretation of medical coders	Many large-scale electronic databases including administrative databases, do not include radiology reports

Figure 2.

Methods for Identification of Patients with Pulmonary Embolism in Electronic Databases and Their Tradeoffs.

Table 1.

ICD-10 Codes for Pulmonary Embolism*

ICD-10 Codes	Definition
I26	Pulmonary embolism
I26.0	Pulmonary embolism with acute cor pulmonale
I26.02	Saddle embolus of pulmonary artery with acute cor pulmonale
I26.09	Other pulmonary embolism with acute cor pulmonale
I26.9	Pulmonary embolism without acute cor pulmonale
I26.92	Saddle embolus of pulmonary artery without acute cor pulmonale
I26.93	Single subsegmental pulmonary embolism without acute cor pulmonale†
I26.94	Multiple subsegmental pulmonary emboli without acute cor pulmonale†
I26.99	Other pulmonary embolism without acute cor pulmonale
O88.2	Obstetric thromboembolism
Z86.711	Personal history of pulmonary embolism

*Note that the codes can be placed in the discharge records as a Principal Discharge Diagnosis or Secondary Discharge Diagnosis, and that for research studies, either or both these locations can be queried, with tradeoffs between sensitivity and specificity. These issues will be investigated in depth in the PE-EHR+ study. Cases of amniotic fluid embolism or fat embolism, if identified by the PE codes, will be flagged. Although the code I82 and its sub-categories denote venous embolism and thrombosis, the subcodes are mostly related to deep vein thrombosis and were not included in the current study. If false negatives are identified in PE-EHR+, we will assess if a subset of them includes this code.

†Subsegmental PE is a challenging diagnosis.[50] Independent validation of the diagnosis in this subset will be attempted if the resources allow.

Table 2.

Existing studies that assessed the accuracy of ICD-10 codes for PE *

Study	ICD-10 codes assessed	Metrics assessed	Summary of Findings	Comments
Burles et al.[33]	I26.0 I26.9	Sensitivity Specificity PPV NPV	Using data from 4 emergency departments in Alberta, CA, the authors reported the accuracy of codes for detecting PE against chart review. Sensitivity was 91.1%, specificity was 99.9%, PPV was 82.3%, and NPV was 99.9%. No distinction was made between primary vs secondary codes.	Among 479,937 visits, 1,453 patients with PE codes we found. The authors ran keyword search of the physician discharge diagnosis field among patients without PE codes to identify false negatives.
Casez et al.[34]	I26.0 I26.9 O88.2	Sensitivity	Among 1375 patients with suspected DVT/PE, ICD-10 codes were compared with diagnosis based on imaging studies. Sensitivity for PE was 88.9%. Specificity could not be assessed.	The authors assessed codes placed in Principal or secondary discharge position. Sufficient details about the breakdown were not provided.
Alotaibi et al.[35]	I26.0 I26.9	Sensitivity Specificity PPV NPV	The authors sampled 1361 patients with probable VTE: 147 had a PE and 105 had a DVT. Predefined ICD codes were applied to the 1361 patients to see who were coded correctly and who should not have been coded. Sensitivity for PE was 74.83%, specificity was 95.77%, PPV was 70.51%, and NPV was 93.35%.	Study from emergency departments in Canada. The ICD-10 PE codes were used in any position. Sufficient details about the breakdown were not provided.
Lawrence et al.[36]	I26.02 I26.09 I26.92 I26.99	Sensitivity Specificity PPV NPV	Charts of 487 patients receiving anticoagulation in a single institution were reviewed. For ICD-10 PEs, sensitivity was 100%, specificity was 79.3%, PPV was 17.1%, and NPV was 100%.	The authors assessed codes placed in Principal or secondary discharge position. Sufficient details about the breakdown were not provided.
Prat et al.[37]	I26.0 I26.9	Sensitivity Specificity PPV	In a study of 970 patients who had a CTPA, ICD-10 codes and NLP were compared to manual review (13% of patients had PE). Sensitivity of ICD-10 codes for PE was 92.9%, specificity was 91.0% and PPV was 60.6%.	Compared NLP to ICD-10 codes. Compared NLP and ICD-10 codes for saddle PE and for subsegmental PE.
Johnson et al.[38]	I26, I26.01, I26.02,I26.09, I26.0, I26.90, I26.92,I26.93, I26.94,I26.99, I26.9, I27.24, I27.82,Z86.711	Sensitivity Specificity PPV NPV	In a study of 1000 random hospitalizations, NLP algorithms, and ICD-10 codes were compared to manual review. Sensitivity of ICD-10 codes for PE was 63%, specificity was 99%, PPV was 70%, and NPV was 99%.	The authors assessed ICD-10 codes in any position and did not assess the codes in Principal Discharge position, separately. NLP tools were also assessed in this study. See Table 3.
Verma et al.[39]	I26, O88.2	Sensitivity Specificity PPV NPV	In a study from 5 hospitals in Canada, the authors reported the accuracy of an NLP algorithm that they developed, compared with simpleNLP and ICD-10 codes. For PE, they reported sensitivity of 57%, specificity of 1, PPV of 0.92 and NPV of 0.99.	The study also assessed accuracy of codes and NLP for DVT. However, detailed information about cohort breakdown for PE was not provided. Information not available for location of codes.
Andersson et al.[40]	I26.0–I26.9	PPV	In a study of 559 patients with ICD-10 codes for PE from Sweden, chart review confirmed acute PE in 435 patients (PPV 78.9%). In 11 patients the codes were completely incorrect, and in another 47, the codes indicated prior diagnosis of PE but not acute PE.	The study did not provide sufficient discrimination between primary vs secondary ICD-10 codes and did not assess sensitivity, specificity, or negative predictive values.

*Data are based on a systematic search and review of the literature. See supplementary material for the search query. CTPA: Computed tomography pulmonary angiography, DVT: Deep vein thrombosis, NPV: Negative predictive value, PE: Pulmonary embolism, PPV: Positive predictive value,

Table 3.

Natural language processing (NLP) algorithms used for assessment of PE in prior studies

Study	NLP Method used	NLP performance metrics	NLP Technique and Methods Summary	Comments
Pham et al.[27]	Generate ML features by using N-gram and manual annotation with Brat.	Precision Recall F-measure	CT angiography reports from 573 patients in a single French institution were used. An NLP algorithm was designed, trained with 100 reports, and tested in the remaining reports. There was 99% precision for PE. Details about positive predictive value and sensitivity were not mentioned.	The study was from France. Applicability to charts in English is uncertain.
Raja et al.[28]	General Architecture for Text Engineering	Sensitivity Specificity PPV NPV	General Architecture for Text Engineering (GATE) tool was applied to 179 CT angiography reports to identify PE, and compared against manual review. Sensitivity and positive predictive value of the NLP algorithm were, both, 91.3%. Specificity and NPV were, both, 98.7%.	Sample size was fairly small.
Tian et al.[29]	Symbolic NLP classifiers	Sensitivity Specificity PPV	Using the imaging reports in a Canadian health system, the authors derived and validated an NLP algorithm for PE against manual review of the radiology reports. NLP achieved 94% sensitivity and 80% positive predictive value for PE and 96% specificity.	
Selby et al.[30]	Bag of words, N-gram	Sensitivity Specificity PPV NPV	In a study using radiology reports and the WEKA machine learning toolkit, an NLP tool for detection of post-operative PE was developed. Among 703 patients in the validation set, sensitivity for PE was 90%, specificity was 98.7%, PPV was 81..8%, and NPV was 99.3%.	The study focused on post-operative PE.
Chen et al.[31]	Convolutional Neural Network (CNN)	Sensitivity Specificity Accuracy	In a single-center study, convolutional neural network with unsupervised learning using TensorFlow (a deep learning library) and an NLP algorithm (PeFinder) were compared against imaging reports. TensorFlow had a sensitivity of 95.2%, specificity of 90.5%, and accuracy of 92.1%. PeFinder had a sensitivity of 94.5%, a speicificty of 92.9%, and accuracy of 93.5%.	Positive predictive values were not reported.
Johnson et al[38]	Rule-based NLP	Sensitivity Specificity PPV NPV	In a study of 1000 random hospitalizations, NLP algorithms, “simpleNLP“ tool, and ICD-10 codes were compared to manual review. Sensitivity of NLP was 96.0% and specificity was 97.7%. Positive and negative predictive values were 86.3% and 99.4%, respectively.	ICD-10 codes were also assessed in this study. See Table 2. The authors identified better discrimination for saddle PE and for sub-segmental PE with NLP, compared with ICD-10 codes.
Verma et al.[39]	Rule-based NLP	PPV?	In a study from 5 hospitals in Canada, the authors reported the accuracy of an NLP algorithm that they developed, compared with simpleNLP and ICD-10 codes.	The study also assessed accuracy of codes and NLP for DVT. However, detailed information about cohort breakdown for PE was not provided. ICD-10 codes were also assessed in this study. See Table 2.

CT: Computed tomography. Other abbreviations as in Table 2. See Table 2 for the study by Johnson et al.[38]

Table 4.

Operational definitions for the assessment of the accuracy of ICD-10 codes for PE, subsegmental PE, and cor pulmonale according to chart review[*]

Condition by the ICD-10 codes	Definition according to chart review	Comment
PE [†]	Mentioning of PE in medical notes such as discharge summary, verified by sufficient confirmatory findings for PE in radiology reports from the index hospitalization (such as reports for filling defect in CTPA, high-probability V/Q scan, direct verification of pulmonary thrombi/emboli in invasive angiography, or presence of new proximal DVT in conjunction with symptoms and signs of PE).	The abstractors will be blinded to the ICD-10 code results.
Subsegmental PE [†]	Report of sub-segmental filling defects consistent with the diagnosis of PE in radiology reports, without involvement of segmental, lobar, or central pulmonary arteries.	A sub-component of the PE-HER study plans to assess 50 CTPA studies with an initial radiology report for sub-segmental PE by a core laboratory.
Acute cor pulmonale[†§]**in the setting of PE**	Evidence of newly-identified RV dysfunction evidenced by at least one of the following: • Radiology report indicating RV/LV ratio≥1.0[¶], or enlarged RV, or bowing of the interventricular septum, or the term "RV strain", or a combination of these. • Echocardiographic report indicating RV/LV ratio≥0.7[¶], or enlarged RV, or bowing of the interventricular septum, or the term "RV strain", or TAPSE<16, or RV free wall hypokinesis, or the term McConnell sign, or newly identified elevated RVSP (>30mmHg) without another cause, or a combination of these. • Elevation of cardiac troponins above the normal assay values.[β]	Several of the ICD-10 codes refer to cor pulmonale. However, major expert guidelines do not use this terminology, and there is no universal definition for the term exists. In the PE-EHR we considered acute cor pulmonale if there was evidence of newly identified RV dysfunction.

[*] The main goal of this study is not to re-adjudicate the initially identified events during routine clinical care, but rather to assess the success of ICD-10 codes to accurately capture the information related to PE as occurred in the index routine care hospitalization. Therefore, routine core laboratory assessment of individual imaging studies is not considered. For a subset of patient, core laboratory assessment may be considered as a supplemental goal of the project. See text for details.

[†] If patients are transferred from other facilities and there is no existing report for their original CTPA or V/Q scan, the study Principal Investigator will attempt to verify the diagnosis of PE from the original imaging studies. However, further attempt assessment for subsegmental PE or acute cor pulmonale will not be made to keep the assessment criteria uniform.

[¶] Different cutoffs have been used for CTPA assessment and echocardiographic assessment of RV/VL ratio. A higher threshold is associated with higher specificity for identification of RV dysfunction as a prognosticator of adverse clinical outcomes. In echocardiographic assessment, RV/LV ratios >0.6 have been assessed in some studies. Since in PE-EHR+ there is no a priori plan to independently re-measure the values –but rather to rely on reports of CTPA and echocardiography, to facilitate the process, the abstractors will be advised to look for an RV/LV ratio cutoff >0.9 in the CTPA or echocardiographic reports.

[§] Since $S_1Q_3T_3$ pattern is nonspecific, it was not considered.

[β] For patients with estimated creatinine clearance <60mL/min, troponin levels may be chronically elevated. At least a 20% elevation than the prior recorded troponin would be required. Fifth generation (high-sensitivity) troponin assays detect very modest elevations in troponin. However, the clinical significance of very modest elevations in troponin (undetected by fourth generation assays) in patients with PE remains uncertain. By consensus among coauthors (BB, DJ, GP), high-sensitivity troponin values beyond 30 ng/L not explained by another cause will be considered positive in the PE-EHR+ study.

CTPA: Computed tomography pulmonary angiography, ICD-10: International Classification of Diseases, 10th revision, PE: Pulmonary embolism, RV: Right ventricular, RVSP: Right ventricular systolic pressure, V/Q scan: Ventilation/perfusion scan.

Table 5.

Basic definitions related to natural language processing as they relate to identification of pulmonary embolism in medical charts

Concept	Definition
Corpus	The unstructured large body of text. Examples include medical notes or imaging reports.
Tokenization	A token represents linguistic units, including single words and spaces. Tokens can be combined to form larger units including phrases. Examples include pulmonary, and embolism.
Stop words	Stop words are some most common used words in the free text. They may be prepositions, pronouns, conjunction…etc. Stop words are typically removed during the data preprocessing stage of NLP since they do not frequently contribute additive information to the text. Examples include "*the*", "*is*", and "*and*".
Acronyms/ abbreviations	The same acronym may have different meanings in the chart. PE can denote pulmonary embolism, but may be used to refer to physical examination. Others may use the acronym 'PTE' to refer to pulmonary thromboembolism. However, PTE can be used to refer to pulmonary thromboendarterectomy.
Word Stemming	This process groups the tokens with similar root meanings. Examples include embolism and embolic for which the stem is 'emboli'.
Lemmatization	This process converts words to *dictionary* forms. The lemma for *better* and *best* is 'good'.
Polyesmy/ word sense disambiguation	Multiple meanings from the same word. A general example is 'cold'. It can refer to the viral illness, or cold temperature.
Lexicon	It is a collection of information about the words and the lexical categories to which they belong (noun, verb, adjective, adverb, preposition). To avoid missing the concept of pulmonary embolism in the clinical text, we will need a dictionary (i.e. Lexicon) to store the possible ways pulmonary embolism is described in the clinical text (e.g., pulmonary embolism, pulmonary emboli, pulmonary thromboembolism, filling defect in the pulmonary artery). Subsequently, the ones deemed relevant, can be programmed to be identified.
N-gram (Bigram)	To better capture the exactly terminology we are looking for, N-gram will be used to identify the contiguous sequence of N items. When the N is equal to two, we will call it as bigram. An example is to look for bigram "***pulmonary embolism***" rather than "Bilateral **pulmonary** infiltrates in the lower lobes. No evidence of paradoxical **embolism**".
Negation	Handling of negation is a common task for NLP process and is quite important in clinical notes since the negation statement is often used in the differential diagnosis process. By considering the context of a sentence, the NLP algorithm can distinguish the concept is truly existing or not in the sentence. An example is to avoid misclassification of "*No pulmonary embolism*" or "*pulmonary embolism not present*" as pulmonary embolism.

Table 6.

Outcome variables for the assessment of the accuracy of the ICD-10 codes[*]

Outcome measure	General definition	Operational definition in PE-EHR+[¶]
Sensitivity	Probability of a patient with the outcome of interest being correctly classified as having the outcome $\left(\frac{True\ positives}{True\ positives + false\ negatives}\right)$	The number of patients correctly identified as having PE according to the test (codes) divided by the entire number of patients who had PE according to manual chart review.
Specificity	Probability of a patient without the outcome of interest of being correctly identified as not the outcome $\left(\frac{True\ negatives}{True\ negatives + false\ positives}\right)$	The number of patients correctly identified as not having PE according to the test (codes) divided by the entire number of patients who did not have PE according to manual chart review.
PPV	Proportion of patients identified as having the outcome according to the test that did, in fact, have the outcome $\left(\frac{True\ positives}{True\ positives + false\ positives}\right)$	The number of patients correctly identified as having PE according to the test (codes) divided by the entire number of patients for whom the test (codes) called a PE.
NPV	Proportion of patients identified as not having the outcome of interest that did not, in fact, have the outcome $\left(\frac{True\ negatives}{True\ negatives + false\ negatives}\right)$	The number of patients correctly identified as not having PE according to the test (codes) divided by the entire number of patients for whom there was no code for PE.
Accuracy	Proportion of the total number of cases examined that were correctly identified as having or not having the outcome of interest $\left(\frac{True\ positives + true\ negatives}{All\ patients}\right)$	The number of patients correctly identified as having PE plus the number of patients correctly identified as not having PE according to the test (codes) divided by the entire pool of patients.

[*] A similar approach will be used for assessing the accuracy of NLP tools.

[¶] The main analyses will be performed on a weighted sample, in which patients with ICD-10 codes for PE and patients without ICD-10 codes for PE are weighed according to the actual frequency of the codes in the entire database. In a sensitivity analysis, we will assess the accuracy metrics only in the studied sample, without weighting.