Development of a machine learning model using electrocardiogram signals to improve acute pulmonary embolism screening

Sulaiman S Somani, M.D.¹, Hossein Honarvar, Ph.D.¹,², Sukrit Narula, Ph.D.³,⁴, Isotta Landi, Ph.D.¹,², Shawn Lee, MD⁵, Yeraz Khachatourian, M.D., M.P.H.⁶, Arsalan Rehmani, M.D.³, Andrew Kim, M.D.⁶, Jessica K De Freitas, B.S.¹,², Shelly Teng, B.S.¹, Suraj Jaladanki, B.S.¹, Arvind Kumar, B.S.¹, Adam Russak, M.D.¹,⁶, Shan P Zhao, M.D., Ph.D.¹,⁷, Robert Freeman, R.N.⁸,⁹, Matthew A Levin, M.D.⁷, Girish Nadkarni, M.D., M.P.H.¹,¹⁰, Alexander C Kagen, M.D.¹¹, Edgar Argulian, M.D., M.P.H.⁵,¹², Benjamin S Glicksberg, Ph.D.¹,²*

¹ The Hasso Plattner Institute for Digital Health at Mount Sinai, Icahn School of Medicine at Mount Sinai, New York, New York, USA
² Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, New York, USA
³ Population Health Research Institute, David Braley Cardiac, Vascular and Stroke Research Institute, Hamilton, Ontario, Canada
⁴ Department of Health Research Methods, Evidence, and Impact, McMaster University, Hamilton, Ontario, Canada
⁵ Department of Cardiology, Icahn School of Medicine at Mount Sinai, New York, New York, USA
⁶ Department of Internal Medicine, Icahn School of Medicine at Mount Sinai, New York, New York, USA
⁷ Department of Anesthesiology, Perioperative, and Pain Medicine, Icahn School of Medicine at Mount Sinai, New York, New York, USA

© The Author(s) 2021. Published by Oxford University Press on behalf of the European Society of Cardiology. This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com
Department of Population Health Science and Policy, Icahn School of Medicine at Mount Sinai, New York, New York, USA

Institute for Healthcare Delivery Science, Icahn School of Medicine at Mount Sinai, New York, New York, USA

The Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, New York, USA

Department of Radiology, Icahn School of Medicine at Mount Sinai, New York, New York, USA

Mount Sinai Heart, Icahn School of Medicine at Mount Sinai, New York, New York, USA

* Corresponding author

Prof. Benjamin S Glicksberg

Department of Genetics and Genomic Sciences

Hasso Plattner Institute for Digital Health at Mount Sinai

Icahn School of Medicine at Mount Sinai

1 Gustave L Levy Pl.

New York, NY 10029

Email: benjamin.glicksberg@mssm.edu

Phone: +1 212-731-7078
Abstract

Aims: Clinical scoring systems for pulmonary embolism (PE) screening have low specificity and contribute to CT pulmonary angiogram (CTPA) overuse. We assessed whether deep learning models using an existing and routinely collected data modality, electrocardiogram (ECG) waveforms, can increase specificity for PE detection.

Methods and Results: We create a retrospective cohort of 21,183 patients at moderate- to high-suspicion of PE and associate 23,793 CTPAs (10.0% PE-positive) with 320,746 ECGs and encounter-level clinical data (demographics, comorbidities, vital signs, and labs). We develop three machine learning models to predict PE likelihood: an ECG model using only ECG waveform data, an EHR model using tabular clinical data, and a Fusion model integrating clinical data and an embedded representation of the ECG waveform. We find that a Fusion model (area under receiver-operating characteristic [AUROC] 0.81 ± 0.01) outperforms both the ECG model (AUROC 0.59 ± 0.01) and EHR model (AUROC 0.65 ± 0.01). On a sample of 100 patients from the test set, the Fusion model also achieves greater specificity (0.18) and performance (AUROC 0.84 ± 0.01) than four commonly evaluated clinical scores: Wells’ Criteria, Revised Geneva Score, Pulmonary Embolism Rule-Out Criteria, and 4-Level Pulmonary Embolism Clinical Probability Score (AUROC 0.50-0.58, specificity 0.00-0.05). The model is superior to these scores on feature sensitivity analyses (AUROC 0.66 to 0.84) and achieves comparable performance across sex (AUROC 0.81) and racial/ethnic (AUROC 0.77 to 0.84) subgroups.

Conclusion: Synergistic deep learning of electrocardiogram waveforms with traditional clinical variables can increase the specificity of PE detection in patients at least at moderate suspicion for PE.

Keywords: pulmonary embolism; electrocardiogram; machine learning; deep learning;
Introduction

Pulmonary embolism (PE) is a life-threatening complication of venous thromboembolism with high short- and long-term cardiovascular morbidity and mortality globally\textsuperscript{1,2}. Given the lack of specificity in presenting symptoms and existing clinical decision rules, diagnostic uncertainty in PE remains high and requires clinicians to use thoracic imaging modalities, most commonly computed tomography pulmonary angiography (CTPA), to confirm diagnosis, especially in those who are at moderate to high suspicion of PE\textsuperscript{3}. Overreliance on CTPA has been documented across health systems and has important implications for resource utilization. In a multi-center analysis of medical centers in the United States, only 3.1\% of CTPA scans were positive for PE\textsuperscript{4}. More importantly, widespread CTPA confers risk to patients through large radiation doses and may be contraindicated in specific subpopulations\textsuperscript{5,6}. An effectively developed framework improving CTPA diagnostic yield for PE using routinely collected clinical information would have crucial implications for PE detection and management, increasing patient safety, and mitigating systemic inefficiencies.

Integration of different data modalities and analytic approaches may provide a means to improve prediction. For example, deep learning models leveraging non-traditional data forms (such as raw ECG waveforms) set themselves apart from other classes of models by their ability to computationally derive features without pre-specification\textsuperscript{7}. Previous work has demonstrated that such ECG models can be used to predict future arrhythmias from normal sinus rhythm ECGs, characterize right- and left-sided ventricular function, and increase detection of LV dysfunction in a primary care setting\textsuperscript{8-11}. The results of these investigations suggest that deep learning models can be used to detect subtle signals indiscernible to clinicians and effectively augment the clinical workflow.

In this paper, we provide a preliminary fusion modeling framework that integrates ECG waveform and clinical data to predict the likelihood of acute PE. Our approach differs from previous PE prediction approaches because it integrates a deep learning-based embedding representation of ECG waveforms into...
the prediction framework\textsuperscript{12-15}. We hypothesize that ECG waveforms and clinical data can be combined synergistically in a machine learning modeling framework to detect PE in those patients with at least moderate PE suspicion.
Methods

Data Sources and Study Population

We conducted a retrospective cohort study to predict the probability of a pulmonary embolism (Figure 1A) in patients at moderate- to high-suspicion of PE using data between January 1, 2003 to June 3, 2020 from five hospitals in the Mount Sinai Health System (New York City, New York) serving a diverse and urban population. Dataset curation, study design, and model development are reported using the The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) and the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) statements16,17 (Supplementary Tables 1-2). The study was approved by the Mount Sinai Institutional Review Board.

For cohort selection, we extracted all the patients who had CTPA, for whom radiologist annotated CTPA reports (report text, time of acquisition) were obtained from the Mount Sinai Data Warehouse (MSDW). Details on the patient encounter selection in our dataset is described further in the Supplement, (Supplementary Methods, Supplementary Figures 1, 6). In addition to CTPA reports, we also extracted the following data and temporally linked them to CTPA reports: ECG morphology parameters (e.g., PR-interval length, ventricular rate), physician-confirmed diagnostic reads, and raw signals or waveforms were parsed from ECG data from the General Electric (GE) MUSE data management system. Waveforms recorded at 500 Hz for 10 seconds (total of 5,000 data points) were available for all linearly independent leads: I, II, V1, V2, V3, V4, V5, and V6. Tabular EHR data (demographics, comorbidities, vitals, labs). Database codes representing unique comorbidities, vital signs, and important labs were manually consolidated across the five hospital sites. ECG metadata was used to fill in missing demographics (age, sex, race) for patient encounters missing those data from MSDW. Patients were only coded as having a comorbidity if the corresponding diagnostic codes were created before the start of the CTPA-related
encounter (Supplementary Table 3). Vital signs (heart rate, blood pressure, respiration rate, temperature, oxygen saturation) and labs (D-Dimer, Brain-natriuretic peptide (BNP), and troponin-I) were also extracted.

CTPA Report Acquisition and Labeling

The MSDW database was queried by IT personnel for all thoracic CT scans acquired between January 1, 2003 to June 3, 2020 in adult patients (age ≥ 18). We use Natural Language Processing (NLP) of semi-structured reports to determine PE status and to exclude scans ordered for an indication other than assessment of the pulmonary vasculature. Pretrained NLP models from SciSpacy, which are trained using biomedical, scientific, and clinical text\(^\text{18}\), were used to preprocess, tokenize, and stem all text in the CTPA reports. Full details on CTPA report preprocessing may be found in Supplementary Methods (page 2). To annotate CTPA reports, we used a two-stage approach predicated on two assumptions: a) that the majority of CTPA reports are negative for a PE and b) that the absence of a PE in the ‘Impression’ section is reported in a standardized format (Supplementary Figure 1). In the first stage, we identify all reports negative for a PE by matching those containing a highly specific pattern (Supplementary Figure 2). In the second stage, we used a team of four Internal Medicine residents and one medical student to annotate the presence, chronicity, and vascular location of a PE in the remaining reports. Chronic PEs without any acute or subacute changes that may be associated with the presentation were classified as PE-negative to restrict the PE-positive class to only acute PEs causing the current symptomology. Given controversy over their clinical significance, we also excluded CTPAs that only documented subsegmental PEs\(^\text{19–22}\). Further details may be found in Supplemental Methods (page 2).

Linking ECG and EHR Data with CTPA Labels

For a given patient encounter, ECGs and CTPAs were linked differently with one another depending on whether the CTPA was PE-positive or PE-negative (Figure 1C). If the CTPA was PE-positive, ECGs recorded within 24 hours of the CTPA were retained and labeled as PE-positive. If an ECG was taken after 24 hours of the CTPA, it was discarded to minimize the impact of medical intervention started after the
positive scan. ECGs taken before 24 hours but less than six months of a positive CTPA were also
conservatively discarded, since knowledge of when the PE began could not be assessed through ordinary
means. If the CTPA was PE-negative, all ECGs were labeled as PE-negative. Example scenarios are shown
in Supplementary Table 4.

Experimental Design

Given the low number of PE-positive CTPAs and the resulting class imbalance, we used a stratified k-fold
approach to create our cross-validation (training and validation) and testing sets. We first split all patients
by those who have at least one positive CTPA (“positive”) and those who have only negative CTPAs
(“negative”) to ensure no leakage of patient data between the training, validation, and testing sets. Next, we
split each subset into 90% for cross-validation and 10% for testing. We further stratify the cross-validation
set into nine folds to maintain an overall 80-10-10 training-validation-test split. The positive and negative
patients from each split are combined to yield the final 9-fold cross-validation sets and holdout testing set.

Given the prognostic severity of PEs, thresholds for classification were set at the greatest value (to
maximize specificity) at which the models achieved 100% sensitivity (i.e. no false negative PEs) on the
validation set.

Modeling Overview

Details of ECG preprocessing to remove baseline drift and exclude poor quality waveforms are found in
Supplementary Methods (page 2) and Supplementary Figures 4-5. Each ECG is treated as a series of 8
contiguous leads (I, II, V1 - V6) of length 5000 data points (10 seconds). To predict the probability of PE
from raw ECG waveforms, we employ a convolutional neural network (CNN) model (“ECG”) composed
of a series of 12 residual blocks followed by a fully-connected layer and an output sigmoid activation layer
(Figure 2a). Details on model architecture and training are discussed in Supplementary Methods (pages
2-3).
For the EHR-only (“EHR”) models, all demographics, comorbidities, vitals, labs, and ECG morphology parameters (PR interval, QRS interval length, QTc, R-wave axis) presented in Table 1 were used as inputs. Vitals, labs, and ECG-derived parameters were associated with a patient encounter if they were collected within 24 hours of the time the CTPA scan was ordered. For encounters with multiple vitals and labs within the 24 hour time period, the closest one to the CTPA order time was used. To predict PE from the EHR dataset, we evaluated four classes of models (logistic regression, ElasticNet, random forest, and Extreme Gradient Boosting (XGBoost)) under different conditions (Supplementary Table 5).

We use a fusion modeling approach to combine ECG waveform information and tabular EHR data. To represent ECG waveform data, we use the output after all convolutional blocks (i.e., “flatten” layer) from the ECG model. Since the dataset size for fusion modeling is much smaller (i.e., samples of patient-encounters versus samples of ECGs), we use principal components analysis (PCA) to reduce the dimensionality of this intermediate output to 20 components. This lower dimensional representation of ECG waveforms after PCA is combined with tabular EHR data and used as the model’s input. Independent PCA components are also beneficial for interpretability analysis used below. For these fusion models (“Fusion”), ECGs for PE-negative patients recorded outside of the 24-hour window were not discarded, since these prior ECGs may still contain relevant information about a patient’s clinical status. To predict PE from this fusion dataset, we use the best performing model from the EHR experiments as our scaffold and assess the model’s robustness through feature sensitivity studies against a holdout test set (Supplementary Figures 10-11).

All models were developed in Python (version 3.5.1) and PyTorch (version 1.5.1). Model Interpretability

We evaluated feature contributions toward model prediction using Shapley Additive Explanations (SHAP) scores for the EHR and Fusion models. SHAP scores are a game-theoretic approach to model interpretability; they provide explanations of global model structures based upon combinations of several
local explanations for each prediction. To interpret and rank the significance of input features toward the final prediction of the model, mean absolute SHAP values were calculated for each feature across all observations in the holdout test set for each model trained on a cross-validation fold. These values are visualized using SHAP dependency plots (Figure 2c), which show how different values (dot color) of features (y-axis) affect model output (x-axis). We also perform extensive feature sensitivity analyses to investigate the effect of features that may cache clinician driven bias (e.g. a BNP lab may be drawn if competing diagnoses, like CHF, are high in suspicion or may be a sign of PE prognostication) by assessing change in model performance after their removal from the model input feature set.

Comparison to Existing Clinical Scores

To understand the performance of these models relative to existing clinical screening tools in the diagnostic workflow particularly patients with CTPA (who are at moderate- to high suspicion of PE), we compare the performance of our model against common clinical criteria for assessing patient likelihood of having PE: Wells’ Criteria for PE (Wells’)\textsuperscript{27}, the Revised Geneva Score for PE (Geneva)\textsuperscript{28}, PE Rule-Out Criteria (PERC)\textsuperscript{29}, and 4-Level PE Clinical Probability Score (4PEPS)\textsuperscript{30}. However, three of the clinical scores used (Wells’, Geneva, and 4PEPS) require D-dimer in their framework to determine whether a CTPA is warranted. To restrict clinical bias that may be embedded in those patients in our dataset with missing D-dimers and ensure a fairer comparison, we compose a manual validation sample of 100 patient encounters selected from the holdout test set, 50 of whom have D-dimer and 50 of them are randomly selected without replacement. Details on the chart review for examining these cases are found in Supplementary Methods.

Statistical Analysis

Unique CTPAs were the unit of analysis chosen to make univariate comparisons. Cohen’s kappa statistic was used to assess for interrater reliability on the test set of PE annotations. To assess model performance, mean receiver operating characteristic (ROC) and mean precision-recall (PRC) curves with 95% confidence intervals (alpha = 0.05) on the holdout test set across each fold were plotted and the mean and standard
deviation areas under the ROC curve (AUROC) and PRC curve (AUPRC) were calculated. Analysis of variance (ANOVA) was used to compare inter-model performance. Sensitivity, specificity, positive predictive value, and negative predictive value were estimated using the optimal threshold ensuring maximal sensitivity while maximizing specificity on the validation set. Model calibration plots showing the fraction of positives at each average predicted probability value for the fusion model were created. Brier score loss was used to quantitatively assess fusion model calibration. McNemar’s test was used to compare the diagnostic performance of the fusion models against the clinical scores. All statistical analyses were performed using the statsmodels library (version 0.12.0) in Python (version 3.5.1).

Role of the Funding Source

The funders had no role in study design, data collection, data analysis, data interpretation, writing the report, or the decision to submit the paper for publication. The corresponding author had full access to all data and the final responsibility to submit for publication.
Results

Study Population

A total of 30,109 potential CTPAs were retrieved from our healthcare system. After NLP-augmented annotation of these reports (Supplementary Figure 1), a total of 28,496 CTPA reports were labeled, of which 25,099 reports (88.1%) were PE-negative and 3,397 reports (11.9%) were PE-positive. Mean interrater reliability on the test set for annotation was 0.97 with a standard deviation of 0.03. After preprocessing and linking with the ECG datasets (Supplementary Figure 4-6), a total of 23,793 CTPAs (10.0% positive for PE) and 320,746 ECGs (12.8% positive for PE) across 21,183 unique patients were available for model development and testing. Within the PE-positive CTPAs, there were 126 (5.2%) truncal, 515 (21.1%) main, 692 (28.4%) lobar, and 1102 (45.3%) segmental PEs. In our sample (Table 1), PE-positive CTPAs were more common in patients who were older (mean age 60.7 vs. 57.6 years), were more likely to have a history of DVT or PE (14.9% vs. 8.2%), and had higher heart rates (93.1 vs. 90.3 beats per minute) and D-dimer levels (6.5 vs. 2.3 µg/mL FEU) on admission. Additional univariate analysis of our dataset by data split, PE location, and feature missingness are provided in Supplementary Tables 6-8.

Model Evaluation

For predicting PE from ECG waveforms, the deep neural network achieved modest performance for the cross-validation set (AUROC 0.69). Results for optimization experiments to fine-tune the sampling method, learning rate, kernel size, and model depth can be found in Supplementary Table 9. XGBoost was chosen as the scaffold for both EHR and fusion models since it performed best-in-class in the EHR-only experiments (AUROC 0.64 on the selected cross-validation set). For predicting PE from both ECG waveforms and EHR data, the fusion model achieved an AUROC of 0.82 on the selected cross-validation set.
On the holdout test set (Figure 2b), the fusion model performed the best (AUROC of 0.81 ± 0.01, AUPRC of 0.35 ± 0.01), followed by the EHR model (AUROC 0.65 ± 0.01, AUPRC 0.17 ± 0.01) and the ECG model (AUROC 0.59 ± 0.01, AUPRC 0.18 ± 0.02). For ECG model, we have also predicted the performance using PCA components only and the results (AUROC 0.59, AUPRC 0.14) are close to the original CNN results. EHR model SHAP scores on this holdout set (Figure 2c) demonstrated that higher values of age, male gender, history of DVT or PE, absence of COPD, higher heart rates, lower oxygen saturation, and lower QRS duration on ECG shifted the model output to predict PE, with the largest impact stemming from a positive history of DVT or PE (towards predicting PE) and positive history of COPD (towards predicting not PE). Fusion model SHAP scores on the holdout set showed model output being driven more by continuous features, such that model output was shifted to predict PE with older age, higher D-dimer, higher QTc, lower QRS duration, lower values of the first principal component from the ECG waveform representation (PCA 1), and higher values of PCA 3. In subgroup analysis of model performance by PE location, all models detected truncal PEs the best, followed in order by main, lobar, and segmental PEs (Supplementary Figure 7). Finally, we found comparable performance in fusion model performance by gender (mean AUROC 0.81) and race (mean AUROC 0.77 – 0.84) (Supplementary Figure 8).

To assess the impact of missingness and data derived from clinician decision-making on model prediction, we perform a series of feature sensitivity experiments (Supplementary Figure 9). Model performance on the holdout test set remained the same when removing race and sex (AUROC 0.81 ± 0.01), comorbidities (AUROC 0.81 ± 0.01), and vitals (AUROC 0.81 ± 0.01) but decreased with removal of D-Dimer (AUROC 0.78 ± 0.01), all labs (AUROC 0.76 ± 0.01), and ECG morphology parameters (AUROC 0.75 ± 0.01). The fusion model was overall well-calibrated (Brier score 0.076) and particularly had greater accuracy for negative-PE cases, for which the prediction probability of the model is lower and maps to label of 0 (Supplementary Figure 10).
Model Performance Against Clinical Scores

On all patients in the sample holdout set, we benchmark our model performance against standard-of-care scoring systems through a threshold-independent metric (AUROC) and a threshold-dependent metric (specificity). On threshold-independent evaluation (Figure 3a), the fusion model with D-dimer (AUROC 0.84 ± 0.02) and without D-dimer (AUROC 0.70 ± 0.02) outperformed the clinical scores: Wells’ criteria (AUROC 0.54), revised Geneva score (AUROC 0.52), PERC (AUROC 0.50), and 4PEPS (AUROC 0.58).

When thresholding these scores based on published standards to exclude PE without the need for a D-dimer, the Wells’ criteria score missed the most PEs (8), followed by the Geneva Score (7) and 4PEPS (3). Only PERC missed no PEs. For equitable comparison, the fusion model trained without D-dimer (since this is not available to these clinical scores) using cross-validation set achieved perfect sensitivity (1.00) with greater specificity (0.18) than the Wells’ criteria (0.00) and PERC (0.03) (Table 2). Even when setting classification thresholds to achieve perfect sensitivity by adjusting the threshold for PE using the holdout test set, the fusion model without D-dimer had greater specificity (0.30) compared to 4PEPS (0.05), PERC (0.03), Wells’ criteria (0.00), and the revised Geneva score (0.00) (Table 2). To assess the impact of missingness on model performance, another feature sensitivity study was performed on the fusion model, which demonstrated better performance of these fusion model (AUROC 0.63 to 0.84) relative to the clinical scoring systems (AUROC 0.50 to 0.58) (Supplementary Figure 11). Notably, model performance suffered minimally with removal of demographics (AUROC 0.83), comorbidities (AUROC 0.83), vitals (AUROC 0.83), and ECG morphology parameters (AUROC 0.81).
Discussion

In this study, we present a novel approach for detecting PEs through a fusion modeling framework built on a large dataset of readily available clinical data and routinely collected ECG waveforms linked to annotated CTPA reports. We demonstrate that raw ECG waveform embeddings act synergistically with tabular clinical data to detect PE with greater specificity (AUROC 0.84, specificity 0.18) in those at moderate- to high-suspicion relative to commonly employed clinical scores (AUROC 0.50-0.58, specificity 0.00-0.05) in our retrospective cohort.

Improving PE detection remains a crucial modeling task for informaticians and clinical researchers. Despite the steady decrease in mortality attributable to PE, improving diagnostic yield remains important given that the condition remains underdiagnosed globally. However, given the high morbidity and mortality risk of PE, current clinical prediction models overpredict further workup with CTPA to reduce missed PEs. The downside of this approach is the overreliance on CTPAs and decreased diagnostic yield of PEs, which results in greater patient exposure to contrast and ionizing radiation, contributes to increased patient time in the hospital, and worsens systemic inefficiencies. Furthermore, this CTPA utilization is more selectively increasing in older populations, who are at increased cancer risk from ionizing radiation.

Our work represents an integrated deep learning approach on a clinically relevant subset of patients suspected of having acute PE. Past work integrating machine learning approaches to PE has relied on integrating EHR data on a smaller subset of CTPA images. This past work also relies on clinical data alone, while we provide the first integration of raw ECG waveforms and show that it improves outcome prediction. For example, we note that traditional ECG measurements such as QTc and QRS duration had an important influence on model output. But we also found that embeddings from raw ECG waveforms also had a large contribution on model prediction, suggesting thereby that these ECG signals provide information beyond what is contained in traditional ECG measurements. This is a novel finding and suggests that further work must be done to identify relevant clinical features our deep learning models are
identifying from ECGs that otherwise remain indiscernible to clinicians. For example, we suspect that our framework may be picking up subtle electrocardiographic signs of increased predisposition to thrombus formation (such as morphologic characteristics representing underlying hypertensive heart disease) or more acute findings suggestive of subclinical PE manifestation (such as right heart strain). This, however, remains difficult to parse out and we are undergoing further post-hoc analysis to better understand the implications of our model’s outputs.

Moreover, while deep learning-based systems may lack straightforward interpretation, we remain encouraged by patterns emerging from our outputs in analysis of the holdout data. First, we observe that our model demonstrates monotonically increasing performance in more proximal vascular territories, which have the most severe clinical implications. In previous machine learning based analysis of PEs, it is unclear whether there was a differential preponderance for the detection of more anatomically relevant lesions. Second, we note that our interpretability analysis prioritized features traditionally associated with increased PE incidence and risk, including age, previous VTE, and vital sign abnormalities. The model also appropriately predicted lower PE likelihood in those with COPD, a feature inversely predictive of PE probability and a prominent feature of other PE prediction scores such as the 4PEPS algorithm. Third, our model performance is comparable across gender and race stratifications, likely reflective of the diverse population on which we trained our model. Increased representation of minority groups and women is imperative for increasing equity in AI where white males tend to be overrepresented in training data. We additionally note that big data approaches, particularly ones relying on routinely collected EHR data, are often hampered by data quality issues, impeding the progress of algorithmic approaches. However, our extensive feature sensitivity experiments attest to the robustness of our framework, even in the setting of missing or noisy data. Indeed, even in the absence of strong prognostic markers such as D-dimer, our approach outperforms other available clinical prediction rules (Supplementary Figure 11).

As a case study, we examined in greater detail a 53-year-old male patient with diabetes mellitus, hypertension, and chronic stable angina in our clinical validation cohort with a saddle PE. On admission,
the patient’s chief complaint was substernal chest pain that began one day prior and resolved with nitroglycerin with a weeklong history of exertional fatigue, dyspnea on exertion, and lightheadedness. His physical exam on admission was remarkable only for tachycardia (heart rate of 123) but notably without hypoxia, tachypnea, or any lower extremity edema or pain. While a troponin-I was normal at 0.16 ng/mL and BNP of 19.78 pg/mL, D-dimer was notably remarkably elevated at 13.7 µg/mL FEU. Sinus tachycardia with bialtrial enlargement and new concordance of previously discordant T-waves in V5 and V6 were present on the admission ECG, while right ventricular hypokinesis was noted on bedside transthoracic echocardiography. In the absence of any reported estrogen use, a history of recent immobilization or active cancer, hemoptysis, or other PE risk factors, this gentleman had remarkably low initial clinical likelihood scores for PE (Wells’ criteria of 1.5, Geneva score of 5, PERC of 2, and 4PEPS of 1). From admission notes, the clinical suspicion for PE from providers in the Emergency Department was low, with the triad of troponin, BNP, and D-dimer initially drawn nonspecifically and CTPA diagnosis ascertained 10 hours after presentation. However, both ECG (model output 0.26, threshold=0.05), EHR (model output 0.18, threshold=0.103), and fusion models (model output of 0.20, for both with and without admission labs as features) provided outputs well above their decision thresholds for recommending evaluation by CTPA (Table 2).

Our current study faces some limitations. First, given the retrospective nature of this dataset, the association of all input data modalities as being PE-positive and PE-negative within a specific time frame to strike an appropriate balance between missingness and timeliness remains a source of modeling error. While our framework attests to the value of combining ECG waveforms and routinely collected clinical data, data collection will need to reflect information more optimally at the point of presentation. However, this current framework may serve as an effective pre-trained scaffold for this future analysis. While overall discrimination was favorable for the fusion model, the model was only well-calibrated up until a moderate clinical predicted probability. This is consistent with the low overall prevalence of PE in our development data. It is also important to mention the limitation of comparing our models with the clinical scores, which
have robust performance in the general ED population (where the bulk of the patients are at low-suspicion of PE) but perform poorly in patients with an elevated pretest probability of PE. Since this dataset contains only patients who have already had CTPA and met some threshold for PE suspicion, the clinical scores have a poor performance. This comparison only serves to demonstrate how limited these scores are at discriminating moderate- to high-risk PE patients. But we have shown that our models can improve the performance for this group of patients. Inferences about how the model compares to the scores in a general ED population requires further research. Future work will focus on assessing the algorithm’s performance on patients with low suspicion for PE that were excluded by a negative D-dimer and ultimately never received a CTPA to compare against existing clinical scoring systems more accurately. Despite demonstrations of model robustness against missingness and cached clinician decision making, there is a risk that the institutional and provider bias in the diagnostic workup a patient undergoes, or just as importantly fails to undergo, has been embedded in this algorithm. However, it is difficult to discern whether the marginal loss in performance from feature ablation stems from inability to capitalize on this embedded clinical decision-making bias or an informative signal helping to provide information about the pathophysiological state of the patient. Given that clinical decision-making tools have existed for decades, we assumed that medically appropriate diagnostic workup was undertaken since our source data comes from an academic hospital that reflects the ESC guidelines for PE diagnosis. To improve generalizability, we plan to validate the algorithm at external sites that have different patient populations, clinical workflows, and different data collection modalities (e.g., ECG machines, blood chemistry assays).

In summary, we present a fusion modeling framework that uses raw electrocardiogram waveforms and traditional clinical variables to predict acute PE more accurately and specifically than existing clinical scores in those patients with at least a moderate clinical suspicion of acute PE.
Acknowledgements

We acknowledge Manbir Singh and Mark Shervy with IT support. We thank Shelly Kaur and Yovanna Roa for administrative and project management support. We thank Zahi Fayad, Thomas Fuchs, Erwin Bottinger, and Alexander Charney for infrastructural support. We thank GE and Percy LaRosa, Adeyinka Fadehan, Thareesh Dondapati, and Ejaz Siddiqui for facilitating data access. This study was supported by the National Center for Advancing Translational Sciences, National Institutes of Health (U54 TR001433-05).
References


   Desmaison A, Köpf A, Yang E, DeVito Z, Raison M, Tejani A, Chilamkurthy S, Steiner B, Fang L, Bai J, 


   Excluding pulmonary embolism at the bedside without diagnostic imaging: management of patients with 

   embolism in the emergency department: the revised Geneva score. Ann Intern Med American College of 
   Physicians; 2006;144:165–171.

29. Kline JA, Mitchell AM, Kabrhel C, Richman PB, Courtney DM. Clinical criteria to prevent unnecessary 
   diagnostic testing in emergency department patients with suspected pulmonary embolism. J Thromb Haemost 

   and Validation of a 4-Level Clinical Pretest Probability Score for Suspected Pulmonary Embolism to Safely 
   Decrease Imaging Testing. JAMA Cardiol 2021;

31. statsmodels. Github;

   Konstantinides SV. Trends in mortality related to pulmonary embolism in the European Region, 2000-15: 


Figure Legends

**Figure 1:** Study Design

**A:** Our pipeline for creating models to detect PE consists of using three data modalities: electrocardiograms (ECGs), clinical data (EHR) including patient demographics, comorbidities, vital signs, and relevant labs, and computed tomography pulmonary angiograms (CTPAs) that are labeled using a two-stage approach combining natural language processing (NLP) pattern matching and manual clinician annotations. These data are linked together to develop, analyze, and benchmark models to predict PE. **B:** We split our dataset for training, validation, and testing first by first identifying all unique patients (not unique CTPA or unique ECG) and separating them based on whether they have at least one PE-positive CTPA scan (PE+) or not (PE–). This stratum is further split into 90% for 9-fold cross-validation (89% for training, 11% for model selection and model development) and 10% for testing to assess model performance and benchmark against clinical scores. **C:** ECGs are labeled as PE-positive (PE+) if they are recorded within 24 hours of a PE-positive CTPA. ECGs recorded 24 hours after or between 6 months and 24 hours before a positive CTPA are discarded. ECGs not meeting the above criteria for PE-positive CTPAs are labeled PE-negative. EHR data is retained if collected within 24 hours of the CTPA and labeled equally with the CTPA finding.

**Figure 2:** Modeling Overview, Performance, and Interpretability

**A:** The ECG model, which is a convolutional neural network with residual connections, trains and infers PE likelihood using 10-second long waveform from 8 leads (I, II, V1-V6) recorded at 500 Hz. The EHR model is an Extreme Gradient Boosting (XGBoost) model that uses tabular clinical data (demographics, comorbidities, labs, and vital signs) and ECG morphology parameters to predict the likelihood of PE. Finally, the fusion model is an XGBoost model that uses a principal component decomposition of an ECG waveform embedding from the ECG model, tabular clinical data, and ECG morphology parameters in an XGBoost framework to predict the likelihood of PE. **B:** Mean receiver-operating characteristic (top) and precision-recall (bottom) curves with 95% confidence intervals for the ECG (red), EHR (blue), and Fusion
models, with the mean and standard deviations for the area under each respective curve (AUROC, AUPRC) in the figure legend. In top plot, the horizontal and vertical lines correspond to optimal threshold. The Fusion model outperforms both the ECG and EHR models. C: SHAP dependency plots for the EHR model (top) and Fusion model (bottom), representing the marginal contribution from patient encounters in the test set (dots, colored by value of feature) of different features (y-axis, in descending order of importance) on the model output (x-axis, positive favors increased PE likelihood). Gray dots represent samples with missing data points.

**Figure 3: Clinical Benchmark and Integration**

A: Mean receiver-operating characteristic (ROC, left) and precision-recall (PRC, right) curves with 95% confidence intervals for the Fusion model with (pink) and without (brown) D-dimer, whereas ROC and PRC are shown for the clinical scores – Wells’ Criteria (yellow), Revise Geneva Score (green), PERC (red), and 4PEPS (purple). In top plot, the horizontal and vertical lines correspond to optimal threshold. Mean and standard deviations for the area under each respective curve (AUROC, AUPRC) for the Fusion models are displayed in the legend, whereas area under each respective curve (AUROC, AUPRC) are shown for the clinical scores. B: The Fusion model may be used to recommend CTPA or exclude PE in patients with moderate to high likelihood of PE after clinical stratification or those at low suspicion with an abnormal D-dimer.
A. Pipeline Overview

EHR
- Demographics
- Vitals
- Comorbidities
- Labs

CTPA
- NLP
- Clinician Annotations

Pipeline Overview

Robustness Experiments

Model Development

Clinical Validation

B. Stratified Folds for Modeling

All patients

90% 10% 10% 90%

9-fold Cross-Validation

Model selection

Model development

Holdout (Testing)

Model performance
Clinical validation

C. Linking CTPAs

Screening for PE

Triage
H&P, labs, ECG

AI-guided stratification

Clinician-guided likelihood stratification

CTPA

Other Workup

Earliest CTPA

By guest on 06 December 2021
C.

EHR

Demographics

Comorbidities

Vitals

Labs

CTPA

NLP

Clinician Annotations

Robustness Experiments

Model Development

Clinical Validation

A.

Clinical Benchmarks

Graphical Abstract

B.

Suggested Workflow

PE Suspected

Assess clinical probability of PE

P(PE) low

D-Dimer normal

P(PE) high

D-Dimer abnormal

Fusion

No CTPA Recommended

CTPA Recommended

CTPA Recommended

Exclude

CTPA

Features

Models

ECG (Nb x 8 x 5000)

-ECG Model-

CNN

Residual block

FCN

ECG Metrics

Waveforms

EHR

Demographics

Comorbidities

Vitals

Labs

CTPA

NLP

Clinician Annotations

Robustness Experiments

Model Development

Clinical Validation

No CTPA

Recommended

CTPA

Recommended

PE Suspected

Assess clinical probability of PE

P(PE) low

D-Dimer normal

P(PE) high

D-Dimer abnormal

Fusion

No CTPA Recommended

CTPA Recommended

Exclude

CTPA

CTPA

Downloaded from https://academic.oup.com/ehjdh/advance-article/doi/10.1093/ehjdh/ztab101/6440044 by guest on 06 December 2021
A. Model Architectures

- ECG Model -
  - CNN K=32
  - Residual block
  - FCN
  - σ

- EHR Model -
  - PCA

- Fusion Model -
  - σ

B. Model Performance

Receiver-Operating Characteristic Curve

Precision-Recall Curve

C. Model Interpretability

EHR Model

Fusion Model

Features

Models

Outputs

ECG (Nb x 8 x 5000)

EHR (Nb x 30)

PCA

EHR Model

Fusion Model

PC 1

D-Cimer

B-NP

Age

PC 2

PC 3

DVT or PE

Troponin

PE | ECG

PE | EHR

PE | EHR, ECG

Figure 2

Click here to access/download: Figure PE_ECG_Figure2_Revised_103.png
A. Clinical Benchmarks

Receiver-Operating Characteristic Curve

Precision-Recall Curve

B. Suggested Workflow

PE Suspected

Assess clinical probability of PE

D-Dimer

D-dimer normal

D-dimer abnormal

P(PE) low

P(PE) high

No CTPA Recommended

CTPA Recommended

Fusion

Exclude

CTPA

Figure 3

Click here to access/download Figure PE_ECG_Figure3_Revised.eps

Downloaded from https://academic.oup.com/ehjdh/advance-article/doi/10.1093/ehjdh/ztab101/6440044 by guest on 06 December 2021
Table 1: Baseline characteristics in the PE-positive and PE-negative cohorts.

<table>
<thead>
<tr>
<th></th>
<th>CTPA or PE-Encounter*</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Missing</td>
</tr>
<tr>
<td>n</td>
<td>23793</td>
</tr>
<tr>
<td>Age (years), mean (SD)</td>
<td>57.9 (17.7)</td>
</tr>
<tr>
<td>Sex, n (%)</td>
<td>13441 (63.0)</td>
</tr>
<tr>
<td>Female (code=0)</td>
<td>1345 (55.3)</td>
</tr>
<tr>
<td>Male (code=1)</td>
<td>8994 (37.8)</td>
</tr>
<tr>
<td>Race, n (%)</td>
<td>13786 (62.2)</td>
</tr>
<tr>
<td>Asian</td>
<td>1345 (55.3)</td>
</tr>
<tr>
<td>Black</td>
<td>6376 (27.0)</td>
</tr>
<tr>
<td>Hispanic</td>
<td>3486 (14.8)</td>
</tr>
<tr>
<td>White</td>
<td>8083 (34.2)</td>
</tr>
<tr>
<td>Other</td>
<td>5068 (21.5)</td>
</tr>
<tr>
<td>Arrhythmia, n (%)</td>
<td>2371 (11.2)</td>
</tr>
<tr>
<td>Coronary Artery Disease, n (%)</td>
<td>2371 (11.2)</td>
</tr>
<tr>
<td>Cancer, n (%)</td>
<td>2371 (11.2)</td>
</tr>
<tr>
<td>Chronic Kidney Disease, n (%)</td>
<td>2371 (11.2)</td>
</tr>
<tr>
<td>Coagulopathy, n (%)</td>
<td>2371 (11.2)</td>
</tr>
<tr>
<td>Chronic Obstructive Pulmonary Disease, n (%)</td>
<td>2371 (11.2)</td>
</tr>
<tr>
<td>Diabetes Mellitus, n (%)</td>
<td>2371 (11.2)</td>
</tr>
<tr>
<td>History of DVT or PE, n (%)</td>
<td>2371 (11.2)</td>
</tr>
<tr>
<td>Congestive Heart Failure, n (%)</td>
<td>2371 (11.2)</td>
</tr>
<tr>
<td>Hypertension, n (%)</td>
<td>2371 (11.2)</td>
</tr>
<tr>
<td>Pregnancy, n (%)</td>
<td>2371 (11.2)</td>
</tr>
<tr>
<td>Pulmonary Hypertension, n (%)</td>
<td>2371 (11.2)</td>
</tr>
<tr>
<td>Rheumatological Disease, n (%)</td>
<td>2371 (11.2)</td>
</tr>
<tr>
<td>Heart Rate (bpm), mean (SD)</td>
<td>90.6 (22.6)</td>
</tr>
<tr>
<td>Systolic Blood Pressure (mmHg), mean (SD)</td>
<td>128.7 (26.8)</td>
</tr>
<tr>
<td>Diastolic Blood Pressure (mmHg), mean (SD)</td>
<td>71.2 (15.1)</td>
</tr>
<tr>
<td>Respiration Rate (breaths per minute), mean (SD)</td>
<td>19.6 (4.4)</td>
</tr>
<tr>
<td>Oxygen Saturation (%), mean (SD)</td>
<td>96.9 (14.8)</td>
</tr>
<tr>
<td>BNP (pg/mL), mean (SD)</td>
<td>437.7 (1078.4)</td>
</tr>
<tr>
<td></td>
<td>Mean (SD)</td>
</tr>
<tr>
<td>--------------------------</td>
<td>--------------------</td>
</tr>
<tr>
<td>D-dimer (µg/mL FEU), mean (SD)</td>
<td>16248 (2.7 (7.2))</td>
</tr>
<tr>
<td>Troponin (ng/mL), mean (SD)</td>
<td>10375 (0.2 (4.3))</td>
</tr>
<tr>
<td>PR Interval Length (ms), mean (SD)</td>
<td>7737 (151.5 (29.5))</td>
</tr>
<tr>
<td>QRS Duration (ms), mean (SD)</td>
<td>6402 (89.3 (19.9))</td>
</tr>
<tr>
<td>QTc (ms), mean (SD)</td>
<td>6402 (450.0 (37.4))</td>
</tr>
<tr>
<td>Cardiac Axis (degrees), mean (SD)</td>
<td>6402 (26.1 (50.0))</td>
</tr>
<tr>
<td>PE Location, n (%)</td>
<td>No PE</td>
</tr>
<tr>
<td>0</td>
<td>21358 (89.8)</td>
</tr>
<tr>
<td></td>
<td>21358 (100.0)</td>
</tr>
</tbody>
</table>

*Unit of analysis is unique patient encounter or, by extension, unique CTPAs.
Table 2: Fusion model performance benchmarked against clinical scores.

<table>
<thead>
<tr>
<th>Threshold Source</th>
<th>Model</th>
<th>Threshold</th>
<th>TP</th>
<th>TN</th>
<th>FP</th>
<th>FN</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>PPV</th>
<th>NPV</th>
<th>McNemar Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cross-Validation Fold</td>
<td>Fusion Model</td>
<td>0.037</td>
<td>22</td>
<td>14</td>
<td>62</td>
<td>0</td>
<td>1.00</td>
<td>0.18</td>
<td>0.26</td>
<td>1.00</td>
<td>(Base)</td>
</tr>
<tr>
<td></td>
<td>Fusion Model (No D-Dimer)</td>
<td>0.040</td>
<td>22</td>
<td>14</td>
<td>62</td>
<td>0</td>
<td>1.00</td>
<td>0.18</td>
<td>0.26</td>
<td>1.00</td>
<td>1.000</td>
</tr>
<tr>
<td>Wells et al. [26]</td>
<td>Wells' Criteria</td>
<td>1</td>
<td>14</td>
<td>24</td>
<td>52</td>
<td>8</td>
<td>0.64</td>
<td>0.32</td>
<td>0.21</td>
<td>0.75</td>
<td>0.005</td>
</tr>
<tr>
<td>Le Gal et al. [27]</td>
<td>Revised Geneva Score</td>
<td>4</td>
<td>15</td>
<td>20</td>
<td>56</td>
<td>7</td>
<td>0.68</td>
<td>0.26</td>
<td>0.21</td>
<td>0.74</td>
<td>0.041</td>
</tr>
<tr>
<td>Kline et al. [28]</td>
<td>PERC</td>
<td>1</td>
<td>22</td>
<td>2</td>
<td>74</td>
<td>0</td>
<td>1.00</td>
<td>0.03</td>
<td>0.23</td>
<td>1.00</td>
<td>0.185</td>
</tr>
<tr>
<td>Roy et al. [29]</td>
<td>4PEPS</td>
<td>0</td>
<td>19</td>
<td>19</td>
<td>57</td>
<td>3</td>
<td>0.86</td>
<td>0.25</td>
<td>0.25</td>
<td>0.86</td>
<td>0.002</td>
</tr>
<tr>
<td>Holdout Test Set</td>
<td>Fusion Model</td>
<td>0.091</td>
<td>22</td>
<td>43</td>
<td>33</td>
<td>0</td>
<td>1.00</td>
<td>0.57</td>
<td>0.40</td>
<td>1.00</td>
<td>(Base)</td>
</tr>
<tr>
<td></td>
<td>Fusion Model (No D-Dimer)</td>
<td>0.072</td>
<td>22</td>
<td>21</td>
<td>55</td>
<td>0</td>
<td>1.00</td>
<td>0.28</td>
<td>0.29</td>
<td>1.00</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td></td>
<td>Wells' Criteria</td>
<td>0</td>
<td>22</td>
<td>0</td>
<td>76</td>
<td>0</td>
<td>1.00</td>
<td>0.00</td>
<td>0.22</td>
<td>0.00</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td></td>
<td>Revised Geneva Score</td>
<td>0</td>
<td>22</td>
<td>0</td>
<td>76</td>
<td>0</td>
<td>1.00</td>
<td>0.00</td>
<td>0.22</td>
<td>0.00</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td></td>
<td>PERC</td>
<td>1</td>
<td>22</td>
<td>2</td>
<td>74</td>
<td>0</td>
<td>1.00</td>
<td>0.03</td>
<td>0.23</td>
<td>1.00</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td></td>
<td>4PEPS</td>
<td>-1</td>
<td>22</td>
<td>4</td>
<td>72</td>
<td>0</td>
<td>1.00</td>
<td>0.05</td>
<td>0.23</td>
<td>1.00</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>

*TP: true positives; TN: true negatives; FP: false positives; FN: false negatives; PPV: positive predictive value; NPV: negative predictive value; PERC: Pulmonary Embolism Rule-Out Criteria; 4PEPS: 4-Level Pulmonary Embolism Clinical Probability Score. Encounters are classified as PE-positive if the clinical score or model likelihood is greater than or equal to the threshold value.