A IRB Approval and Data De-identification

Release of INSPECT was approved by the Stanford University Institutional Review Board (IRB), given data privacy review via a standardized workflow conducted by the Center for Artificial Intelligence in Medicine and Imaging (AIMI) and the University Privacy Office. Our study was approved by the Stanford University Administrative Panel on Human Subjects Research, protocol #24883, and included a waiver of consent. All included patients from SHC signed a privacy notice, which informs them that their records may be used for research purposes given approval by the IRB.

All INSPECT data (CT scans, DICOM metadata, radiology impression sections, EHR timelines) are manually reviewed by AIMI to confirm any protected health information (PHI) is removed before public release. We de-identify each modality as follows:

EHR Timelines: All dates are anonymized by using per-patient time jittering. We apply the same date transformation procedure used by MIMIC-III, specifically: "*[d]ates were shifted into the future by a random offset for each individual patient in a consistent manner to preserve intervals, resulting in stays which occur sometime between the years 2100 and 2200*" [33]. We remove all patients >89 years of age. We further remove all unstructed text fields that do not map to controlled vocabularies (e.g., SNOMED, LOINC) to prevent PHI leakage We use OHDSI Athena [1] ontologies to describe our data, which includes both public ontologies like ICD-10 as well as OHDSI specific ontologies such as Race/Gender. The full list of ontologies used is in Table 6..

Ontology
OMOP Extension
Medicare Specialty
CPT4
CVX
ICD9Proc
RxNorm
SNOMED
RxNorm Extension
Cancer Modifier
ICD10PCS
CMS Place of Service
Visit
Ethnicity
Gender
ICDO3
Race
LOINC
HCPCS

Table 6: OHDSI Athena ontolgies used in our benchmark

CT Scans: DICOM image data were converted to NumPy pixel arrays and all DICOM tags exported as metadata to remove patient identifying information. Each CT scan slice is manually reviewed for PHI, with slices containing patient information removed from the CT scan.

Radiology Notes: Radiology notes are preprocessed to include only the impression section, i.e., the description of radiologist findings in the coresponding CT scan. Each note is processsed to tag names, locations, dates, telephone numbers and other HIPAA protected idenifiers such as MRNs and accession numbers. These tags are then replaced wth anonymized placeholder text. All deidentified notes are then manually reviewed to remove any additonal PHI.

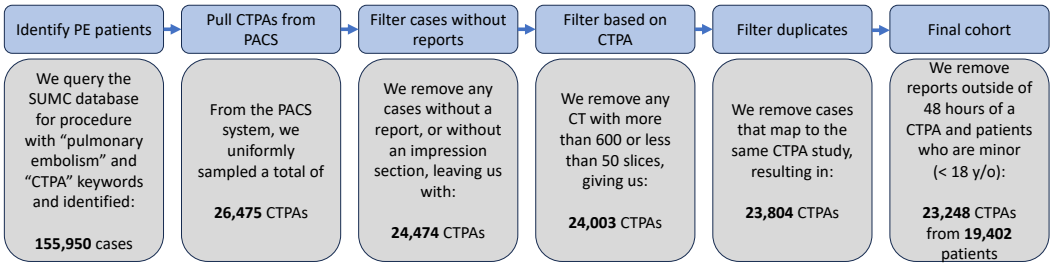

Figure 3: A flowchart of our cohort definition process.

B Cohort Definition

The flowchart of our cohort definition protocol is illustrated in Figure 3. With the approval from Stanford Institutional Review Board's (IRB), we identified 155,590 cases with the CT pulmonary angiography (CTPA) procedure code from the STAnford medicine Research data Repository (STARR) [15]. STARR data contains routinely collected EHR data from Stanford Health Care covering the time period of 2000 to 2021. We mapped each of these studies to their respective CTPA based on the procedure date. In instances where there was no exact match (1,296 cases) we extended the search to 10 days post-procedure date. From the mappable cases, we sampled uniformly at random 26,475 CT scans (chosen due to file storage constraints) and their corresponding radiology report.

The data cleaning phase followed, where we removed cases without a report or an impression section. This refinement process resulted in 24,474 cases for further analysis. We then selected the most relevant CTPA series per study by enforcing a slice thickness constraint between 1.0mm and 3.0mm, favoring thicker slice series. Additionally, CTs with over 600 or under 50 slices were removed from consideration. This filtering resulted in 24,003 studies. We further eliminated the test and validation data from RSNA and Radfusion in our training split and the training data thereof in our test and validation split. The patients who are minors (age < 18 years old) are removed due to privacy policy. The final selection phase involved eliminating cases with corrupted DICOMs and cases from the same patient that mapped to an identical study. The final result is a collection of 23,248 CTPA studies from a set of 19,402 unique patients.

C Dataset Documentation

C.1 Hosting, Access, License, and Long-Term Preservation

We share data and trained model weights under Data Use Agreement (DUA) for non-commercial, research use. The Stanford AIMI Center will host and ensure long-term preservation of all data. Complete licensing terms for dataset and models are provided below.

INSPECT is available at https://stanfordaimi.azurewebsites.net/. We provide a preview subset of the entire dataset for reviewers. Before public release, the entire dataset is currently undergoing manual review by the AIMI Center to ensure no leakage of patient indentifying information.

As authors of the submitted dataset and corresponding manuscript, we hereby affirm that we take full responsibility for its contents. We ensure that this dataset and manuscript are original and that all data collection procedures were carried out ethically, respecting all relevant rights and regulations. We confirm that we have procured all necessary permissions for the use of the data included in the dataset, and that the data does not infringe upon any existing copyright, proprietary, or personal rights of others.

C.2 License Terms of Use

By registering for downloads from the INSPECT Dataset, you are agreeing to this Research Use Agreement, as well as to the Terms of Use of the Stanford University School of Medicine website as posted and updated periodically at http://www.stanford.edu/site/terms/.

Permission is granted to view and use the INSPECT Dataset without charge for personal, non-commercial research purposes only. Any commercial use, sale, or other monetization is prohibited.

Other than the rights granted herein, the Stanford University School of Medicine ("School of Medicine") retains all rights, title, and interest in the INSPECT Dataset.

You may make a verbatim copy of the INSPECT Dataset for personal, non-commercial research use as permitted in this Research Use Agreement. If another user within your organization wishes to use the INSPECT Dataset, they must register as an individual user and comply with all the terms of this Research Use Agreement.

YOU MAY NOT DISTRIBUTE, PUBLISH, OR REPRODUCE A COPY of any portion or all of the INSPECT Dataset to others without specific prior written permission from the School of Medicine.

YOU MAY NOT SHARE THE DOWNLOAD LINK to the INSPECT Dataset to others. If another user within your organization wishes to use the INSPECT Dataset, they must register as an individual user and comply with all the terms of this Research Use Agreement.

You must not modify, reverse engineer, decompile, or create derivative works from the INSPECT Dataset. You must not remove or alter any copyright or other proprietary notices in the INSPECT Dataset.

The INSPECT Dataset has not been reviewed or approved by the Food and Drug Administration, and is for non-clinical, Research Use Only. In no event shall data or images generated through the use of the INSPECT Dataset be used or relied upon in the diagnosis or provision of patient care.

THE INSPECT Dataset IS PROVIDED "AS IS," AND STANFORD UNIVERSITY AND ITS COLLABORATORS DO NOT MAKE ANY WARRANTY, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE, NOR DO THEY ASSUME ANY LIABILITY OR RESPONSIBILITY FOR THE USE OF THIS INSPECT Dataset.

You will not make any attempt to re-identify any of the individual data subjects. Re-identification of individuals is strictly prohibited. Any re-identification of any individual data subject shall be immediately reported to the School of Medicine.

Any violation of this Research Use Agreement or other impermissible use shall be grounds for immediate termination of use of this INSPECT Dataset. In the event that the School of Medicine determines that the recipient has violated this Research Use Agreement or other impermissible use has been made, the School of Medicine may direct that the undersigned data recipient immediately return all copies of the INSPECT Dataset and retain no copies thereof even if you did not cause the violation or impermissible use.

In consideration for your agreement to the terms and conditions contained here, Stanford grants you permission to view and use the INSPECT Dataset for personal, non-commercial research. You may not otherwise copy, reproduce, retransmit, distribute, publish, commercially exploit or otherwise transfer any material.

Limitation of Use: You may use INSPECT Dataset for legal purposes only.

You agree to indemnify and hold Stanford harmless from any claims, losses or damages, including legal fees, arising out of or resulting from your use of the INSPECT Dataset or your violation or role in violation of these Terms. You agree to fully cooperate in Stanford's defense against any such claims. These Terms shall be governed by and interpreted in accordance with the laws of California.

C.3 Data Format

We detail and define our cohort in the data using a master cohort CSV file (inspect_cohort.csv), with the following primary columns.

1. patient_id: The de-identified patient id
2. procedure_time: The date of the CTPA procedure, in ISO 8601 format
3. split: A string, either "train", "valid", or "test" that indicates the data split for this patient

The primary keys for our cohort are patient_id and procedure_time. Every case, label, and feature set is associated with a patient_id / procedure_time pair.

This file also includes the following demographic columns: age, gender, race, ethnicity.

We additionally include various columns for labels.

First, we include three NLP based pulmonary embolism diagnostic label columns. pe_positive_nlp is the main PE label used in all of our experiments and described as "Positive PE" in Appendix D.1, pe_acute_nlp and pe_subsegmentalonly_nlp are the other two label columns that are similarly described in Appendix D.1. Every case in our cohort is assigned either "True" or "False" for each of these columns.

Second, we include seven prognostic label columns. These label columns correspond to the seven prognostic tasks in our paper and have the following names: 1_month_mortality, 6_month_mortality, 12_month_mortality, 1_month_readmission, 6_month_readmission, 12_month_readmission 12_month_PH. Every case in our cohort is assigned either "True", "False", or "Censored" for each of these columns.

Finally, we include indicators for whether or not each case in this dataset is also present in either of the RNSA or Radfusion datasets using the rnsa and radfusion columns respectively.

C.3.1 CTPA

The CTPAs are made available in NumPy format. Initially, every CTPA slice undergoes resizing to dimensions of 512x512, after which it is rescaled according to the equation $x = x * r_s + r_i$, with x representing the CTPA slice, r_s signifying the RescaleSlope, and r_i denoting the RescaleIntercept. The RescaleSlope and RescaleIntercept values are directly obtained from the DICOM headers. Post these preliminary processing stages, slices are organized in the order of the patient's position, subsequently stacked, and preserved as numpy arrays.

C.3.2 DICOM Headers

We manually selected the following DICOM headers and released them as a csv file: ['InstanceNumber', 'ImagePositionPatient', 'PixelSpacing', 'RescaleIntercept', 'RescaleSlope', 'WindowCenter', 'WindowWidth', 'Manufacturer','SliceThickness']

C.3.3 Radiologist Report Impressions

Radiologist reports with impressions sections, after the anonymization process, are included in a CSV file with the name INSPECT_anon_impression.csv.

It contains columns with the names: PatientID, StudyTime and anon_impression, where the first two columns are used to map to the master cohort file and anon_impression is the anonymized impression section.

C.3.4 Structured EHR Data

Structured EHR data is released as gzipped CSV files in the FEMR format, which is documented at `https://github.com/som-shahlab/femr/blob/main/tutorials/2b_Simple_ETL.ipynb`. We release all known diagnoses, procedures, lab tests, medications, visits, and death records for patients in our cohort. The FEMR format is a simplified subset of OMOP 5.3 [57], with a subset of the columns and tables. It can be processed with any CSV reader as well as with the FEMR software package.

C.4 Dataset Statistics

C.4.1 CTPA

Based on our inclusion criteria, each CTPA study can have between 50 to 600 slices. On average, each CTPA has 220.6 slices, giving us a total of 5,164,472 CT slices in our dataset. The CTPA studies range from 1.00mm to 3.00mm (Table 7) collected from CT scanners by 3 different manufacturers (Table 8).

C.4.2 Structured EHR Data

Distribution of history and follow-up times Our released EHR data contains all of Stanford's records for each patient in our cohort. As such, we have relatively substantial history before and follow-up time after each CTPA procedure in our cohort. Figure 4 provides the distributions for both the amount of history and the amount of follow-up time in days for our dataset. For the patients who underwent multiple CTPA scans we also calculate some basic statistics of the intervals, shown in Table 9.

Distribution of data types Here we will show types of data (clinical events) in EHR patient timeline, including what OMOP or Clarity table they are from. The OMOP and Clarity table distribution across

Slice Thickness	Count
3.00	8,600
2.50	2,840
2.00	8
1.50	5,366
1.25	7,834
1.00	16,911

Table 7: Slice Thickness Distribution

Manufacturer	Count
SIEMENS	11,357
GE MEDICAL SYSTEMS	3,786
TOSHIBA	3,072

Table 8: CT Scanner Manufacturer Distribution

patients is shown in Figure 5 and Figure 6. We can see the measurement table dominates the OMOP table distribution and is larger than others by an order of magnitude. The Clarity table follows a long-tail power law distribution.

C.5 Model Releases

To aid reproducibility, we release all models trained in our experiments as part of the dataset.

EHR models are in the form of pickle objects, either LightGBM Classifiers for the LightGBM models or sklearn LogisticRegression for the linear probes for MOTOR. MOTOR itself is being released separately at https://huggingface.co/StanfordAIMI.

CT models are saved in the form of PyTorch checkpoints.

D Task Label Definitions And Validation

As part of our project, we developed and validated a set of diagnostic labels for pulmonary embolism and a set of prognostic labels for the risk of pulmonary hypertension for every case in our dataset.

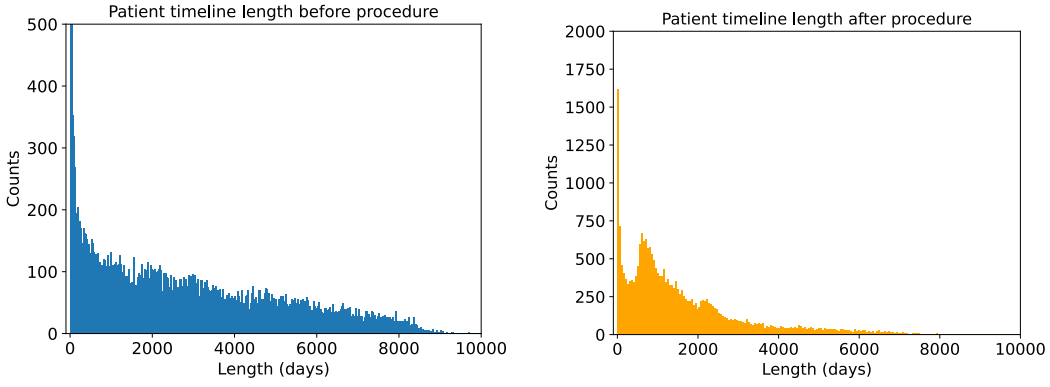

Figure 4: Patient timeline length distributions

	Min	Max	Average	Standard deviation
Interval (in days)	0	6887	448.19	764.87

Table 9: Statistics of CTPA scans intervals across all patients

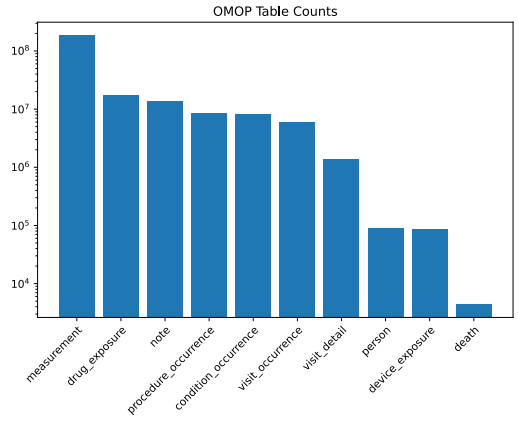

Figure 5: OMOP table distribution

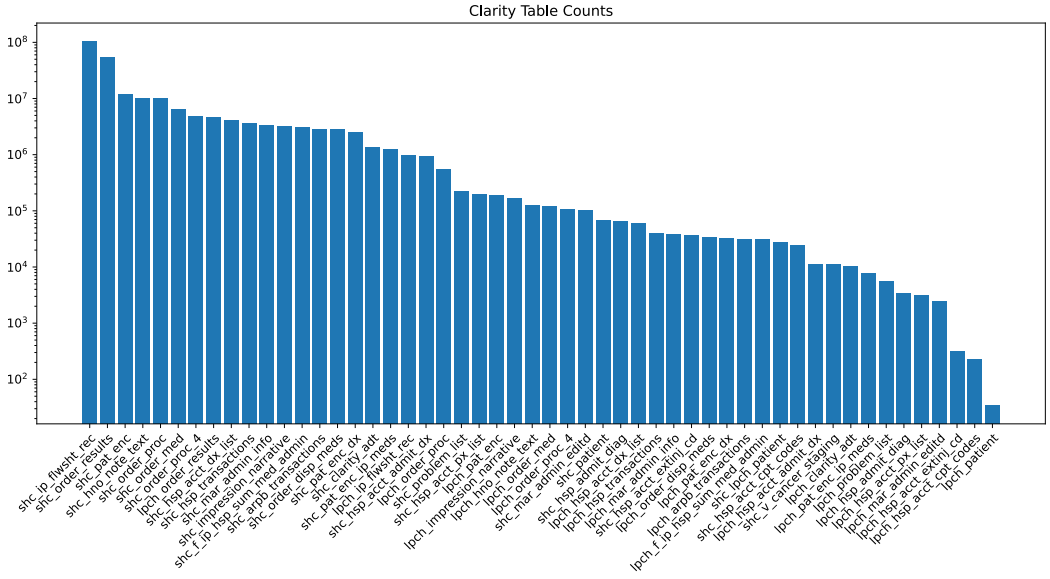

Figure 6: Clarity table distribution

D.1 Pulmonary Embolism

Task Label Definition

We construct three sets of pulmonary embolism labels ("Positive PE", "Subsegmental PE", and "Acute PE"). All the primary analysis in the benchmark is done using "Positive PE", but we release all three as part of our dataset. We define these labels using the impression section of radiology reports as ground truth. If the radiology report contains evidence of the label, we consider that to be a positive example. If the radiology report is either unclear or contains evidence against the label, we consider that a negative. We shifted the prediction time of PE to 24 hours before the CTPA exam time to avoid feature leakage, following [30].

To develop our NLP PE labeler, we use the annotated dataset described in Banerjee et al. 2019 [6]. This dataset includes 4,351 CTPA reports obtained from Stanford Healthcare Center between 2000-2016. All reports were manually annotated by board-certified radiologists according to the following labels:

- **Positive PE**: This label is critical as it signifies the presence of a PE, a potentially life-threatening condition where one or more of the pulmonary arteries in the patient's lungs is blocked by a blood clot. Accurate identification of PE in radiology reports is a crucial step towards timely treatment and patient recovery.
- **Subsegmental PE**: This label indicates a PE that affects the subsegmental branches of the pulmonary arteries, the smaller vessels within the lung. This label is essential in tailoring patient treatment as subsegmental PE sometimes have different treatment protocols compared to PE located in the larger pulmonary arteries. The classification of PE down to the subsegmental level is vital for precision medicine.
- **Acute PE**: This label marks a sudden onset of PE. Acute PE is particularly significant due to its immediate risk to the patient. Rapid identification and treatment of acute PE can mean the difference between life and death. As such, the Acute PE label serves as an urgent signal in the patient's radiology report, prompting immediate medical intervention.

Using these reports and labels, we train a text-based labeling model that can automatically label the impression sections of radiology reports with "Positive PE", "Subsegmental PE", and "Acute PE" labels. We utilize a pretrained version of the Clinical Longformer model [45], as the backbone of our NLP labeler. This model is then finetuned, validated, and tested using our hand-labeled cases. After finetuning, this model is applied to generate labels for every case in our dataset, with the "Positive PE" label in particular used for all analyses. Each label was deemed positive if the model's prediction probability exceeded 0.5; otherwise, the label was classified as negative. The labeling process is shown in Figure 7.

Task Label Validation

We validate our PE NLP labels using the test set of the manual labels. The performance of the model can be found in Table 10. The precision and recall are quite high, especially for the main "Positive PE" label, indicating that our NLP-generated labels are high quality.

Even with satisfactory performance, we are interested to know the error modes of our NLP labeler so we manually examine the predictions of it against the ground truth. The confusion matrices are shown in Figure 8. When comparing against human annotation on notes, for false positive cases, we observed that the NLP might have mistakenly used the wording '*...thrombus within the superior vena cava...*' as an indicator of positive pulmonary embolism when it is not. For false negative cases, the NLP labeler might have incorrectly used the wordings '*...possible pulmonary emboli are incompletely evaluated. Consider ct pe protocol if clinically indicated...*' as positive PE. Overall the misclassifications when compared to experts' annotated notes are relatively low (12 out of 682).

D.2 Pulmonary Hypertension

Task Label Definition

We construct a set of prognostic pulmonary hypertension labels that attempt to define whether or not a patient has a future incidence of pulmonary hypertension in the next year.

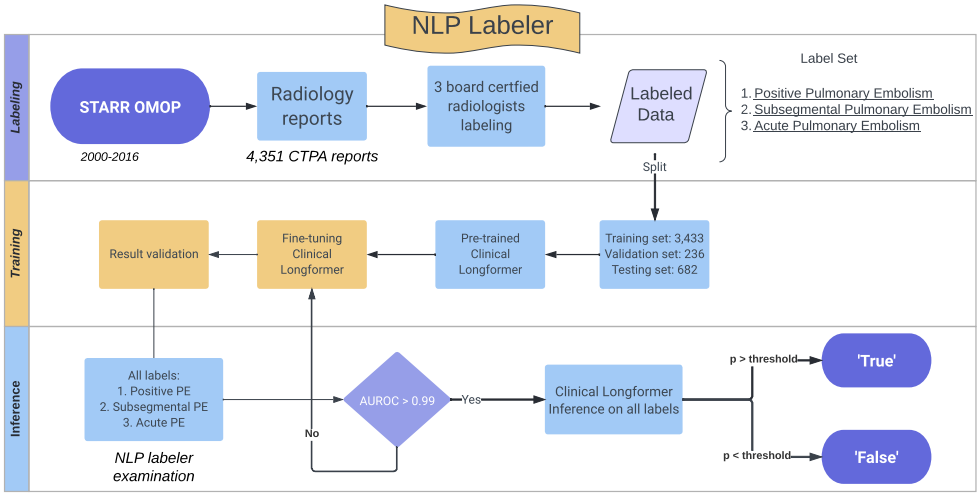

Figure 7: A flowchart of our NLP labeling process.

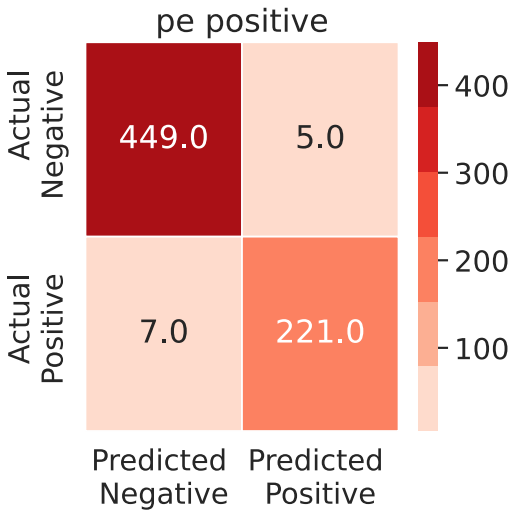

Figure 8: Confusion matrix between NLP labeler and human experts on notes under test set of our training data for NLP labeler

	Positive PE	**Subsegmental PE**	**Acute PE**
AUROC	0.99	0.99	0.99
F1	0.97	0.95	0.96
Precision	0.97	0.98	0.98
Recall	0.98	0.93	0.94
Accuracy	0.98	0.99	0.98
Class counts (pos)	228	43	201
Class counts (neg)	454	639	481
Total support	682	682	682

Table 10: NLP PE labeler performance. Each label was deemed positive if the model's prediction probability exceeded 0.5; otherwise, the label was classified as negative.

	# Patients
Positive PH	97
Negative / Unknown PH	23

Table 11: Statistics for our ground truth hand-labeled pulmonary hypertension labels.

We start by having a board-certified clinician create a manually annotated label set for this task on 120 patients in our cohort. We define ground truth for this task based on the review of a subset of notes for those patients. Each of these notes is either labeled "Positive" or "Negative", where "Positive" is that the patient has pulmonary hypertension and "Negative" is that the patient either doesn't have pulmonary hypertension or it is unknown. We then aggregate these labels at the patient level, labeling a patient with any "Positive" label as "Positive" and "Negative" otherwise. The statistics for these labels are in table 11.

Hand-labeling all of the cases in our dataset is not viable so we use this seed set of manual labels to develop a structured data-based phenotyping algorithm that can then be applied to all of the patients in our dataset. From manual review and expert assistance, we derive Table 12 which contains a comprehensive list of ICD and internal Stanford codes that can be used to identify pulmonary hypertension. This phenotyping algorithm is applied to obtain the pulmonary hypertension labels that we use for our primary analysis.

Table 12: Concepts of pulmonary hyptertension and their ICD9/10 codes

Concept Name	Vocabulary ID	Code
Pulmonary hypertension	STANFORD_CONDITION	1029634
Secondary pulmonary arterial hypertension	ICD10CM	I27.21
Pulmonary hypertension due to left heart disease	ICD10CM	I27.22
Chronic pulmonary heart disease	ICD9CM	416
Kyphoscoliotic heart disease	ICD9CM	416.1
Chronic pulmonary embolism	ICD9CM	416.2
Other secondary pulmonary hypertension	ICD10	I27.2
Other secondary pulmonary hypertension	ICD10CM	I27.29
Eisenmenger's syndrome	ICD10CM	I27.83
Primary pulmonary hypertension	ICD10	I27.0
Primary pulmonary hypertension	ICD9CM	416.0
Other chronic pulmonary heart diseases	ICD9CM	416.8
Other specified pulmonary heart diseases	ICD10CM	I27.89
Chronic pulmonary embolism	ICD10CM	I27.82
Pulmonary hypertension due to alveolar hypoventilation disorder	STANFORD_CONDITION	1170535
Kyphoscoliotic heart disease	ICD10CM	I27.1
Pulmonary hypertension, unspecified	ICD10CM	I27.20
Pulmonary hypertension due to lung diseases and hypoxia	ICD10CM	I27.23
Chronic pulmonary heart disease, unspecified	ICD9CM	416.9
Cor pulmonale (chronic)	ICD10CM	I27.81
Pulmonary heart disease, unspecified	ICD10CM	I27.9
Pulmonary heart disease, unspecified	ICD10	I27.9
Primary pulmonary hypertension	ICD10CM	I27.0
Kyphoscoliotic heart disease	ICD10	I27.1
Secondary pulmonary hypertension	STANFORD_CONDITION	67294
Chronic pulmonary heart disease (CMS-HCC)	STANFORD_CONDITION	142308
Chronic thromboembolic pulmonary hypertension	ICD10CM	I27.24
Other secondary pulmonary hypertension	STANFORD_CONDITION	2065632
Other secondary pulmonary hypertension	ICD10CM	I27.2

Task Label Validation

We validate our pulmonary hypertension phenotyping algorithm by testing it using the hand-labeled set. Our hand labels don't incorporate time, so we can't directly compare the 12-month PH task used in our analysis to them. Instead, we compare a slightly modified algorithm that uses the same

	Positive PH
F1	0.88
Precision	0.85
Recall	0.91
Accuracy	0.80

Table 13: Structured data-based PH labeler performance.

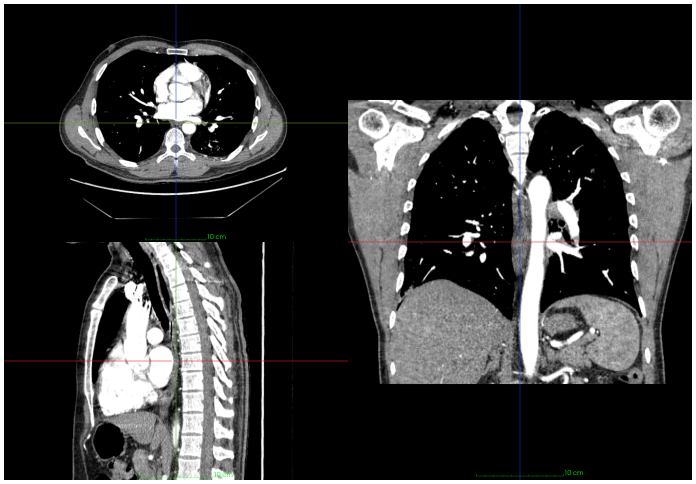

Figure 9: An example of CTPA examination scan in multi-planar reconstruction

ICD/Stanford code list to identify patients who have ever had PH and compare that set of patients to the set of "Positive" patients in our hand-labeled set. The precision, recall, F1 and accuracy are in Table 13. Our phenotyping algorithm has a very high recall, of 0.91, with slightly worse precision. The reduced precision is likely due to how we only hand labeled a subset of notes, so our ground truth here has poor recall. Regardless, this demonstrates that the structured data phenotyping algorithm we are using is effective.

E Example of CTPA scan

For the readers who are unfamiliar with CTPA, we also attached an example in Figure 9. This scan demonstrates an MPR (multi-planar reconstruction) format rendering of the 3D volumetric CTPA scan from our INSPECT cohort.

F Additional Model Details And Hyperparameters

Hyperparameters are selected through grid search on the validation set. Table 14 contains the hyperparameter grids, and the software versions used for each model.

Table 14: Hyperparmeter search grids of methods under comparison in our experiments. The software version for implementing each method is also shown.

Hyperparameters	**Values**
LightGBM	
max_depth	3, 6, -1
learning_rate	0.02, 0.1, 0.5
num_leaves	10, 25, 100
software_version	LightGBM 3.3.5
MOTOR	
linear_probe_l2_strength	automatic between 10 and 0
dropout	0
learning_rate	10^{-5}
num_time_bins	8
survival_dim	512
inner_dim	768
layers	12
max_sequence_length	16,384
vocabulary_size	65,536
software_version	femr 0.1.8
CTPA Slice Encoder	
learning_rate	0.0005
optimizer	AdamW
loss	BCEWithLogitsLoss
architecture	resnext101_32x8d
pretrain data	BigTransfer
software_version	timm 0.9.2
CTPA Sequence Encoder	
learning_rate	0.001, 0.0005, 0.0001, 0.00005
n_epochs	50
slice aggregation	max, mean, attention, attention+max
sequence encoder type	LSTM, GRU, Transformer
hidden size	128, 256, 512
bidirectional	True, False
num_layers	1, 3, 5
dropout_prob	0.0, 0.25, 0.5, 0.75
weighted_sampling	True, False
pretrain data	RSNA RESPECT
input_sie	256
PE NLP Labeler	
max_sequence_length	1536
learning_rate	2e-5
n_epochs	15
architecture	Longformer
pretrain type	Clinical-Longformer
software_version	hugging face 4.30.1

F.1 CTPA model

Windowing We here begin to describe the viewing window for our CTPA imaging model. (window center = -600, window width = 1500), pulmonary embolism (window center = 400, window width = 1000), and the mediastinum (window center = 40, window width = 400). Specifically, for each viewing window, we clipped the Hounsfield Unit (HU) pixel values to fall within the range $[windowcenter - windowwidth/2, windowcenter + windowwidth/2]$

Slice Encoder Augmentations After the windowing operation, every CT scan is resized to dimensions of 256x256 followed by a random cropping operation to yield a 224x224 size. Before inputting into the model, each slice is normalized using the mean and standard deviation values from ImageNet. Once the slice encoder training phase is concluded, each slice is inputted into the trained model for the extraction of a latent representation. In this phase, center cropping is applied as opposed to random cropping for retrieving slice representations.

Sequence Encoder Augmentations Before the slice representations are input into a sequence encoder, we ensure each series is standardized to the same input size through either random sampling or padding. Specifically, if a series has a higher slice count than num_slices, a random sampling of the slices is conducted to equalize with num_slices. Alternatively, if a series possesses fewer slices, padding is executed with zero vectors to complete the series.

F.2 Structured Electronic Health Records Models

Gradient Boosted Tree Model

For our featurization, we use count featurization augmented by ontologies. For count featurization, we count each occurrence before the prediction time of every medical code (diagnoses, procedures, lab orders, and medications) and have a column containing the count for each code. Normally, this is a very sparse matrix as each code individually is relatively rare, so we take advantage of the standard *ontology expansion* technique, where we count higher level concepts in addition to the raw codes themselves. For instance, we will have a column both for the number of very specific ICD/I27.29 codes as well as a column for the more generic ICD/IXX (and I class ICD code) concept.

These features are then fed into a hyperparmeter tuned LightGBM model [35].

MOTOR Model

MOTOR [62] is a self-supervised transformer model designed for long-term medical prediction. For our experiments, we use a version that was already pretrained on de-identified Stanford data. We explicitly construct our training, validation, and test cohorts in sync with that pretrained model such that there is no overlap between its pretraining data and our test and validation data.

We use the linear probe method for adapting MOTOR to our tasks. Aka, we extract the final patient representation from the last transformer layer and then train a logistic regression model with L2 regularization on those representations.

F.3 Model Fusion

For model fusion, we apply a simple late fusion strategy of taking a weighted average of the outputs of each source model. We implement this by first converting all output probabilities to logits, and then fitting a logistic regression model on those logits using the validation set. We do not use any regularization for that logistic regression model as it only has at most 3 features in our setup.

Furthermore, we examine the agreement and disagreement between the three source models by computing the Spearman correlations between their output probabilities on the 8 tasks in our dataset. Figure 10 contains the corresponding heatmaps. As expected, the two EHR based models, MOTOR and GBM, are much more correlated with each other than the CT based model.

G Experiment Compute Environment

EHR experiments are performed in a local on-prem university compute environment using 24 Intel Xeon 2.70GHz CPU cores and 1 Nvidia V100 GPU.

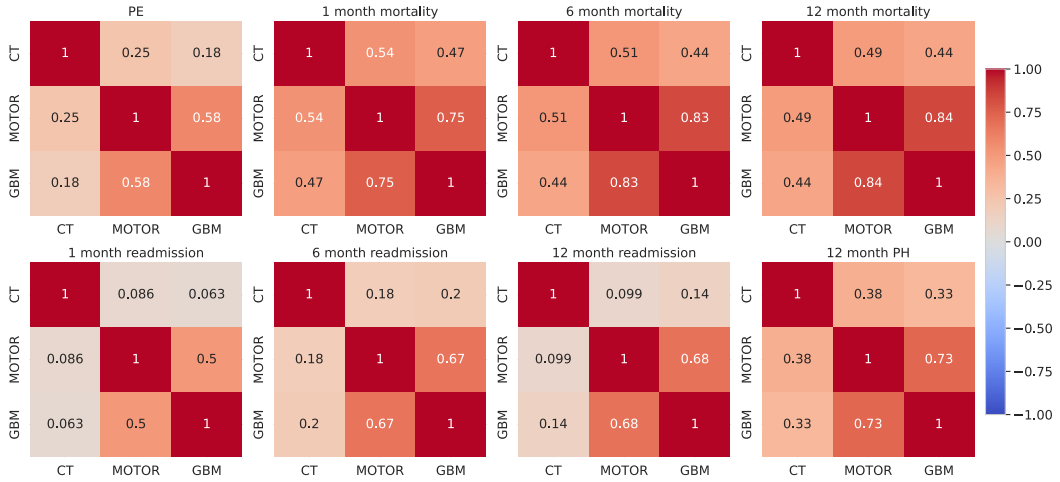

Figure 10: Spearman correlation matrices for each pair of models' output probabilities. **CT** is the CTPA based LRCN model, **M** is the structured EHR based MOTOR model, and **G** is the structured EHR based gradient-boosted trees model.

Input Modality			**Diagnostic**	**Prognostic**						
Image	**EHR**		**PE**	**In-Hospital Mortality**			**Re-admission**			**PH**
CT	M	G	(+)	1 m	6 m	12 m	1 m	6 m	12 m	12 m
✓			(0.69, 0.75)	(0.76, 0.83)	(0.73, 0.78)	(0.72, 0.77)	(0.50, 0.60)	(0.48, 0.55)	(0.49, 0.56)	(0.63, 0.69)
	✓		(0.66, 0.70)	(0.91, 0.94)	(0.89, 0.92)	(0.88, 0.91)	(0.73, 0.81)	(0.75, 0.81)	(0.74, 0.79)	(0.80, 0.85)
		✓	(0.66, 0.71)	(0.82, 0.87)	(0.85, 0.88)	(0.84, 0.87)	(0.69, 0.78)	(0.71, 0.77)	(0.70, 0.75)	(0.80, 0.85)
✓	✓		(0.74, 0.78)	(0.91, 0.94)	(0.89, 0.92)	(0.88, 0.91)	(0.73, 0.82)	(0.75, 0.80)	(0.74, 0.79)	(0.80, 0.84)
✓		✓	(0.74, 0.79)	(0.84, 0.89)	(0.86, 0.89)	(0.85, 0.88)	(0.69, 0.78)	(0.71, 0.76)	(0.70, 0.75)	(0.81, 0.85)
	✓	✓	(0.68, 0.72)	(0.91, 0.94)	(0.89, 0.92)	(0.88, 0.91)	(0.74, 0.82)	(0.76, 0.81)	(0.75, 0.80)	(0.82, 0.87)
✓	✓	✓	(0.75, 0.79)	(0.91, 0.94)	(0.89, 0.92)	(0.88, 0.91)	(0.74, 0.82)	(0.76, 0.81)	(0.75, 0.79)	(0.82, 0.86)

Table 15: 95% confidence intervals as a function of the test set for model performance in AUROC. **CT** is the CTPA based LRCN model, **M** is the structured EHR based MOTOR model, and **G** is the structured EHR based gradient-boosted trees model.

Image experiments are performed on a HIPAA-compliant Google virtual machine using 4 x Nvidia V100 GPU with 96 Intel Skylake vCPU with 624GB of RAM.

All compute environments supported HIPAA-compliant data protocols.

H Confidence Intervals

We obtain some uncertainty estimates for our results by bootstrapping with respect to the test set.

First, we estimate the uncertainty of the AUROC of each model. For each task, we create 1,000 bootstrap samples and compute the AUROC for each model on each bootstrap sample. We then extract the 2.5% and 97.5% percentiles of the 1,000 samples to obtain 95% confidence intervals.

Table 15 presents the results of this analysis. The widths of the intervals are often around 0.04, indicating that we are able to estimate model performance with reasonable precision.

Second, we estimate the uncertainty of the relative AUROC of each model. We use the same bootstrap samples as in the first analysis, but compute the relative performance between each model and our chosen baseline, which we arbitrarily choose as MOTOR.

Table 16 contains these relative confidence intervals. Most (12/14) of the intervals for individual models exclude zero, indicating that we have enough precision to accurately tell the difference in performance between CT, MOTOR, and gradient-boosted tree models. The fused models have a less clear separation, with about half of the differences being statistically insignificant.

Input Modality			Diagnostic	Prognostic						
Image	EHR		PE	In-Hospital Mortality			Re-admission			PH
CT	M	G	(+)	1 m	6 m	12 m	1 m	6 m	12 m	12 m
✓			**(0.01, 0.07)**	**(-0.16, -0.10)**	**(-0.17, -0.12)**	**(-0.17, -0.12)**	**(-0.29, -0.16)**	**(-0.30, -0.22)**	**(-0.28, -0.20)**	**(-0.19, -0.13)**
	✓		(0.00, 0.00)	(0.00, 0.00)	(0.00, 0.00)	(0.00, 0.00)	(0.00, 0.00)	(0.00, 0.00)	(0.00, 0.00)	(0.00, 0.00)
		✓	(-0.02, 0.03)	**(-0.10, -0.05)**	**(-0.05, -0.02)**	**(-0.05, -0.02)**	(-0.08, 0.01)	**(-0.06, -0.02)**	**(-0.06, -0.02)**	(-0.02, 0.03)
✓	✓		**(0.06, 0.10)**	(-0.00, 0.01)	(-0.00, 0.01)	(-0.00, 0.01)	(-0.00, 0.00)	(-0.00, 0.00)	(-0.01, 0.00)	(-0.01, 0.00)
✓		✓	**(0.06, 0.11)**	**(-0.08, -0.03)**	**(-0.04, -0.01)**	**(-0.04, -0.01)**	(-0.08, 0.01)	**(-0.07, -0.02)**	**(-0.07, -0.02)**	(-0.01, 0.03)
	✓	✓	**(0.01, 0.03)**	(-0.00, 0.00)	(-0.00, 0.00)	(-0.00, 0.00)	(-0.00, 0.02)	**(0.00, 0.01)**	**(0.00, 0.01)**	**(0.01, 0.04)**
✓	✓	✓	**(0.07, 0.11)**	(-0.00, 0.01)	(-0.00, 0.01)	(-0.00, 0.01)	(-0.00, 0.02)	**(0.00, 0.01)**	(-0.00, 0.01)	**(0.00, 0.04)**

Table 16: 95% confidence intervals for the difference in AUROC performance between a particular model and the structured EHR based MOTOR model. **CT** is the CTPA based LRCN model, **M** is the structured EHR based MOTOR model, and **G** is the structured EHR based gradient-boosted trees model. Statistically significant differences at p = 0.05 are **bolded**.

Input Modality			Diagnostic	Prognostic						
Image	EHR		PE	In-Hospital Mortality			Re-admission			PH
CT	M	G	(+)	1 m	6 m	12 m	1 m	6 m	12 m	12 m
✓			0.715, (0.003)	0.741, (0.007)	0.753, (0.001)	0.750, (0.003)	0.549, (0.008)	0.547, (0.014)	0.551, (0.012)	0.658, (0.005)

Table 17: The mean and standard deviation in AUROC for the various tasks when the random seed is changed. We use 5 random seeds to estimate both the mean and standard deviation.

In order to conduct variation study, we have rerun our image-based modality for 5 times for different seeds. Note that our EHR baselines, MOTOR and LightGBM, are deterministic with the hyperparameters we used in our study, so we do not perform reseeding experiments. The results of this analysis are in Table 17.

I Model Performance as a Function of PE Status

In our results section, we present performance statistics on the entire cohort. However, it is sometimes useful to look at performance within patients who test positive for PE (+) vs patients who test negative for PE (-). Table 18 contains the performance on the seven prognostic tasks by PE status.

J Comparison of Models vs. Simplified PESI

Clinical risk scores are heuristics commonly used in medicine to inform treatment decisions. We compare our machine learning-based models against a common PE rule-based risk calculator, the simplified PESI (sPESI) score [55]. sPESI is a 0-6 scoring rule comprised of the following additive criteria (each rule contributes +1 to the overall score) in Table 19.

To ensure that the features used to calculate sPESI reflect the patient's condition at the time of the imaging, we only use data between the most recent 10 days prior to the CT exam and the 2 days after the CT scan for the numeric sPESI features. Patients that are missing data required for sPESI are dropped. In addition, as sPESI is only designed for use with patients that have PE, we further restrict this analysis to patients that have a positive diagnosis for PE. This results in a total of 1,719 cases derived from 1,609 unique patients with the required sPESI features and PE. The amount of patients with each total score is listed in Table 20.

We evaluate the sPESI score by measuring the performance in terms of AUROC of using the score to rank patients for our seven prognostic tasks. Table 21 contains the results of this comparison. Note that sPESI is designed for short-term mortality prediction, and might not be meaningful in the context of other prognostic tasks. We observe relatively low performance for the sPESI. One potential cause of that low performance is that our retrospective data has a much higher degree of missingness than the prospective studies used to generate and validate sPESI. For example, our ability to extract features like the history of Chronic Cardiopulmonary Disease is relatively limited as we are restricted to structured data already within the health record.

Has PE	Input Modality			Prognostic						
	Image	EHR		In-Hospital Mortality			Re-admission			PH
	CT	M	G	1 m	6 m	12 m	1 m	6 m	12 m	12 m
	✓			0.761	0.738	0.726	0.609	0.586	0.629	0.596
		✓		0.914	0.897	0.869	0.782	0.770	0.755	0.761
			✓	0.853	0.850	0.835	0.773	0.763	0.729	0.762
(+)	✓	✓		0.879	**0.902**	0.870	0.752	0.766	**0.760**	0.740
	✓		✓	0.817	0.858	0.844	0.712	0.748	0.732	0.740
		✓	✓	**0.914**	0.897	**0.871**	**0.788**	**0.789**	0.757	**0.772**
	✓	✓	✓	0.879	0.899	0.870	0.762	0.776	0.758	0.752
	✓			0.806	0.758	0.754	0.534	0.504	0.507	0.677
		✓		0.925	0.902	0.897	0.771	0.782	0.770	0.852
			✓	0.847	0.869	0.861	0.728	0.737	0.728	0.852
(-)	✓	✓		**0.927**	0.903	0.900	0.771	0.778	0.764	0.850
	✓		✓	0.872	0.879	0.873	0.729	0.728	0.716	0.859
		✓	✓	0.924	0.905	0.898	**0.775**	**0.787**	**0.775**	**0.871**
	✓	✓	✓	0.926	**0.905**	**0.901**	0.775	0.783	0.768	0.868

Table 18: The performance in AUROC for our different baseline modeling strategies, split by PE status. **CT** is the CTPA based LRCN model, **M** is the structured EHR based MOTOR model, and **G** is the structured EHR based gradient-boosted trees model. The best overall models are **bolded** and the best individual models are underlined.

#	PESI Score criteria
1	Age > 80
2	History of Cancer
3	History of Chronic Cardiopulmonary Disease
4	Heart Rate (bpm) $\geq$ 110
5	Systolic BP (mmHg) < 100
6	O_2 Saturation < 90%

Table 19: Criteria for PESI score

sPESI Score	# Cases With Score
0	169
1	361
2	529
3	450
4	184
5	30
6	1

Table 20: The statistics for the sPESI scores on the 1,719 cases in our cohort that it can be calculated on.

Input Modality				Prognostic						
Image	EHR			In-Hospital Mortality			Re-admission			PH
CT	M	G	P	1 m	6 m	12 m	1 m	6 m	12 m	12 m
✓				0.676	0.634	0.663	0.643	0.478	0.631	0.579
	✓			**0.808**	**0.813**	**0.787**	**0.745**	**0.684**	**0.678**	**0.725**
		✓		0.749	0.749	0.741	0.679	0.620	0.619	0.688
			✓	0.569	0.571	0.571	0.701	0.679	0.614	0.567

Table 21: The performance in AUROC for our different baseline modeling strategies given patients with PE. Note that this set of evaluations is only done on the subset of cases that have both PE and enough data for the simplified PESI risk score. **CT** is the CTPA based LRCN model, **M** is the structured EHR based MOTOR model, **G** is the structured EHR based gradient-boosted trees model, and **P** is the simplified PESI risk score. The best models are **bolded**.

K Additional Metrics

For our main analysis, we compare models in terms of AUROC as it is a low variance and widely used metric. However, additional metrics, especially in the clinical space, can also be important when evaluating the utility of models.

Input Modality			Diagnostic	Prognostic						
Image	EHR		PE	In-Hospital Mortality			Re-admission			PH
CT	M	G	(+)	1 m	6 m	12 m	1 m	6 m	12 m	12 m
✓			<u>0.463</u>	0.189	0.288	0.324	0.056	0.124	0.169	0.230
	✓		0.327	<u>0.396</u>	<u>0.537</u>	<u>0.588</u>	<u>0.160</u>	<u>0.342</u>	<u>0.402</u>	0.485
		✓	0.335	0.234	0.426	0.497	0.145	0.276	0.334	<u>0.582</u>
✓	✓		0.510	0.426	**0.545**	**0.599**	0.164	0.337	0.393	0.481
✓		✓	0.515	0.295	0.447	0.521	0.145	0.271	0.330	0.573
	✓	✓	0.354	0.399	0.542	0.587	0.179	**0.346**	**0.407**	**0.597**
✓	✓	✓	**0.523**	**0.428**	0.542	0.598	**0.180**	0.343	0.399	0.589

Table 22: The performance in AUPRC for our different baseline modeling strategies, including late fusion. **CT** is the CTPA based LRCN model, **M** is the structured EHR based MOTOR model, and **G** is the structured EHR based gradient-boosted trees model. The best overall models are **bolded** and the best individual models are <u>underlined</u>.

We thus perform additional analysis to compare our models in terms of both AUPRC (area under the precision-recall curve) (see Table 22 and ECE (expected calibration error) (see Table 23). AUPRC provides an estimate of the precision of a model under various recall thresholds and ECE provides an estimate of the calibration of a model. We use 10 bins for our ECE estimate.

The relative model performance rankings for both of these additional metrics are very similar to the rankings seen with AUROC, with the EHR models doing better at prognostic tasks while the image models do better at the diagnostic task.

Input Modality			Diagnostic	Prognostic						
Image	**EHR**		**PE**	**In-Hospital Mortality**			**Re-admission**			**PH**
CT	M	G	(+)	1 m	6 m	12 m	1 m	6 m	12 m	12 m
✓			0.278	0.423	0.369	0.265	0.136	0.347	0.346	0.325
	✓		0.026	0.011	0.010	0.015	0.008	**0.004**	**0.007**	0.012
		✓	0.015	0.020	0.053	0.017	0.016	0.016	0.026	0.021
✓	✓		0.024	**0.004**	0.012	0.017	0.009	0.009	0.010	0.016
✓		✓	0.015	0.007	0.019	0.017	**0.004**	0.009	0.014	0.023
	✓	✓	**0.007**	0.004	**0.010**	**0.014**	0.006	0.009	0.014	**0.009**
✓	✓	✓	0.016	0.005	0.011	0.015	0.006	0.009	0.013	0.025

Table 23: The calibration performance in ECE for our different baseline modeling strategies, including late fusion. Lower scores indicate better models with this metric. **CT** is the CTPA based LRCN model, **M** is the structured EHR based MOTOR model, and **G** is the structured EHR based gradient-boosted trees model. The best overall models are **bolded** and the best individual models are underlined.