Prediction of Imaging Outcomes from Electronic Health Records: Pulmonary Embolism Case-Study

Imon Banerjee, Ph.D.[1,2], Miji Sofela, MS.[3], Timothy Amrhein, MD. [4], Daniel L. Rubin, MD. [2,1], Roham Zamanian, MD. [5], Matthew P. Lungren, MD.[2]

[1]Department of Radiology, Stanford University, Stanford, CA; [2]Department of Biomedical Data Science, Stanford University, Stanford, CA; [3]Duke University Health System, Duke University School of Medicine, NC, USA; [4]Department of Neuroradiology, Duke University School of Medicine, NC, USA; [5]Department of Medicine - Med/Pulmonary and Critical Care Medicine, Stanford University, Stanford, CA;

email - imonb@stanford.edu

Introduction

Pulmonary embolism (PE) is a life-threatening clinical problem and CT imaging is the current gold standard for diagnosis. In the past years, a substantial rise in the number of CT examinations for PE evaluation has been observed with a concomitant decrease in imaging yield (as low as 2 – 3% yield). Unnecessary use of CT not only carries risks due to radiation and intravenous contrast, but also the discovery of low impact incidental imaging findings which further expose patients to unneeded procedures, tests, and risks. Clinical decision support rules based on PE risk scoring models to inform CT imaging decisions have been developed but are underutilized, in part due to evolving risk factors for PE that are not included in static scoring systems (Wells, PERC, rGeneve). The purpose of this study is to design and evaluate a machine learning modeling approach for predicting PE imaging outcomes based on patient EMR data captured before the CT exam which includes demographics, vital signs (change from baseline), diagnoses, medications, lab tests, etc. In contrast with the current clinical scoring systems that rely on a very limited set of pre-selected EMR variables, our objective is to build a machine learning solution that can compute a patient-specific risk score for PE by utilizing complex correlation between hundreds of EMR variables without relying on hand crafted feature selection; we evaluate this model on intra- and extra-institutional patient data and compare to existing PE risk scoring systems.

Methods

Data

Internal dataset (SHC) – Using the STAnford medicine Research data Repository (STARR), we randomly retrieved 4512 contrast-enhanced CT chest examinations of adult patients performed between January 1, 1998 and January 1, 2016 at Stanford hospital and clinics (SHC). Each study was manually annotated by three experienced radiologists and assigned two binary class labels (PE present/absent and PE acute/chronic). Inter-rater reliability among the three raters were highly consistent for two categories, "*PE presence*" and "*PE Acute*", with kappa scores of 0.959 and 0.969 respectively. Because we engineered the model to identify acute PE, we dropped chronic cases to generate the final annotated internal cohorts (SHC) of 3,397 annotated PE-CT exams from 3,214 unique patients (1704 women, mean (SD) age, 60.53 (+/- 19.43)). From the EMRs, we also extracted the following phenomic data and time stamps for these patients: (1) all diagnosis codes; (2) all inpatient and outpatient medications (normalized to RxNorm); (3) all laboratory data raw values; (4) all collected vital sign data (i.e. height, weight, BMI, pulse, respiration rate, systolic blood pressure, temperature, etc); (5) all demographics (i.e. age, race, gender).

External dataset (Duke) - As external dataset, we collected contrast-enhanced CT examinations of chest from Duke University Medical Center performed between January 1, 2013 and August 31, 2017. We retrieved similar set of phenomic data of these patients with encounter time-stamp details. The data were normalized according to the standards - RxNorm, IC9 code. In order to create an external annotated dataset for validation, we randomly selected 300 CT exams from Duke and performed manual annotation with the same group of radiologists. After dropping chronic case, the models were validated on 240 unique patients seen at Duke (132 women, mean (SD) age, 70.2 (+/- 14.2)).

Outpatient samples (SHC and Duke) - In addition, we also created separate outpatient dataset for SHC - 100 consecutive patients (67 women, mean (SD) age, 57.74(+/- 19.87)) and Duke - 101 consecutive patients (59 women, mean (SD) age, 73.06(+/- 15.3)), and these cases are independent from the internal SHC and Duke hold-out dataset.

Proposed System:

The PE prediction approach is outlined by the following problem statement: "Given a new patient encounter and access to prior structured EMR data (vitals, demographics, labs, inpatient and outpatient medication, diagnoses) predict the risk of pulmonary embolism (PE)". This framing lends itself to be treated as a probabilistic classification

problem. The proposed workflow parses raw EMR data arranged as a timeline to transform into feature vectors of use in training a machine learning model based on PE imaging outcomes. For each patient, we defined their observation window as the 12 months leading up to a given prediction date (24 window of CT exam). Within the observation window, we created a feature engineering pipeline that computes a vector representation of the EMR snapshot of each patient by considering the temporal sequence within the records. The designed pipeline parse five core components of EMR - (1) all diagnosis codes (except current encounter); (2) all inpatient and outpatient medications; (3) all laboratory data raw values; (4) all collected vital sign data; (5) all demographics. Given the complexity of the EMR data and the requirement of temporality preservation, we carefully designed a EMR feature engineering pipeline able to parse varying types of EMR simultaneously while also tolerant of sparse records (a common limitation).

As machine learner algorithm, we used the same input features and compared between a regularized regression methodology, ElasticNet, and a novel deep learning model (PE Neural model) – an Encoder network with ReLU activation and sigmoid outcome. To optimize the hyperparameters of ElasticNet (regularization rate) and the PE Neural network architecture (number of hidden layers, learning rate, activation function, optimizer, number of epochs, dropout rate), we used Grid search on 10% training SHC training data to reduce overfitting. Trained models are then tested on hold-out intra- and new extra- institutional patient data as test sets as well as tested on separate intra- and extra- institutional outpatient population. We also compared the performance of trained machine learning models against three popular clinical scoring systems for PE- Wells, PERC, and revised Geneva.

Results

The performance of the machine learning models are summarized as AUC-ROC in Table 1. The models are only trained on a sub-sample of the SHC data in order to test the generalizability of the model on a same training population as well as on a different population from another institution. Both machine learning models scored high accuracy on the internal testset of 340 CT exams (PE neural 0.85 and ElasticNet 0.93) while ElasticNet model outperformed the PE Neural model ($p = 0.013$). Both machine learning models' performance dropped on the external duke dataset compared to the internal hold-out SHC testset (PE neural 0.72 and ElasticNet 0.7). However, the AUC-ROC score stayed > 0.7 and both models performed equally well on the external data ($p = 0.165$) which shows the fact that even when trained on the SHC patients the models are generalizable to the Duke patients (Inpatient and Outpatient) population.

Table 1: *Quantitative analysis of the model's performance - measured in-terms of AUC-ROC score*

	AUC-ROC on SHC data	**p-value**	**AUC-ROC on Duke data (external testset)**	**p-value**
Hold-out testing on the internal SHC dataset and external Duke dataset (inpatient and outpatient)				
ElasticNet model	0.93	0.0132	0.7	0.165
PE Neural model	0.85		0.72	
Comparison with clinical scoring systems on outpatients from the internal SHC and external Duke dataset				
Machine learning models				
ElasticNet model	0.73	0.42	0.74	0.011
PE Neural model	0.81		0.81	
Clinical scoring				
Wells score	0.48	N/A	0.51	N/A
PERC Score	0.51		0.6	
rGeneva Score	0.53		0.47	

Given the criteria for usability of the clinical scoring systems for computing pretest probability, we randomly selected 100 outpatient samples from SHC and 100 from Duke, and created another hold-out cohort. We used the models that were trained using SHC patients and tested it on ED patients separately to judge the model performance of the ED cases in parallel with three popular clinical scorings (Table 1). The PE Neural model performed significantly better than the all other models/criteria on the Stanford and Duke hold-out ED patients including the ElasticNet model on the Duke data ($p = 0.01$).

Conclusion

In conclusion, we found that achieving prediction models based on available retrospective structured EMR data can consider multitudes of patient-specific risk factors and dependencies in order to arrive at a PE likelihood recommendation model is possible and these models may be more accurately generalized to new population distributions. Future work is needed in investigating the ideal application of these prediction models for clinical imaging decision support systems in suspected PE and ultimate effect on imaging utilization.