Monitoring AI Performance After Deployment: Demonstrated in Pulmonary Embolism Detection

Alicia Maehara, Student, Marlbourough School
Gordon Guyant; Melody Bounmasanonh; Alexandria Uy; Edward Zaragoza, MD; William Hsu, PhD

Background/Problem Being Solved

Commercially available AI systems have shown promising results in timely detecting pulmonary embolism. However, continuous monitoring of AI systems is necessary for optimal quality of care. This project aims to develop a system for continuously monitoring an imaging AI for detecting PE (Pulmonary Embolism).

Intervention(s)

The system consists of: 1. A data pipeline that retrieves and transforms data from the imaging AI and CT reports from the medical record; 2. An open-source Large Language Model (Llama 3.1 8b) that extracts and structures PE results based on CT reports using a custom prompt; 3. An algorithm that processes and compares results; and 4. A dashboard to facilitate interpretation of the data and enable case review (Figure 1).

Barriers/Challenges

CT Pulmonary Angiogram has the highest specificity and sensitivity for detecting PE compared to V/Q scan, D-dimer, and compression ultrasound. Therefore, we develop a system to extract and structure PE results (PE and incidental PE AI) from contrast-enhanced CT reports by customizing Llama 3.1 8b and comparing those results with the imaging AI's predictions.

Outcome

During the evaluation period between 07/08/2024 and 11/16/2024, the imaging AI detected 368/23,319 (1.6%) PE cases from CT exams. Concordance with CT report result was 98.8% (23,038/23,319). Positive predictive value was higher in PE CTA exams (PE CTA 84.9% vs PE Incidental 50%). See Table 1.

Conclusion/Statement of Impact/Lessons Learned

We have successfully developed an automated system for continuously monitoring an imaging AI tool for detecting PE and incidental PE. The pipeline is modularized to facilitate ingestion of data from diverse sources. This enables future monitoring of disparate AI systems in different domains where downstream data may be derived from clinical systems, registries, and patient-reported outcomes.

Figure(s)

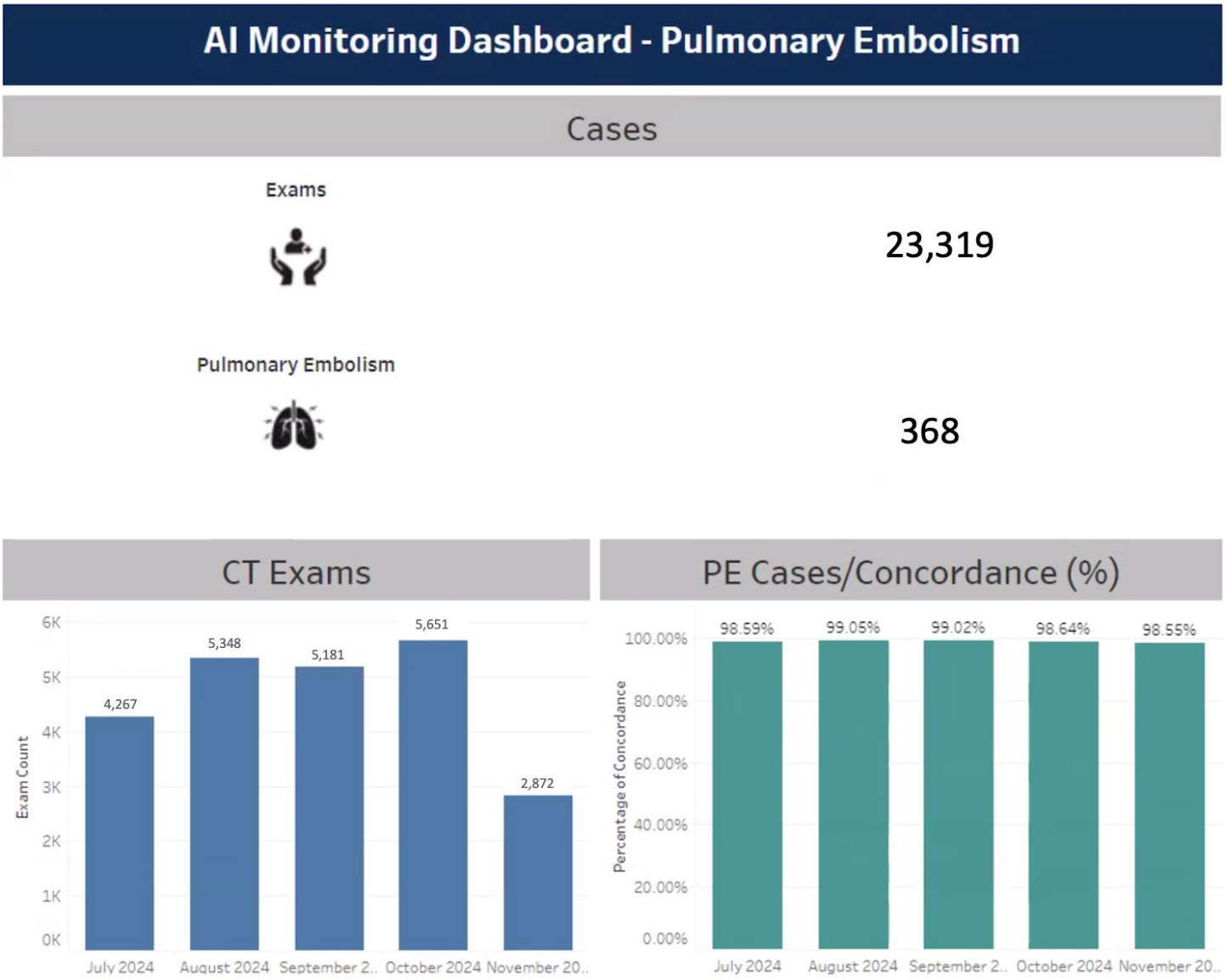

Figure 1. Dashboard showing counts and imaging AI concordance. July exams: 07/08 to 07/31. November exams: 11/01 to 11/16.

		CT reports results			
Imaging AI		PE CTA Positive	PE CTA Negative	Incidental PE Positive	Incidental PE Negative
	PE Positive	180	32	78	78
	PE Negative	70	2,280	101	20,500

Test Parameters		
	PE CTA	Incidental PE
Specificity	98.6%	99.6%
Sensitivity	72.0%	43.6%
Positive Predictive Value	84.9%	50.0%
F-Score	77.9%	46.6%
Accuracy	96.0%	99.1%

Table 1. Comparative effectiveness analytics comparing imaging AI output and CT report results.

Keywords

Artificial Intelligence/Machine Learning; Quality Improvement & Quality Assurance