SIIM CMIMI24
Oct 21-22 | Boston, MA
Conference On Machine Intelligence In Medical Imaging

Beyond FDA Clearance: Automated Post Deployment Monitoring and Validation of Commercial AI Models using Local Large Language Models (LLMs)

Theo Dapamede, MD, PhD; Emory University; Bardia Khosravi, MD, MPH, MHPE; Chad Robichaux, MPH; Aawez Mansuri, MS; Mohammadreza Chavoshi, MD ; Alex Belov; Angela Udongwo, MPH; Chinonyelum Igwe; Frank Li, PhD; Beatrice Brown-Mulry; Hanssen Li, MD; John Moon, MD; Judy Gichoya, MD, MS, FSIIM; Hari Trivedi, MD

Introduction/Background

As AI models are deployed in diverse clinical settings, continuous monitoring and assessment of subgroup performance is critical. Automated techniques to compare radiologist interpretations to model performance must be developed. We used a large language model (LLM) to evaluate the performance of two clinically-deployed commercial AI models for pulmonary embolism and intracranial hemorrhage detection.

Methods/Intervention

We identified 8,966 CT pulmonary embolism exams and 14,637 non-contrast CT head exams conducted between April and October 2023 that were evaluated by the AI model, and extracted the corresponding radiology reports. A locally deployed instance of Llama3 8B was used to extract the PE and ICH labels ground truth labels from the radiology reports, using methods that were previously validated on 500 manually annotated reports (PE: Sn 1.0, Sp: 1.0; ICH: Sn: 0.93, Sp: 1.0). AI model performance was compared to extracted ground truth for multiple subgroups (race, age, sex, and patient location). Overall performance was also compared to the submitted FDA and published performances.

Results/Outcome

For the PE model, sensitivity was 80.3% (95%CI: 77.8% – 83.0%) and specificity was 98.0% (95%CI:97.7% – 98.3%), compared to the published FDA clearance sensitivity of 93.0% (90.2% - 95.1%) and specificity of 93.7% (92.7% - 94.6%). For the ICH model, the sensitivity was 92.2% (91.2%-93.2%) and specificity was 90.3% (89.8%-90.8%), compared to FDA clearance sensitivity of 93.6% (86.6%-97.6%) and specificity of 92.3% (85.4%-96.6%). Both models demonstrated the lowest performance for outpatients as compared to emergency and inpatients, with sensitivities of 77.5% (58.8%-85.0%) and 87.4% (76.8%-95.5%) for PE and ICH models, respectively. Both models demonstrated equitable performance across race, ethnicity, age, and sex subgroups.

Conclusion

We have shown the potential use of LLMs as an automated method for post deployment monitoring and evaluation of clinical AI models. It is notable that the lowest-performing group for both models was outpatients, where advanced detection models can potentially provide the most benefit. Further work and reader studies are required to understand model failure modes and confounders.

Statement of Impact

This study demonstrates a potential automated solution for post deployment monitoring of clinical AI models, which is necessary for ensuring safe and stable model performance after deployment.

A. ICH Model Performance

	TP	TN	FP	FN	Sn	Sp.	PPV	NPV	F1
Overall Performance	2330	10942	1168	197	0.92	0.90	0.67	0.98	0.77
Sex									
Male	1284	4610	607	108	0.92	0.88	0.68	0.98	0.78
Female	1046	6330	561	89	0.92	0.92	0.65	0.99	0.76
Race									
Black or African American	1078	6320	572	86	0.93	0.92	0.65	0.99	0.78
Asian	146	387	75	9	0.94	0.84	0.66	0.98	0.78
Other	281	776	102	24	0.92	0.88	0.73	0.97	0.8q
White	825	3459	419	78	0.91	0.89	0.66	0.98	0.77
Patient Location									
Emergency	411	8264	398	70	0.85	0.95	0.51	0.99	0.64
Inpatient	1796	1894	660	100	0.95	0.74	0.73	0.95	0.83
Outpatient	91	511	86	20	0.82	0.86	0.51	0.96	0.63
Age Group									
20-40 yo	257	2097	129	24	0.92	0.94	0.67	0.98	0.77
41-60 yo	687	2970	335	48	0.94	0.90	0.67	0.98	0.78
61-80 yo	1056	4064	522	95	0.92	0.89	0.67	0.98	0.77
>80 yo	309	1652	169	30	0.91	0.91	0.64	0.98	0.76

B. PE Model Performance

	TP	TN	FP	FN	Sn	Sp	PPV	NPV	F1
Overall Performance	776	7841	160	189	0.80	0.98	0.83	0.98	0.82
Sex									
Male	319	2997	79	81	0.80	0.97	0.80	0.97	0.80
Female	457	4843	81	108	0.81	0.98	0.85	0.98	0.83
Race									
White	241	2508	56	53	0.82	0.98	0.81	0.98	0.82
Black or African American	478	4515	87	124	0.79	0.98	0.85	0.97	0.82
Asian	12	286	6	4	0.75	0.98	0.67	0.99	0.71
Other	45	532	11	8	0.85	0.98	0.80	0.99	0.83
Patient Location									
Emergency	513	5777	98	122	0.81	0.98	0.84	0.98	0.82
Inpatient	222	1485	49	54	0.80	0.97	0.82	0.96	0.81
Outpatient	20	251	5	8	0.71	0.98	0.80	0.97	0.75
Age Group									
(60.0, 80.0]	335	2990	81	71	0.83	0.97	0.81	0.98	0.82
(20.0, 40.0]	113	1526	23	42	0.73	0.99	0.83	0.97	0.78
(40.0, 60.0]	232	2422	34	48	0.83	0.99	0.87	0.98	0.85
(80.0, 100.0]	94	847	22	26	0.78	0.97	0.81	0.97	0.80

Table 1. Peformance metrics for overall model performance and subgroup performance for a commercially deployed PE detection model (A) and ICH detection model (B) as compared to ground truths extracted using Llama3.

Performance metrics for overall model performance and subgroup performance for a commercially deployed PE detection model (A) and ICH detection model (B) as compared to ground truths extracted using Llama3

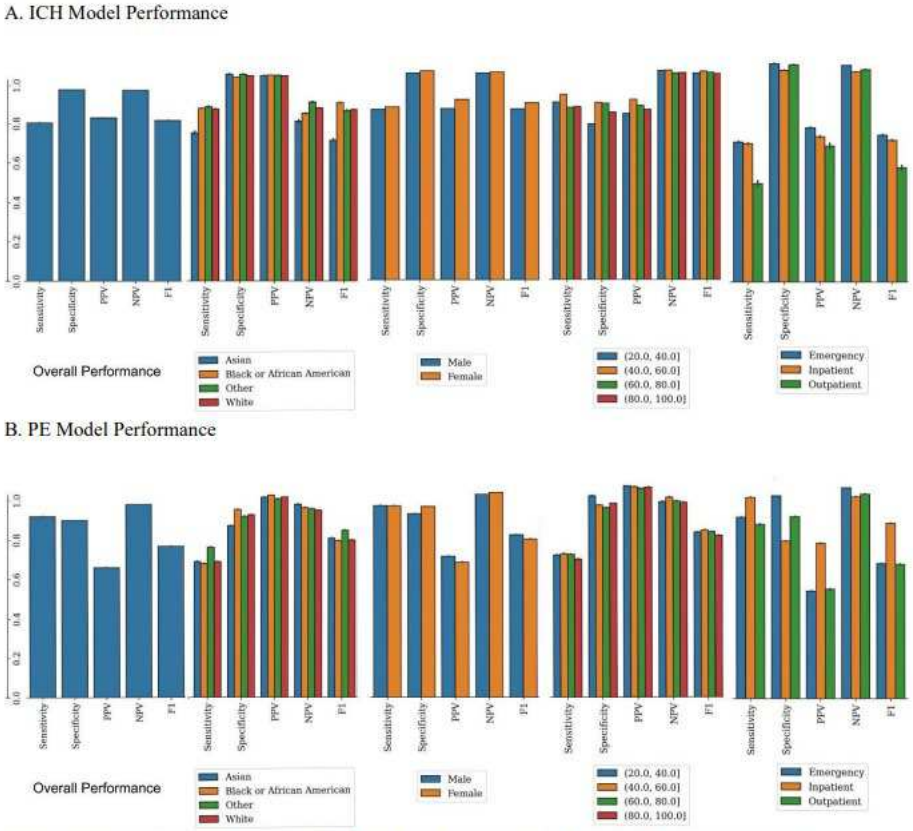

Figure 1. Summary of ICH (A) and PE (B) model performance overall and across various subgroups.

Summary of ICH (A) and PE (B) model performance overall and across various subgroups.

Keywords

Post-Deployment Monitoring; AI Validation; LLM