RSNA 2021
REDEFINING RADIOLOGY
November 28 to December 2

Implementation of a Regular Performance Monitoring Framework for Chest Radiograph Classification at a Radiology Department

Ivan HO MIEN[1,3], Feri GURETNO[1], Oliver NICKALLS[2], Steven WONG[2], Lux ANANTHARAMAN[1], Pavitra KRISHNASWAMY[1]

1. Institute for Infocomm Research, A*STAR, Singapore
2. Sengkang General Hospital, Singapore
3. National Neuroscience Institute, Singapore

Institute for Infocomm Research
I²R

Sengkang General Hospital
SingHealth

Why monitor AI model performance?

- Objectives:
  - Detect data drift and AI model performance deviation
  - Improve audit efficiency by automation
- Impact:
  - Ensures quality and reliability of AI model
  - Saves resources by automating laborious process

- Performance monitoring lets us audit AI results regularly
- Informs need for model refresh, revision, or removal in a timely manner
- Enforces human oversight over AI

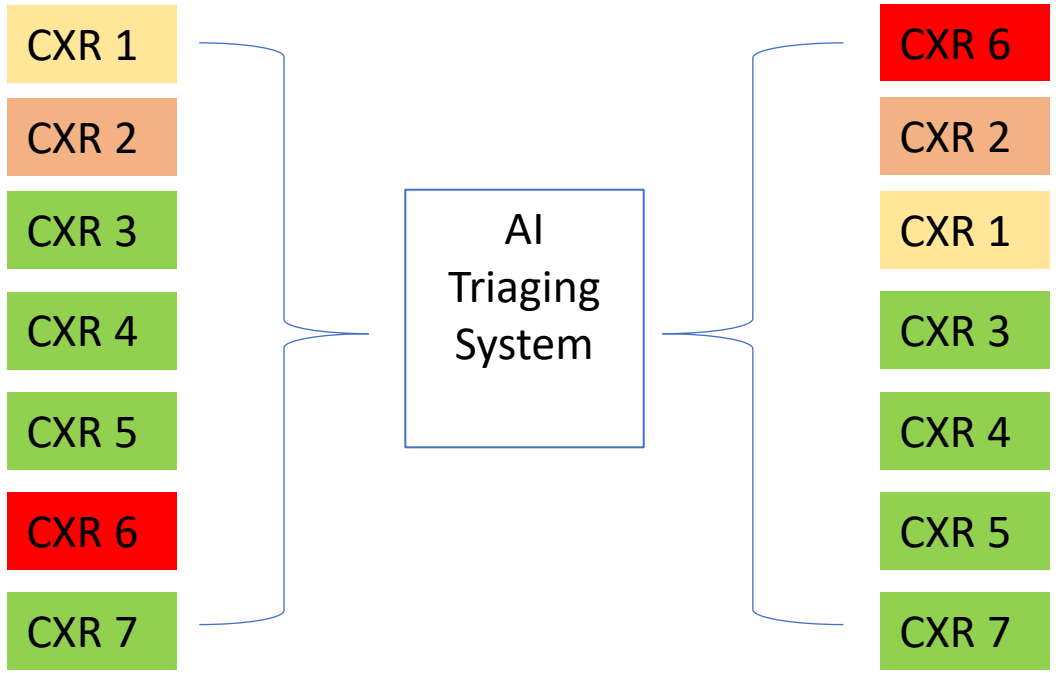
CREATING GROWTH, ENHANCING LIVES

Defining the components

Component	Description	Our project
Task or use case	The deployed AI model's task as applied to the intended clinical use case	Multilabel CXR classification to aid triaging
AI model input	Data used by the AI model to generate output inferences	CXR images
AI model output	Inferences from the AI model after processing input	Presence/absence of each of 14 abnormalities/classes
Ground truth	The standard against which the AI model output is judged to be correct	Radiology reports of respective CXR
Feedback	The method of showing and comparing AI model performance against the ground truth to aid audit and monitoring	Control chart

Use Case: AI-Assisted Triaging of CXR Studies

Objective: Given surge in CXR volume, prioritise CXR studies with significant abnormal findings for early management



Low Priority

Normal

Medium Priority

Atelectasis
Enlarged Cardiomediastinum
Pleural Other
Cardiomegaly
Support Devices

High Priority

Pleural Effusion
Lung Lesion
Lung Opacity
Oedema

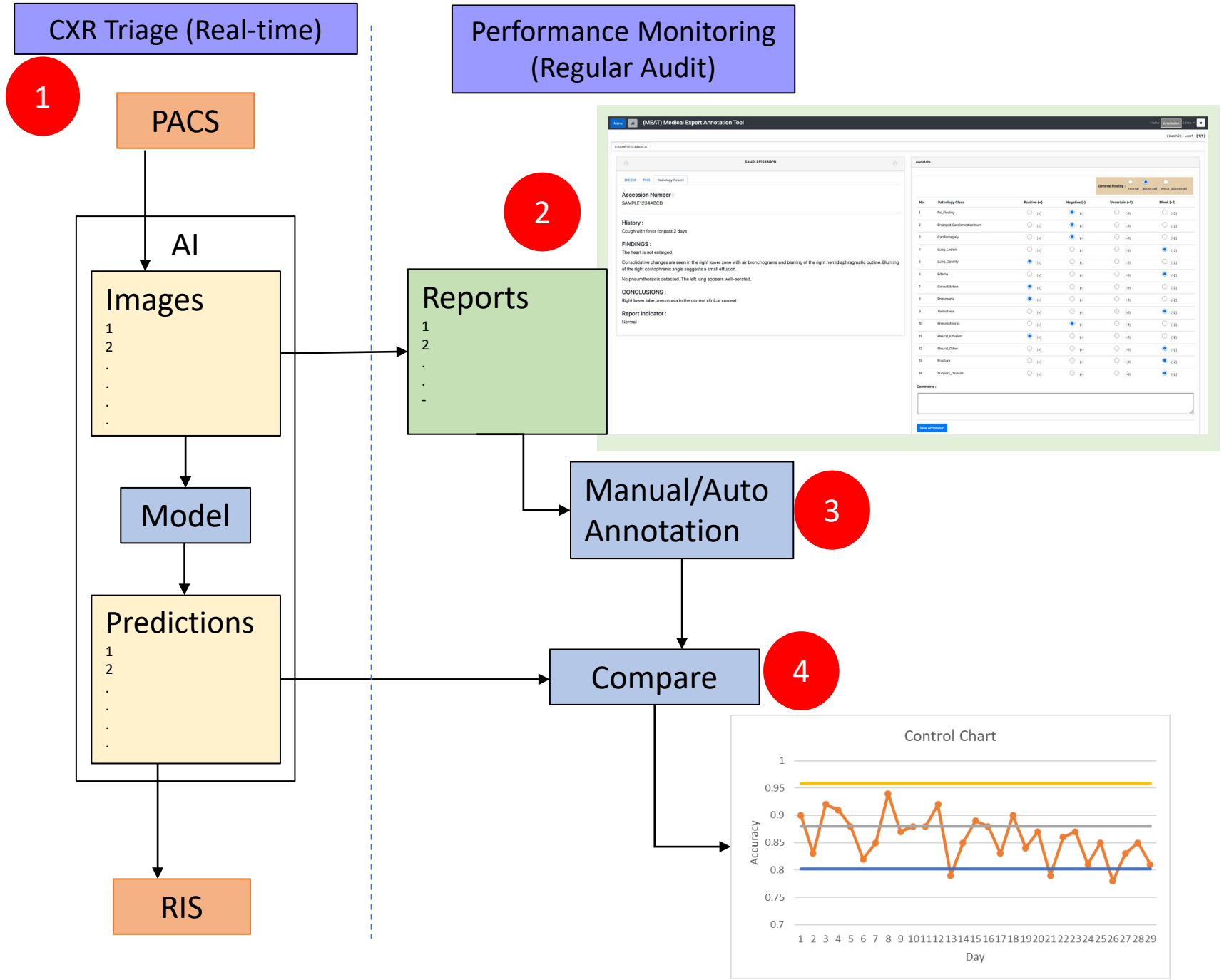
Highest Priority

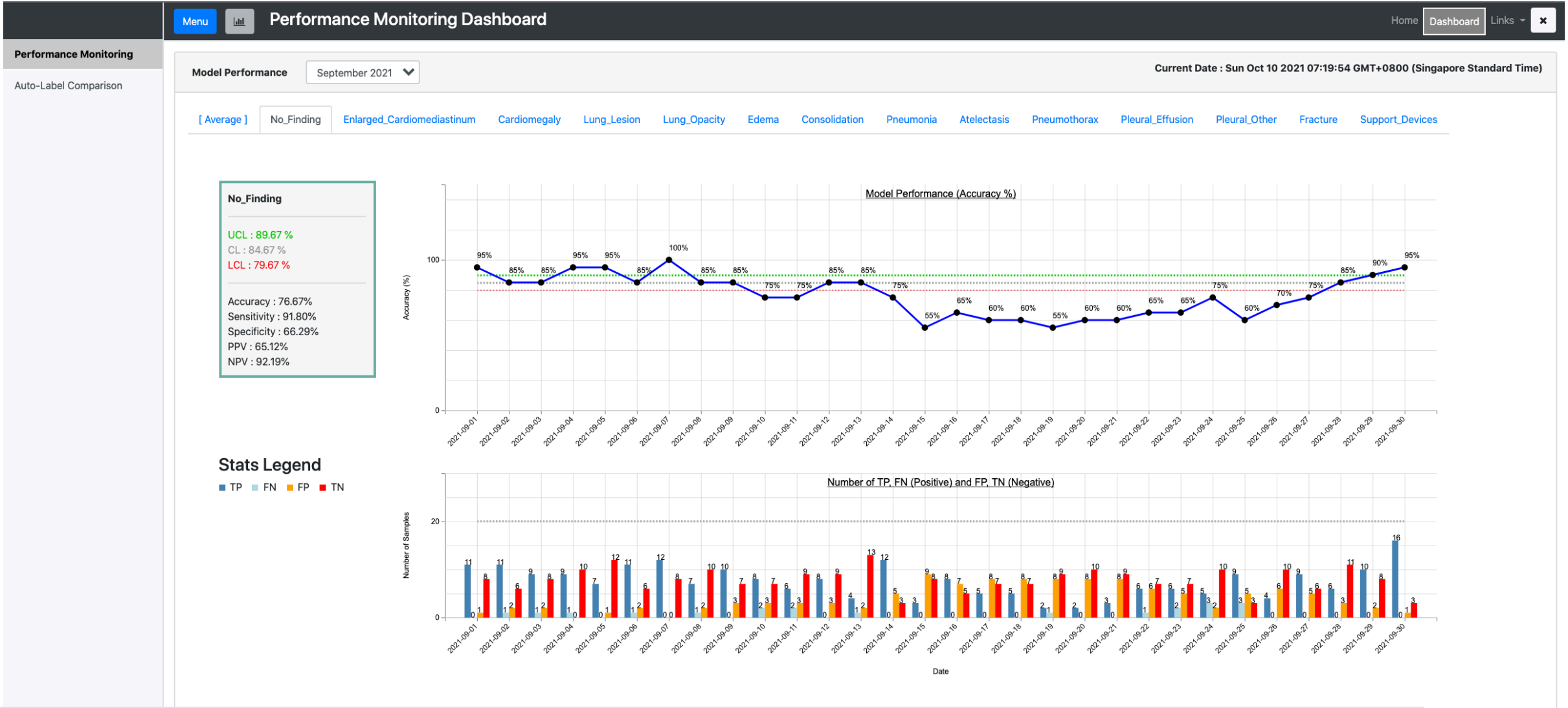
Pneumonia
Pneumothorax
Consolidation
Fracture

*Labels as used in "CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison" by Irvin & Rajpurkar et al. (arXiv:1901.07031v1)

System overview

1. The end-to-end CXR Triage system runs in real-time.
2. After the studies are reported by radiologists, the Performance Monitoring framework generates ground truth labels from the reports.
3. This is done either manually with the aid of an annotation tool or automatically using a natural-language processing (NLP) labeller.
4. By comparing the ground truths with the AI output, we can generate a control chart showing daily change in metrics such as accuracy, sensitivity, false positive rate, etc. for easy visual feedback on AI model performance across time.





Illustrated example from trial deployment in a test environment: Observation of performance degradation midway through the month prompts review and adjustment of model thresholds, leading to restoration of model performance at the end of the month. This increases confidence in safe and accurate AI model performance. In real-world settings, investigation of root cause(s) such as data drift (e.g., changes in disease prevalence, equipment upgrade) will be initiated to assess the need for corrective action including AI model update.

Annotation burden

- Manually annotating/labelling each CXR still requires manpower which we mitigate by:
  - Annotating radiology reports rather than the images (increase ease, speed, and consistency)
  - Using an integrated annotation tool (increase ease and speed)
  - Annotating a daily sample instead of all reports (reduce volume)
- Concurrent annotation using NLP labeller:
  - If performance of NLP labeller is comparable to manual labelling, we can alleviate annotation burden without sacrificing performance.
  - Auto-labelling also mitigates against inter-rater variation from manual labelling.
  - Manual effort can be reduced to quarterly or semi-annual quality checks to maintain accuracy of NLP labeller.

- Example of comparison between labels generated by the NLP labeller and manual annotation using radiology reports

Auto-Label Comparison

image_id	No_Finding [F=0.0%]	Enlarged_Cardiomediastinum [F=35.0%]	Cardiomegaly [F=5.0%]	Lung_Lesion [F=0.0%]	Lung_Opacity [F=25.0%]	Edema [F=5.0%]	Consolidation [F=25.0%]	Pneumonia [F=10.0%]	Atelectasis [F=10.0%]	Pneumothorax [F=0.0%]	Pleural_Effusion [F=0.0%]	Pleural_Other [F=0.0%]	Fracture [F=0.0%]	Support_Devices [F=10.0%]
317					M=[1], A=[-2]									M=[1], A=[-2]
380		M=[1], A=[-2]					M=[1], A=[-2]							
481		M=[1], A=[-2]												
947														
249			M=[1], A=[-1]											
581		M=[1], A=[-2]												
826		M=[1], A=[-2]			M=[1], A=[-2]	M=[1], A=[-1]								
966		M=[1], A=[-2]			M=[1], A=[-2]				M=[1], A=[-1]					M=[1], A=[-2]
484														
514														
577														
583					M=[1], A=[-2]									
680														
717														
018							M=[1], A=[-2]	M=[1], A=[-2]						
510		M=[1], A=[-2]												
006					M=[1], A=[-2]		M=[1], A=[-2]	M=[1], A=[-1]	M=[-2], A=[1]					
174														
390		M=[1], A=[-2]					M=[1], A=[-2]							
510							M=[1], A=[-2]							

Refresh Data

- Green boxes indicate concordance between manual and auto labels
- The contents within the white boxes provide details on the discrepancies (e.g., in the sole discrepancy in the cardiomegaly column shown here, the human annotator interpreted the report as mentioning that cardiomegaly is “present” whereas the NLP labeller interprets this as “uncertain”.)
- (Legend: F = % discordance; M = Manual; A = Auto; 1 = present; 0 = absent; -1 = uncertain; -2 = not mentioned)

CREATING GROWTH, ENHANCING LIVES

Summary and future work

- Performance monitoring is an important yet often overlooked aspect of clinical AI deployment
- Effectively doing so increases confidence in AI systems by letting stakeholders know when we can or cannot rely on AI outputs in a timely manner
- Our framework is applicable across a variety of clinical AI use cases
- We are in the process of migrating deployment from the test environment to the production environment
- We plan to scale and deploy alongside other AI systems in medical imaging while making improvements to system efficiency and robustness, including options to refresh the AI models as and when needed

Acknowledgements

SKH

- Victor Cheong
- Michael Ng
- Uppaluri Srinivas Anandswaroop
- Shawn Kok
- Syed Aftab
- Senior management and clinical champions

IHiS

- Chai Chun Wei
- Richard Sunga
- Frederick Tea

Philips/Carestream

- Rajesh Sajda

Contact Us: Ivan Ho (Ivan_Ho@i2r.a-star.edu.sg)