Open-Source Foundation Models on Plain Musculoskeletal and Chest Radiographs: A Multi-Model Diagnostic Accuracy Study

Katrina Nash[1,2], Anissa Alloula[3], Joseph Salf[4], Katharine Wykes[1,5], James Vaz[1,5], Alex[1,5] Novak, Bartlomiej Papiez[3], Sarim Ather[1,5]

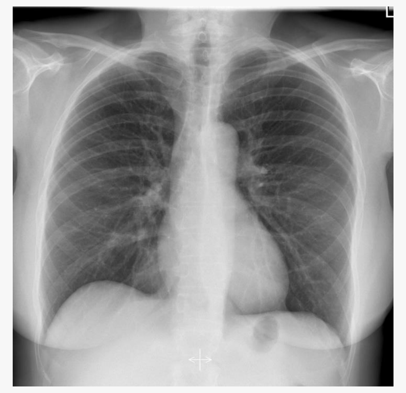
1-John Radcliffe Hospital, Oxford University Hospitals NHS Foundation Trust, 2-Oxford University Clinical Academic Graduate School, University of Oxford, 3-Oxford Big Data Institute, University of Oxford, 4 – NHS Tayside, Dundee, 5-Oxford Clinical Artificial Intelligence Research, Oxford University Hospitals

Oxford University Hospitals NHS Trust | BIG DATA INSTITUTE | UNIVERSITY OF OXFORD | OxCAIR OXFORD CLINICAL ARTIFICIAL INTELLIGENCE RESEARCH

Background

- Studies have demonstrated excellent performance of foundation models in radiology
- Recent claims have suggested the possibility of cross modality diagnostic capabilities
- Plain radiography is the most commonly performed imaging modality
- Detection of pneumothorax and fractures represents clinically significant, time critical tasks, where diagnostic error carries measurable patient risk
- There are limited independent, comparative evaluation of open-source foundation models on real world data

Case courtesy of Jeremy Jones, Radiopaedia.org, rID: 6428



Case courtesy of Usman Bashir, Radiopaedia.org, rID: 18394

Aim: To evaluate the diagnostic performance of multiple open-source foundation models for detection of pneumothorax on chest radiographs (CXR) and fractures on musculoskeletal (MSK) radiographs using an NHS dataset.

Methodology

Retrospective review of 394 CXRs and 500 MSK radiographs

↓

Enriched dataset ensured a broad range of pathology subtypes and robust evaluation of rare pathologies

↓

Eight open-source foundation models were evaluated in a zero-shot setting (no fine tuning)

↓

Reference standard was defined by two consultant radiologists, with a third arbitrator

↓

Binary outputs were compared to evaluate sensitivity and specificity of each model, ROC curves where possible

↓

Subgroup analysis of demographic and imaging characteristic

Results

CXR Dataset

- Mean age: 56 years old (SD 21)
- 156 females (39.6%); 238 males (60.4%)
- 186 positive(47.2%) and 208 negative for pneumothorax (52.8%)
- 135 (35%) AP and 256 (65%) PA radiographs

MSK Dataset

- 219 positive (43.8%) and 281 (56.2%) negative for a fracture
- Mean age: 55 years old (SD 24)
- 256 females (53%); 219 males (43.8%)
- Fewer shoulder X-rays (4%) and slightly more pelvis/hip X-rays (26%)
- Otherwise approximately equal proportions of X-rays by anatomical region

Model Key

- M01 = MAIRA grounded[1]
- M02 = MAIRA ungrounded[1]
- M03 = CXR Foundation Binary[2]
- M04 = Biomedclip[3]
- M05 = ChexAgent Binary Score[4]
- M06 = BiomedGPT[5]
- M07 = Radfm[6]
- M08 = Medgemma[7]

	CXR	MSK
M01	62	n/a
M02	0	n/a
M03	0	0
M04	0	0
M05	0	0
M06	249	180
M07	0	0
M08	193	0

Table 1: Number of scans not processed by each model

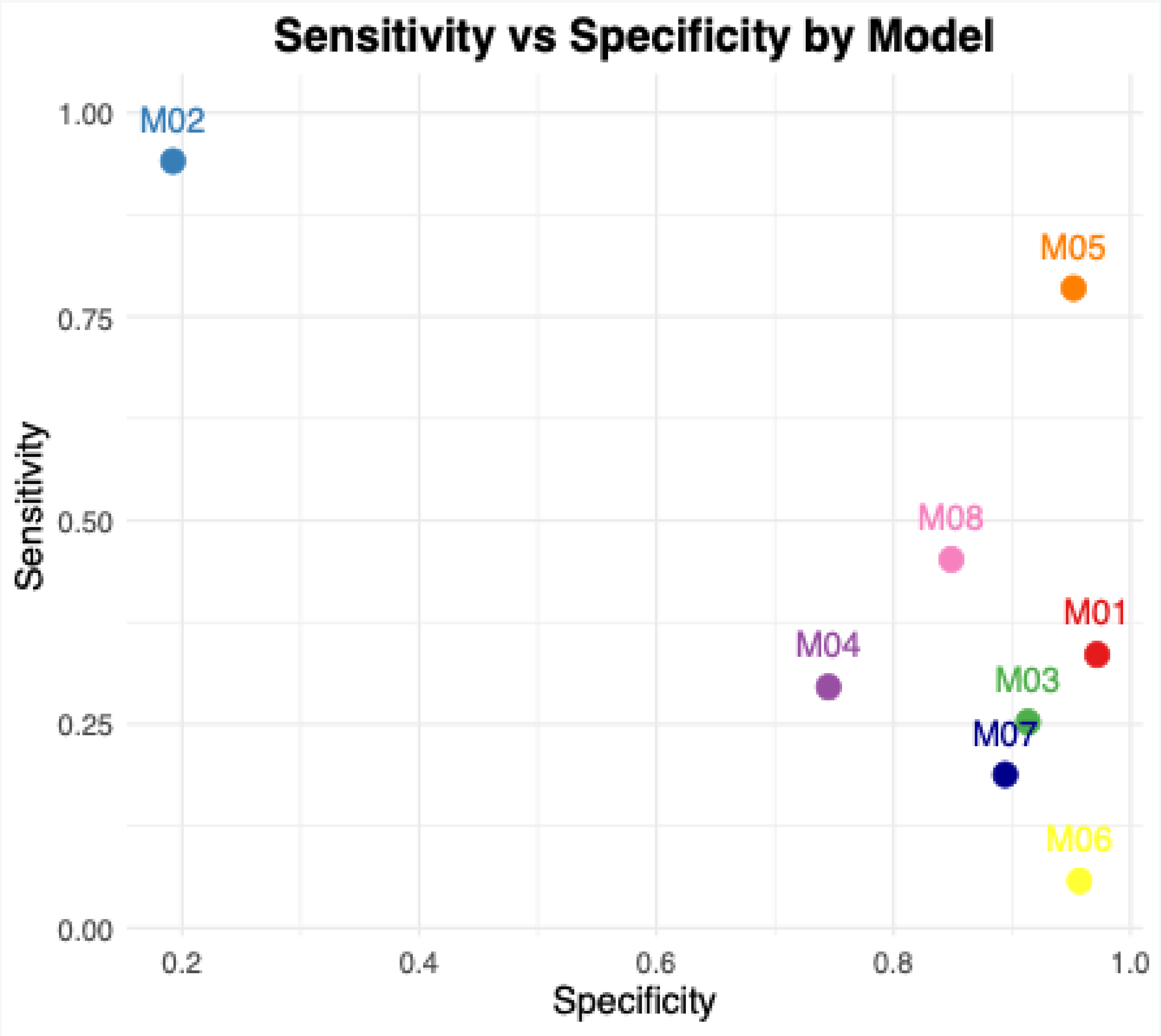


Figure 1: Scatter graph showing sensitivity vs. specificity by model for pneumothorax on CXR

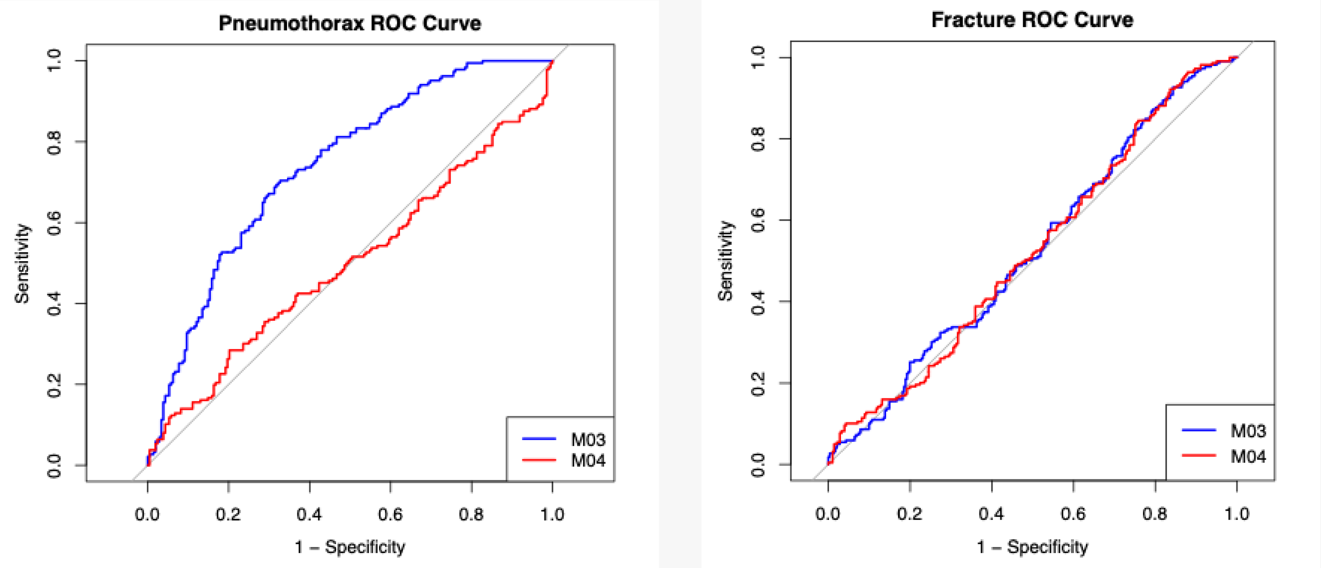


Figure 3: Receiver Operating Characteristic (ROC) Curve for M03 and M04 for pneumothorax on CXR and fracture on MSK radiograph

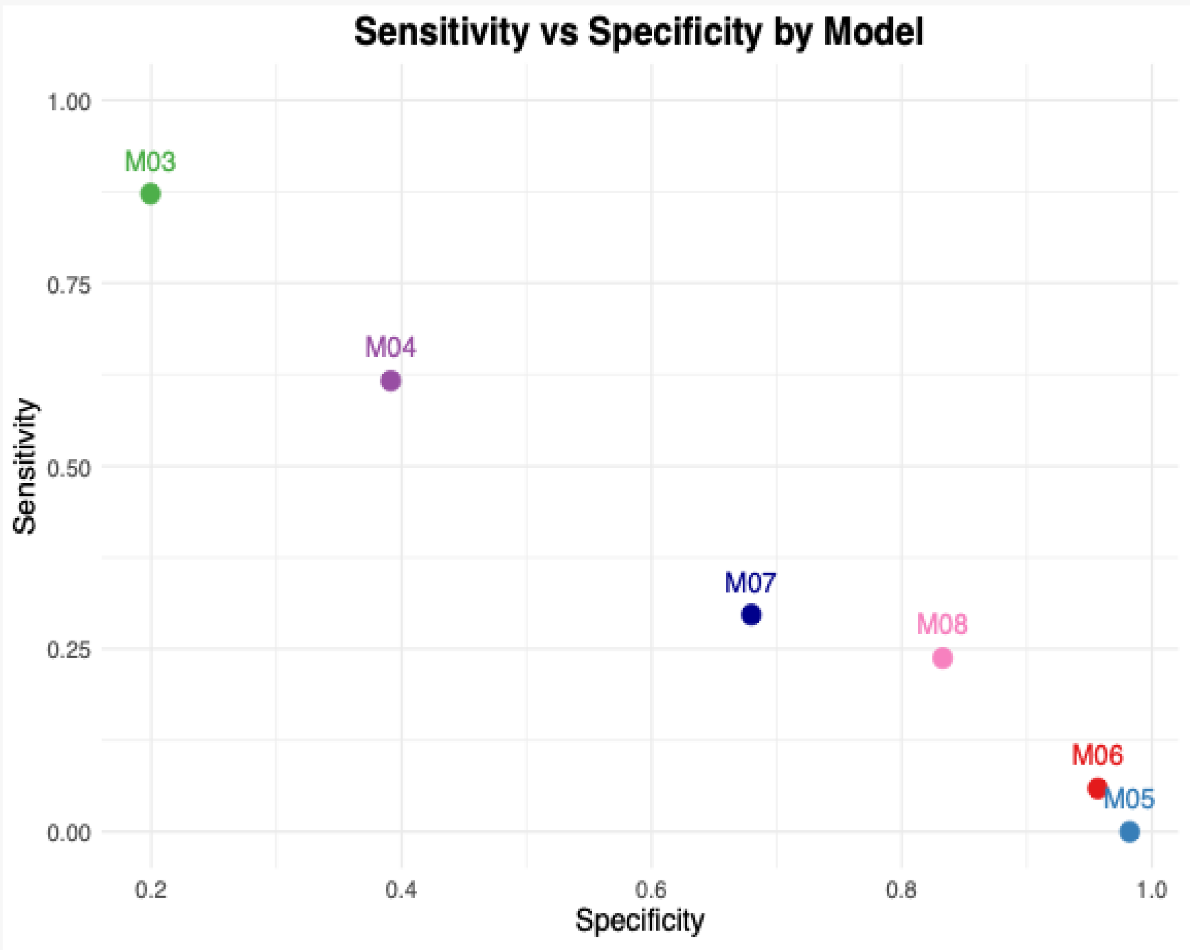


Figure 2: Scatter graph showing sensitivity vs. specificity by model for fracture on MSK radiograph

Key Results:

❖ **Chest Radiographs:** Only one model demonstrated both a high sensitivity and specificity, the remaining models showed accuracies close to chance

❖ **Musculoskeletal Radiographs:** No model achieved both a sensitivity and specificity above 50%

Strengths and Limitations

Strengths

- ✓ First systematic comparative evaluation of foundation models
- ✓ Rigorous reference standard
- ✓ Zero shot deployment → NHS trusts have limited resources for fine tuning

Limitations

- Small datasets → further larger studies should be studied
- Single site → may limit generalisability to other locations
- Limited set of pathologies → performance may differ in other pathologies

Conclusions

Foundation models demonstrated poor performance on an NHS dataset

- The majority of models had a tendency toward predominantly positive or negative classifications, resulting in a high sensitivity with low specificity, or vice versa
- M05 was the only model to demonstrate a high sensitivity and specificity for pneumothorax on CXR

Training datasets are likely vital in underpinning model performance

- M01 and M02 are architects of MAIRA-2 model, but trained using different datasets
- M05

Foundation models are not yet ready for live deployment in zero shot setting

- Future studies should evaluate the performance improvement with fine-tuning

References

1 - Microsoft. Hugging Face [Internet]. 2024 [cited 2026 Feb 24]. MAIRA-2. Available from: https://huggingface.co/microsoft/maira-2
2- Google. Hugging Face [Internet]. 2003 [cited 2026 Feb 24]. CXR Foundation Model. Available from: https://huggingface.co/google/cxr-foundation
3- Microsoft. Hugging Face [Internet]. 2023 [cited 2026 Feb 24]. BiomedCLIP-PubMedBERT_256-vit_base_patch16_224. Available from: https://huggingface.co/microsoft/BiomedCLIP-PubMedBERT_256-vit_base_patch16_224
4- Stanford AIMI. GitHub [Internet]. 2024 [cited 2026 Feb 24]. CheXagent. Available from: https://github.com/Stanford-AIMI/CheXagent
5 - Taokun Zhang. GitHub [Internet]. 2023 [cited 2026 Feb 24]. BiomedGPT. Available from: https://github.com/taokz/BiomedGPT
6 - Wu C. GitHub [Internet]. 2025 [cited 2026 Feb 24]. RadFM. Available from: https://github.com/chaoyi-wu/RadFM
7 - Google. MedGemma [Internet]. 2025 [cited 2026 Feb 24]. Available from: https://huggingface.co/google/medgemma-4b-it