

Foundation Models for Chest Radiography:

*Knowledge Distillation and Impacts on
Performance and Bias Propagation*

Motivation

Fadi Zahar

26/06/2024

Plan

Motivation

1. Deep Learning in Medical Imaging
2. Foundation Models
3. Associated Challenges
4. Need for Knowledge Distillation

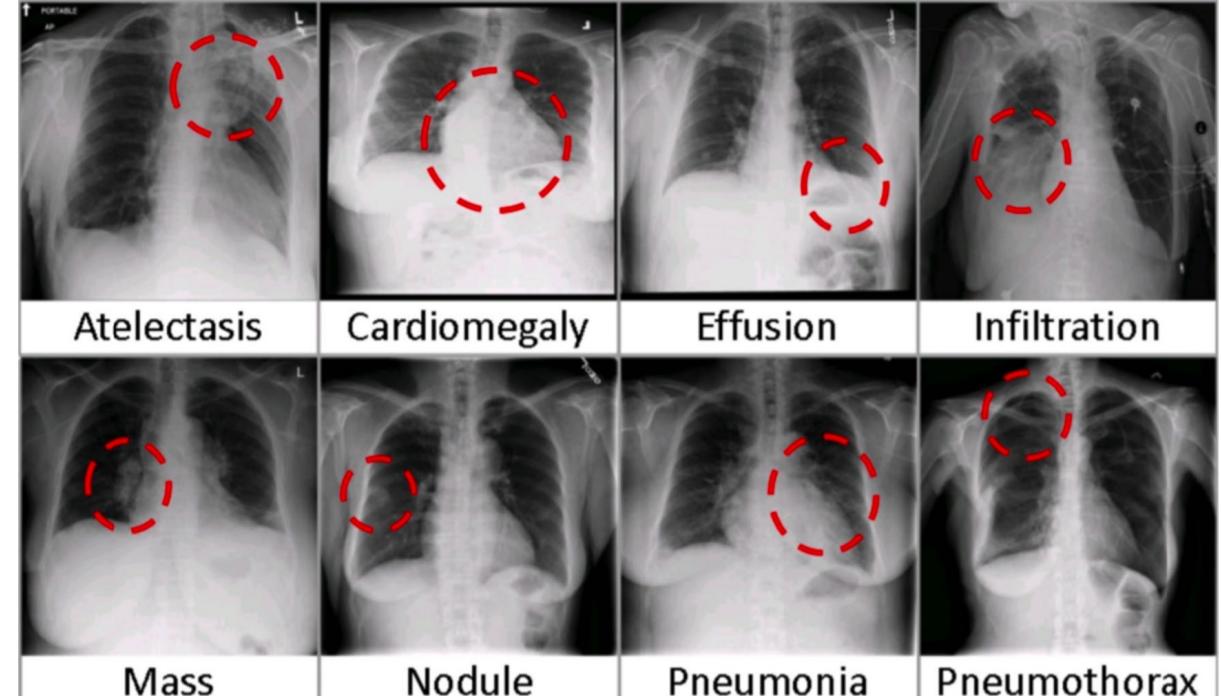
Motivation

1. Deep Learning in Medical Imaging

Deep Learning(DL)-based models have shown significant success in medical imaging applications:

- **Accelerate** medical image analysis (automation)
- Match or **exceed human expert** performance

e.g., thoracic disease detection using chest X-rays (CXRs)



Common thoracic diseases observed in Chest X-rays [1]

1. Deep Learning in Medical Imaging

Deep Learning(DL)-based models have shown significant success in medical imaging applications:

- **Accelerate** medical image analysis (automation)
- Match or **exceed human expert** performance

e.g., thoracic disease detection using chest X-rays (CXRs)

However, they need **very large amounts of representative data** for good generalisation



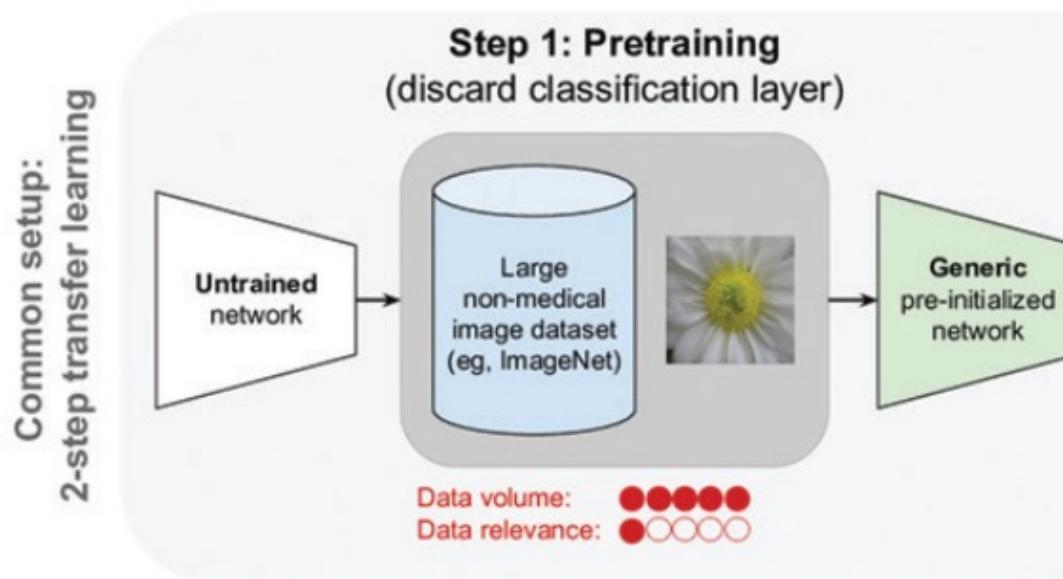
Medical images: Expensive labelling & data privacy concerns

Otherwise, they risk to **fail** due to **data distribution shifts** in different clinical settings:

- Population shifts (*demographics*)
- Acquisition shifts (*imaging modalities*)

2. Foundation Models

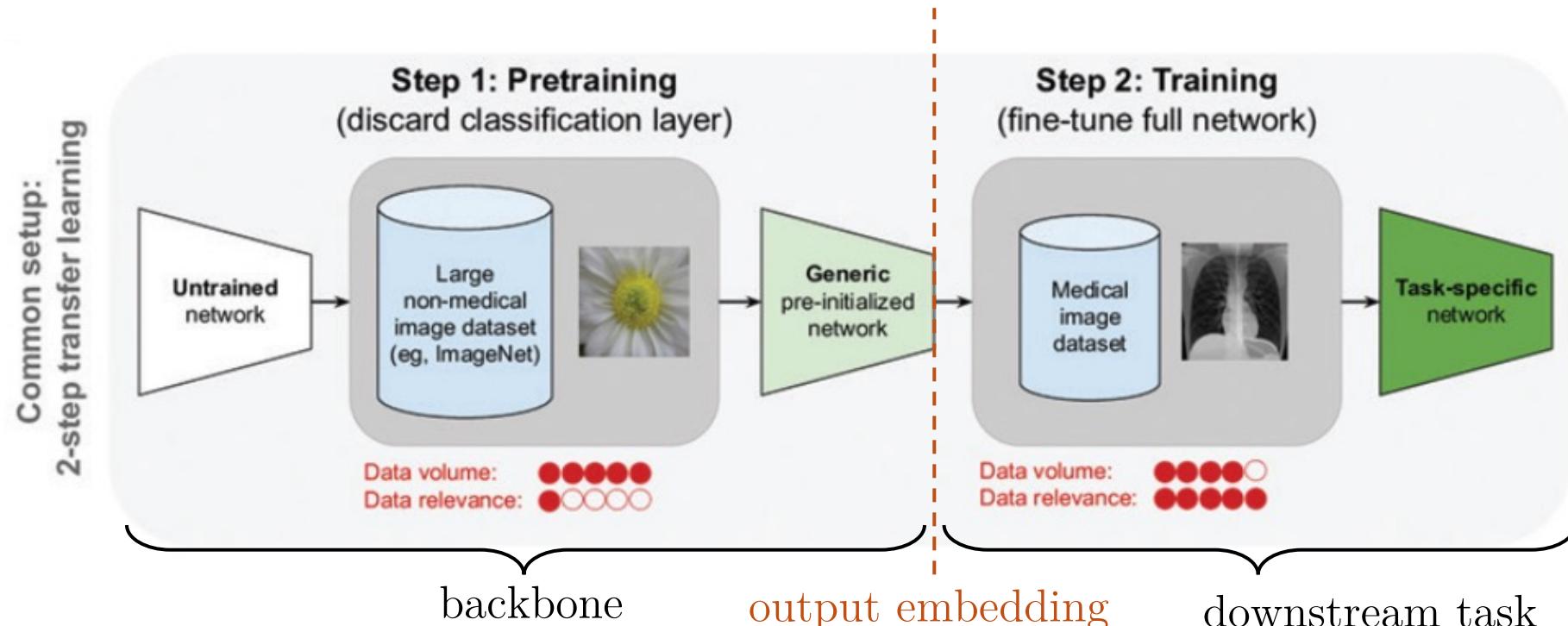
Solution: ***Foundation Models*** → pretrained on heterogeneous, diverse, large-scale datasets;
often via self-supervised or semi-supervised learning



“surrogate” task

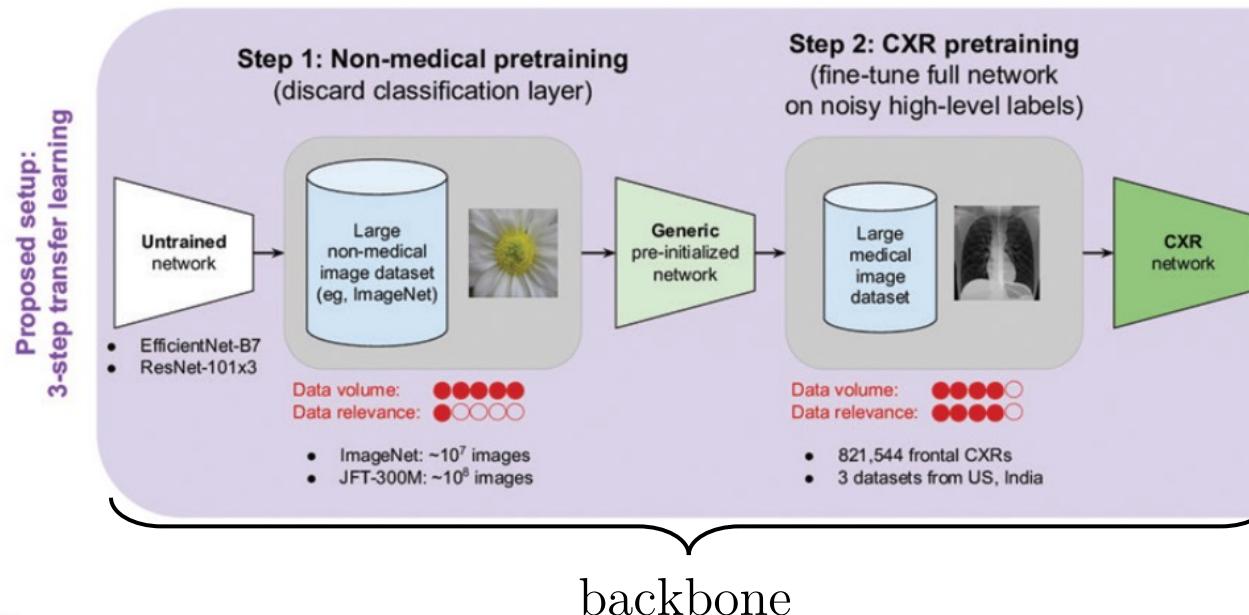
2. Foundation Models

Solution: ***Foundation Models*** → pretrained on heterogeneous, diverse, large-scale datasets;
often via self-supervised or semi-supervised learning
→ generated feature representations (embeddings) then used in downstream task adaptation with much less training data



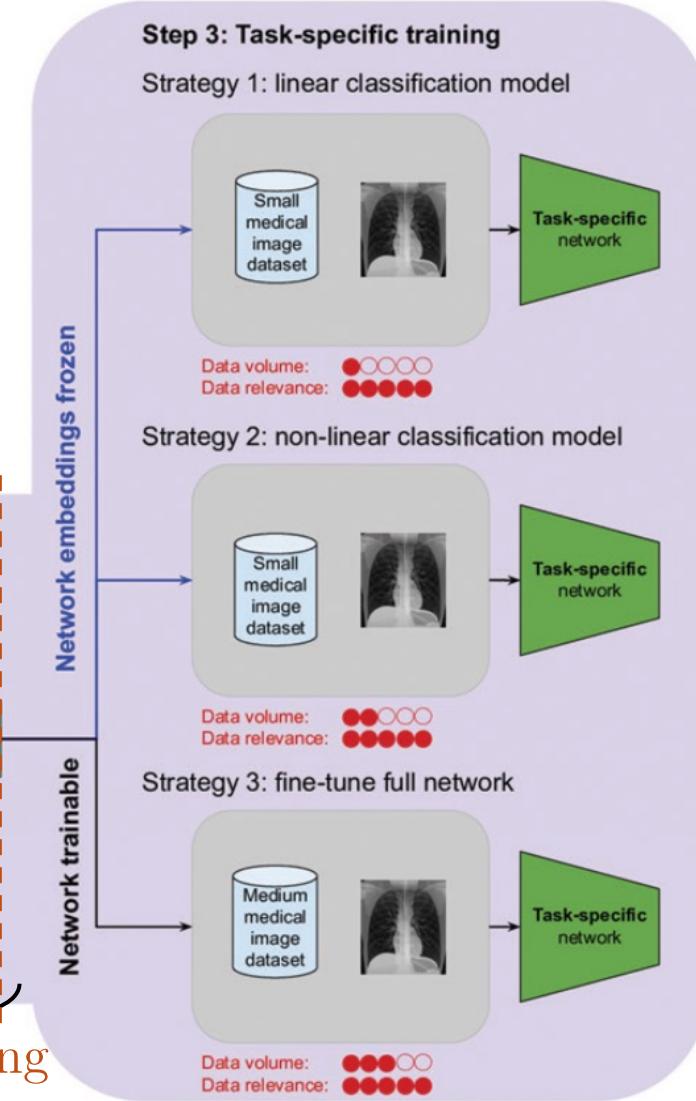
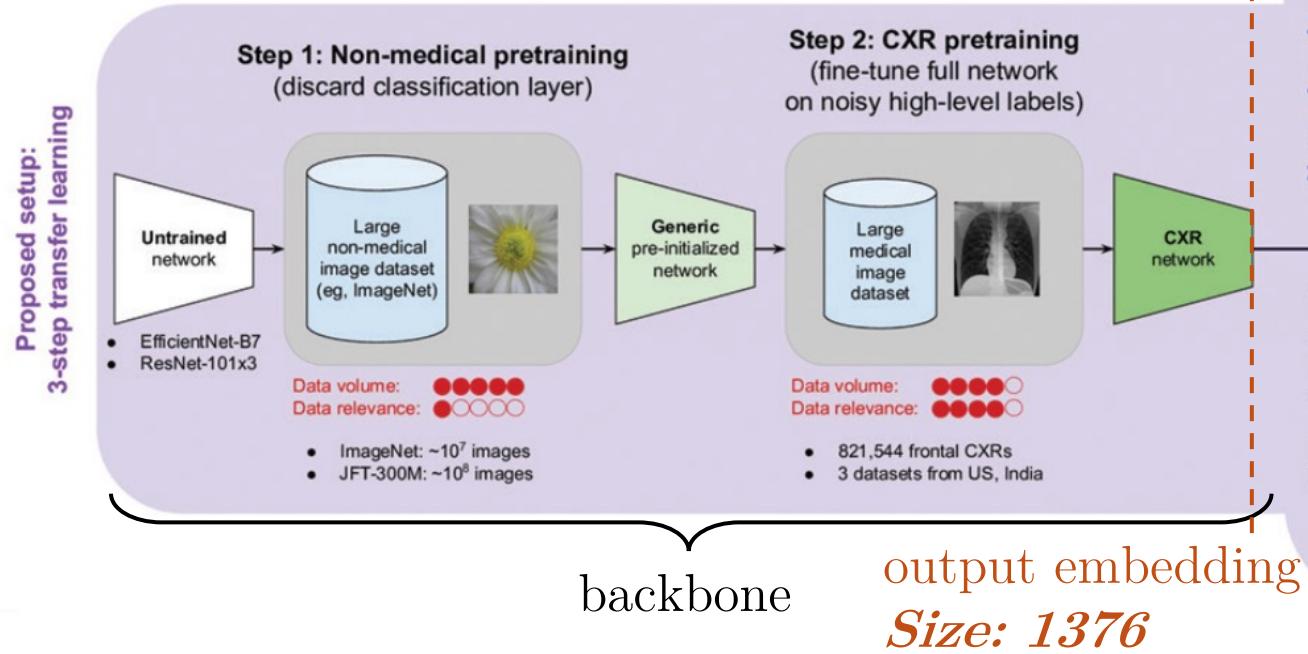
2. Foundation Models

Google's CXR-FM: Adds a CXR pretraining to create a ‘better’ backbone for downstream CXRs analysis



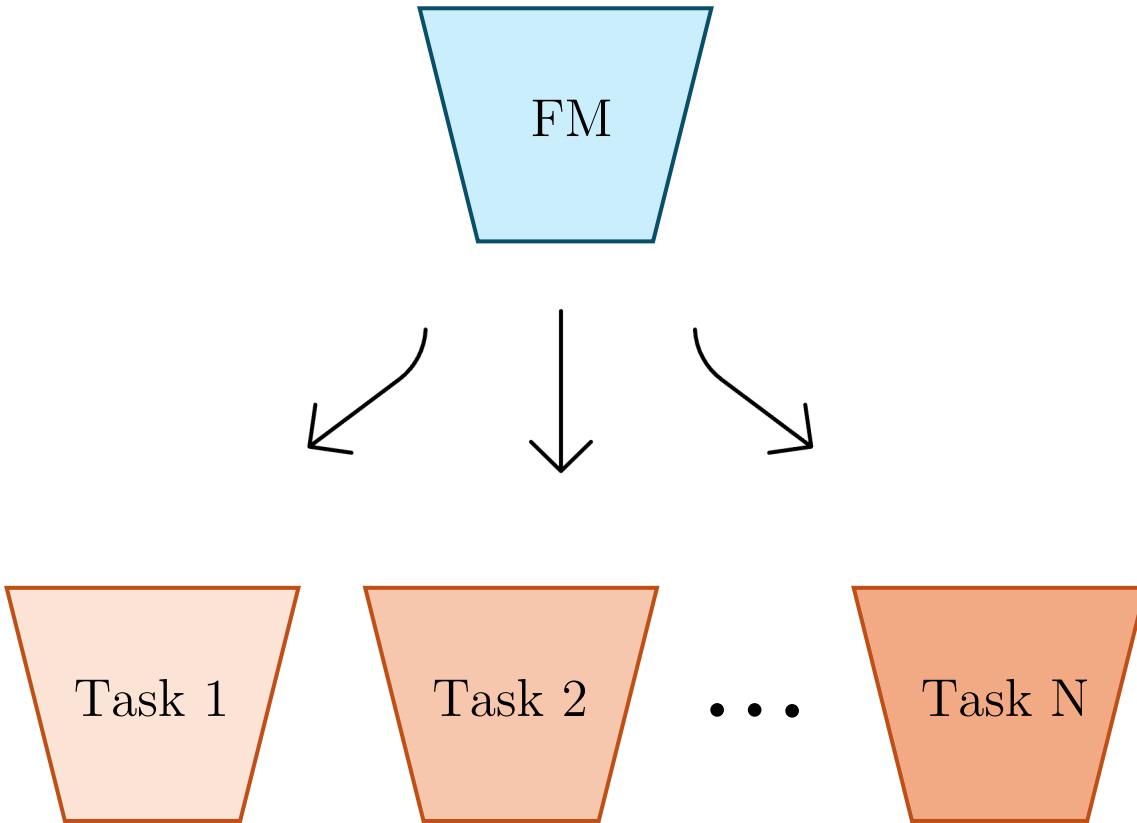
2. Foundation Models

Google's CXR-FM: Adds a CXR pretraining to create a ‘better’ backbone for downstream CXRs analysis



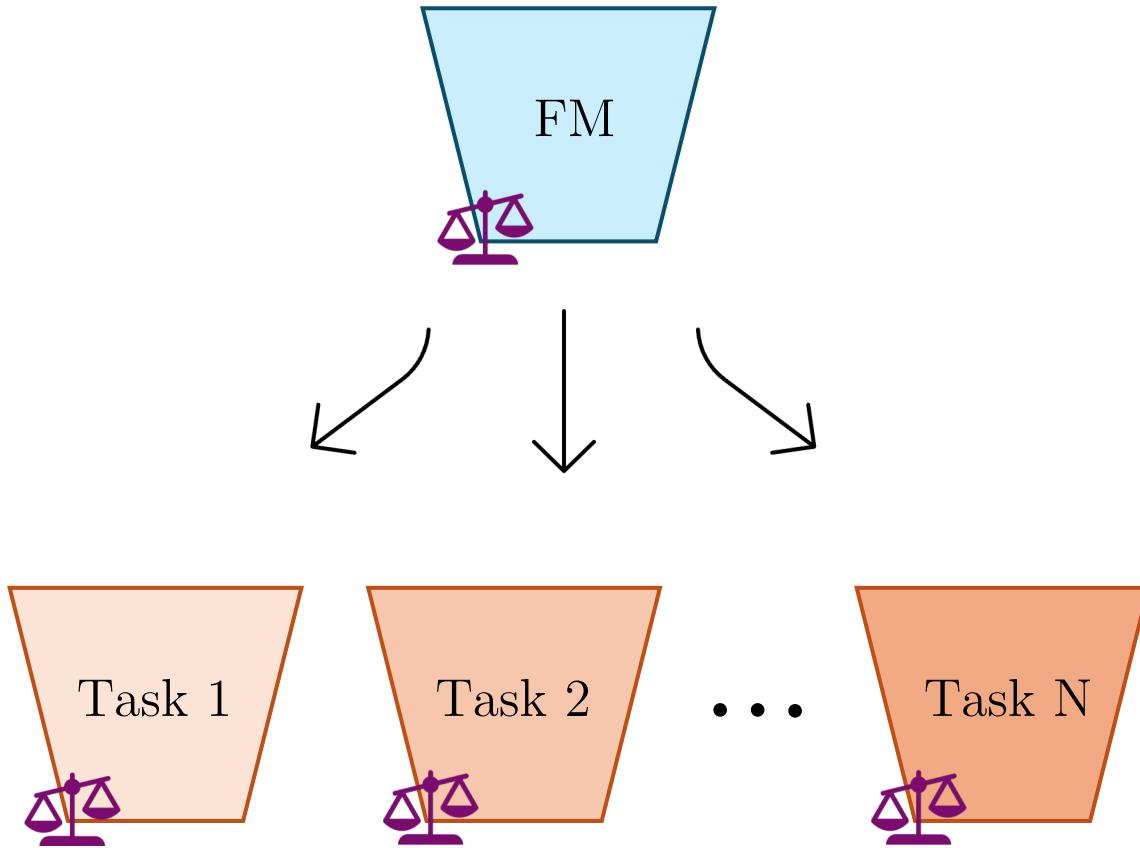
3. Associated Challenges

- General lack of profound understanding of Foundation Models (FMs)' mechanisms and impacts



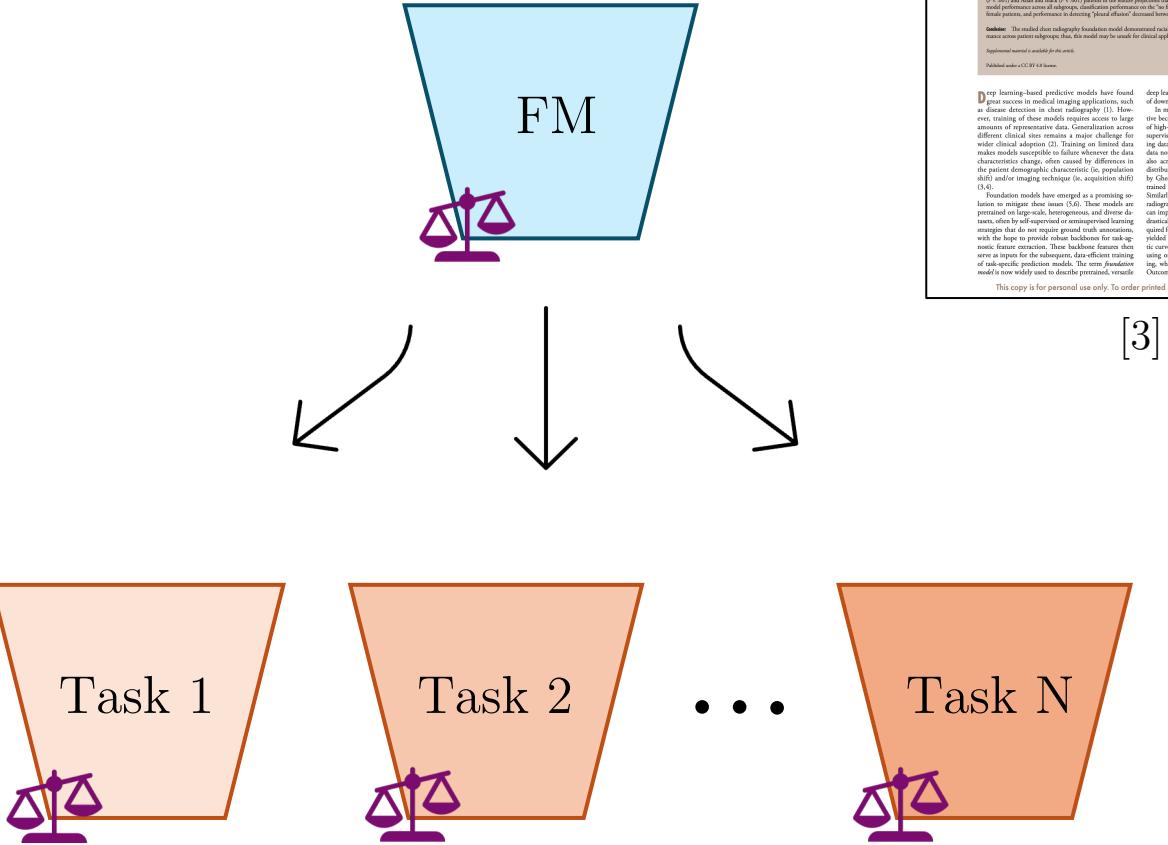
3. Associated Challenges

- General lack of profound understanding of Foundation Models (FMs)' mechanisms and impacts
- Can propagate encoded biases through downstream tasks: *any defect is likely ‘inherited’ by all adapted models.*



3. Associated Challenges

- General lack of profound understanding of Foundation Models (FMs)' mechanisms and impacts
- Can propagate encoded biases through downstream tasks: *any defect is likely ‘inherited’ by all adapted models*
- Indeed, [3] showed that the CXR-FM exhibited biases related to race and biological sex: *making it unsafe for clinical deployment*



[3]

Radiology: Artificial Intelligence

ORIGINAL RESEARCH

Risk of Bias in Chest Radiography Deep Learning Foundation Models

Ben Glocker, PhD • Charles Jones, MEng • Milosav Bošković, MSc • Stefan Winkler, PhD

From the Department of Computing, Imperial College London, South Kensington Campus, London SW7 2AZ, United Kingdom. Received February 26, 2023; revised November 1, 2023; accepted December 1, 2023. Published online January 26, 2024. © 2024 by the Radiological Society of North America, Inc.

Received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement no. 852000).

CJ is supported by Microsoft Research and Engineering and Physical Sciences Research Council. MJB is funded through Imperial College London Prostasis PhD Studentship.

Conflicts of interest are listed at the end of this article.

See also: <https://doi.org/10.1148/radiol.230909> • [CrossMark](https://doi.org/10.1148/radiol.230909) • [PDF](https://doi.org/10.1148/radiol.230909)

Paper To analyze a recently published deep learning foundation model for the presence of biases that could lead to subgroup discrimination in chest radiography.

Method We used the Foundation Model and Accessibility for Computer Tomography study (17118 chest radiographs from 6284 patients (mean age, 63 years ± 17 years); 2062 male, 1516 female) from the ChaLearn dataset as our reference. We compared the performance of this model with a baseline model (the same architecture) trained on a dataset of 100000 randomly selected chest radiographs from the ChaLearn dataset. We also compared the performance of this model with a baseline model trained on a dataset of 100000 randomly selected chest radiographs from the ChaLearn dataset and another deep learning model (demonstrated relevance method) without web-crawled Schlegens-Senior test images. Finally, we compared the performance of this model with a baseline model trained on a dataset of 100000 randomly selected chest radiographs from the ChaLearn dataset and another deep learning model trained on a dataset of 100000 randomly selected chest radiographs from the ChaLearn dataset that were preprocessed to remove any bias in the features to specific disparities in classification performance across patient subgroups.

Results Ten of 12 pairwise comparisons across biologic sex and race showed statistically significant differences in the model's performance. The Foundation Model performed significantly worse than the baseline model in discriminating between White and Black ($P < .001$) and Asian and Black ($P < .001$) patients in the former prediction that privately obscure disease. Compared with average model performance, the Foundation Model had a 10% decrease in performance for White patients and a 17% increase for Black patients, and performance in detecting “pleural effusion” decreased between 15.7% and 11.6% for Black patients.

Conclusion The resulting chest radiography foundation model demonstrated racial and sex-related bias, which led to disparate performance across patient subgroups. The model may be used for clinical applications.

Supplemental material is available for this article.

Published online in *Radiology* first look. DOI: 10.1148/radiol.230909

This copy is for personal use only. To order printed copies, contact reprints@rsna.org.

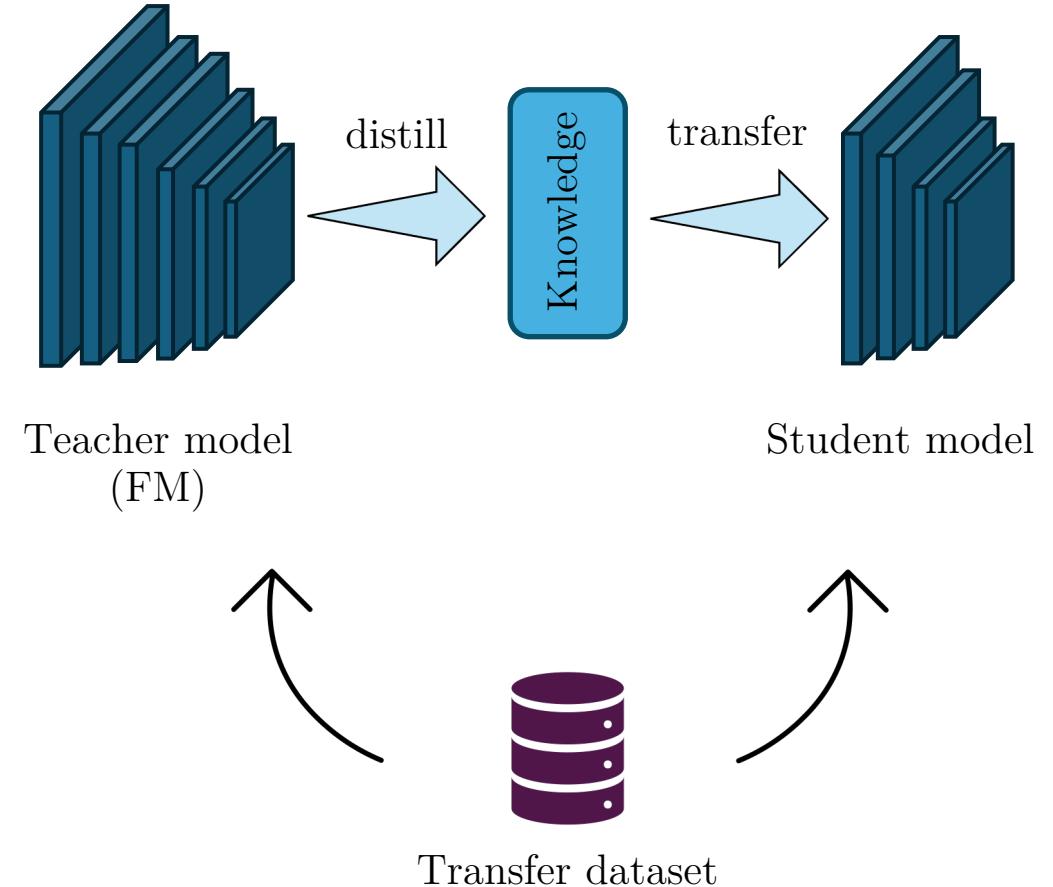
4. Need for Knowledge Distillation

- For effective bias mitigation → full access to the FM is needed [3]
- However, **CXR-FM** for example is only released as an ‘API call’: it takes input images and generate output features BUT the **parameters are not publicly available**

4. Need for Knowledge Distillation

- For effective bias mitigation → full access to the FM is needed [3]
- However, **CXR-FM** for example is only released as an ‘API call’: it takes input images and generate output features BUT the **parameters are not publicly available**

- Interest to explore **Knowledge Distillation** (KD) as a means to transfer the learned knowledge from the *teacher* (e.g., CXR-FM) to a usually smaller *student* model.



References

1. Louati, H., Louati, A., Bechikh, S. et al. Topology optimization search of deep convolution neural networks for CT and X-ray image classification. *BMC Med Imaging* 22, 120 (2022).
<https://doi.org/10.1186/s12880-022-00847-w>
2. Sellergren AB, Chen C, Nabulsi Z, et al (2022) Simplified Transfer Learning for Chest Radiography Models Using Less Data. *Radiology* 305:454–465
3. Glocker B, Jones C, Roschewitz M, Winzeck S (2023) Risk of Bias in Chest Radiography Deep Learning Foundation Models. *Radiol Artif Intell* 5:230060
4. Gou J, Yu B, Maybank SJ, Tao D (2020) Knowledge Distillation: A Survey. *Int J Comput Vis* 129:1789–1819