Letters to the Editor

Reconsidering Conclusions of Bias Assessment in Medical Imaging Foundation Models

From Akshay S. Chaudhari, PhD*†, Christian Bluethgen, MD, MSc‡, and David Ouyang, MD§

Department of Radiology, Stanford University, 1201 Welch Rd, Stanford, CA 94305*

Department of Biomedical Data Science, Stanford University, Stanford, Calif†

Institute for Diagnostic and Interventional Radiology, University Hospital Zurich, Zurich, Switzerland‡

Department of Medicine, Division of Cardiology, Cedars-Sinai Medical Center, Los Angeles, Calif§

email: *akshaysc@stanford.edu*

Editor:

In the November 2023 issue of *Radiology: Artificial Intelligence*, Dr Glocker and colleagues report that a deep learning model for chest radiograph interpretation depicted unequal performance across subgroups of age and sex (1). Such studies are vital to understand the benefits and blind sides of current models that may be used in research studies and prospective clinical applications. Despite the important findings from this study, broader conclusions about foundation models and generalizability should be cautiously considered given the diversity and variability of foundation models. While the title describes the risk of foundation models, only a single, relatively small model was evaluated, and it is challenging to generalize such biases across a broader family of models.

Despite the foundation model terminology being relatively new, there is general consensus that foundation models are large-scale, self-supervised, and ostensibly multimodal (2). A closer evaluation of the examined model suggests three reasons that the underlying model might not be considered a foundation model. First, the model was only trained with fewer than 1 million domain-specific chest radiographs, which is in contrast to current natural-domain foundation models trained with billions of images or data elements (3). Out-of-distribution pretraining has also not shown benefits for training improved medical imaging models (4). Second, the domain-specific training for the model was performed in a supervised nature. Foundation models are trained in a self-supervised manner on all domain-specific data to generate robust and data-efficient feature representations to lower the likelihood of spurious correlations. Third, the model was trained only on images, while most foundation models are multimodal in nature, which is especially powerful given the multimodal nature of radiology images and reports.

Thus, while the conclusions of the study are an important call to action to develop more robust models and benchmarks for evaluating subgroup performance, generalization of conclusions to all foundation models is discordant with the results provided. All models may be susceptible to biases; however, this will require further investigation. In the meanwhile, we urge readers to contextualize the findings of the study as that *this supervised embedding model* has biases across subgroups, and that it is too early to assess whether all foundation models broadly have subgroup biases.

Disclosures of conflicts of interest: A.S.C. Contracts/grants to institution from GE HealthCare, Philips, and NIH; royalties/licenses from LVIS; consulting fees from Subtle Medical and Patient Square Capital; participation on a data safety monitoring board or advisory board for Brain Key and Chondrometrics; stock/stock options in Brain Key, Subtle Medical, and LVIS. **C.B.** No relevant relationships. **D.O.** Grants/contracts to institution from NIH, Alexion, and AHA; consulting fees from AZ, InVision, and Pfizer; payment or honoraria from Japanese Society of Echo and Korean Society of Echo.

References

1. Glocker B, Jones C, Roschewitz M, Winzeck S. Risk of Bias in Chest Radiography Deep Learning Foundation Models. Radiol Artif Intell 2023;5(6):e230060.
2. Bommasani R, Hudson DA, Adeli E, et al. On the Opportunities and Risks of Foundation Models. arXiv 2108.07258 [preprint] https://arxiv.org/abs/2108.07258. Posted August 16, 2021. Updated July 12, 2022. Accessed October 2023.
3. Sellergren AB, Chen C, Nabulsi Z, et al. Simplified Transfer Learning for Chest Radiography Models Using Less Data. Radiology 2022;305(2):454–465.
4. Raghu M, Zhang C, Kleinberg J, Bengio S. Transfusion: Understanding Transfer Learning for Medical Imaging. In: Advances in Neural Information Processing Systems Volume 32. Curran Associates, Inc; 2019.

Response

From

Ben Glocker, PhD, Charles Jones, MEng, Mélanie Roschewitz, MSc, and Stefan Winzeck, PhD

Department of Computing, Imperial College London, South Kensington Campus, London SW7 2AZ, United Kingdom

email: *b.glocker@imperial.ac.uk*

We appreciate the correspondence by Dr Chaudhari and colleagues regarding the question of what conclusions can and *should* be drawn from our study on the Risk of Bias in Chest Radiography Deep Learning Foundation Models. The fact that comprehensive bias and subgroup performance analyses are still largely missing in the literature on medical imaging foundation models is worrying and deserves particular attention. We are grateful for the opportunity to further highlight this issue.

The exact definition of the term *foundation model* seems rather philosophical, but not too important when assessing the potential risks (and opportunities) of the general trend toward using ever larger datasets for pretraining of (generic and domain-specific) deep learning models (1). Neither dataset size alone nor the specifics of the training strategy (be it unsupervised, self-supervised, or fully supervised) provide sufficient assurances for obtaining better or fairer models. In fact, in the absence of a supervised training signal (typically obtained from disease labels), a self-supervised approach may arguably be even at higher risk of learning biased representations due to the often strong yet implicit encoding of protected characteristics in imaging data such as chest radiographs (2). While one may hope

that using large-scale, diverse, and heterogeneous data would improve generalization and fairness, this remains to be seen but most certainly requires thorough validation.

The correspondence argues for a reconsideration of the conclusions drawn from our study, because we only investigated one specific model that was trained on "only" 800 000 images, with a supervised contrastive training objective. First, we would like to note that these limitations were already clearly discussed in the original article and contextualized in our conclusions. The reported performance disparities across biologic sex and race are indeed specific to the studied model and at no point did we suggest that similar disparities will necessarily exist in other foundation models. However, our study highlights these models are at *risk* of bias and crucially calls upon careful evaluation of biases for any model used for high stakes applications such as clinical diagnosis (3), irrespective of the type and amount of data and machine learning strategy used for training.

Disclosures of conflicts of interest: B.G. Grants from EU Commission (Grant Agreement No. 757173, Project MIRA, ERC-2017-STG), Innovate UK/UKRI (UKRI London Medical Imaging & Artificial Intelligence), and Royal Academy of Engineering (Centre for Value Based Healthcare), Research Chair in Safe Deployment of Medical Imaging AI; scientific advisor for Kheiron Medical Technologies (January 2018–September 2021); part-time employment at Kheiron Medical Technologies with stock options as part of the standard compensation package (since October 2021); part-time employment with HeartFlow with stock options as part of the standard compensation package (since September 2018). **C.J.** Microsoft Research/EPSRC, iCASE Microsoft PhD Scholarship. **M.R.** President's PhD Scholarship, Imperial College London. **S.W.** No relevant relationships.

References

1. Zhou Y, Chia MA, Wagner SK, et al; UK Biobank Eye & Vision Consortium. A foundation model for generalizable disease detection from retinal images. Nature 2023;622(7981):156–163.
2. Gichoya JW, Banerjee I, Bhimireddy AR, et al. AI recognition of patient race in medical imaging: a modelling study. Lancet Digit Health 2022;4(6):e406–e414.
3. Ahluwalia M, Abdalla M, Sanayei J, et al. The subgroup imperative: Chest radiograph classifier generalization gaps in patient, setting, and pathology subgroups. Radiol Artif Intell 2023;5(5):e220270.