Are Foundational Models Less Biased Than Specialized Models? An Ultrasound Study

Joris Fournel[1], Jakob Ambsdorf[2], Benjamin Jønch Jürgensen[1], Christopher Boland[1], Aya Elgebaly[1], Kamil Mikolaj[1], Paraskevas Pegios[1], Emilie Pi Fogtmann Sejer[3], Martin Tolsgaard[3], Anders Nymark Christensen[1], Mads Nielsen[2], and Aasa Feragen[1]

[1] Technical University of Denmark, Kongens Lyngby, Denmark
[2] University of Copenhagen, Copenhagen, Denmark
[3] Region Hovedstaden Hospital, Copenhagen, Denmark

Abstract. Foundation models (FMs) pre-trained via self-supervised learning (SSL) on large, diverse datasets are widely expected to offer superior generalization—and, by extension, reduced bias—compared to classically supervised (CS) models. Yet whether this holds in practice, and whether the choice of pretraining data matters, remains largely unresolved. We present a controlled comparison between SSL-pretrained FMs and a CS baseline fine-tuned on identical data across two prenatal ultrasound tasks: spontaneous preterm birth (sPTB) prediction and fetal scan weight estimation. Using visual radar-based bias profiles spanning a comprehensive set of clinical and acquisition subgroups, we show that FMs reproduce the bias profile of their CS counterpart—best- and worst-performing subgroups largely coincide, and mean performance gaps are comparable in magnitude. Pretraining data strategy does influence bias: oversampling cervical images during pretraining outperformed scaling to 13 million images for sPTB in both global performance and subgroup fairness, while naïve oversampling of rare devices degraded overall performance. These results challenge the assumption that large-scale diverse pretraining uniformly reduces bias, and suggest that data alignment with the downstream task matters more than scale.

Keywords: Foundational Models · Generalization · AI Fairness

1 Introduction

Foundation models—pre-trained via self-supervised learning (SSL) on large, diverse, unlabeled datasets—offer a compelling response to the chronic shortage of annotated data in medical imaging [3,16,23]. By transferring broad visual representations through lightweight fine-tuning [1, 2, 4], they are particularly attractive in healthcare, where annotation costs, disease rarity, and privacy constraints severely limit labeled data availability.

Yet the clinical deployment of medical AI remains fundamentally constrained by bias. Models trained on unrepresentative data produce systematic perfor-

mance disparities across patient subgroups [6, 12, 13, 19, 20]. On this topic, regulatory pressure is mounting [5], indicating that bias is not a peripheral concern, but a barrier to deployment.

A natural hypothesis is that SSL pre-training on large, diverse data should reduce susceptibility to the spurious correlations that drive these disparities in classically supervised (CS) models [8, 22]. However, foundation models (FMs) appear to reproduce many of the same biases as their purely-supervised counterparts [10, 11, 21], and Glocker *et al.* [7] found FMs to be *more* biased—though their model was pre-trained in a supervised manner, leaving the SSL case unresolved. Whether SSL pre-training at scale can mitigate fine-tuning biases, and whether pretraining strategy matters, remain open questions of direct practical consequence given the enormous cost of building foundation models.

This study provides a controlled, direct comparison between SSL-pretrained foundation models and a "traditional", purely-supervised baseline fine-tuned on identical data across two prenatal ultrasound tasks.

Our **research questions** are as follows:

1. Are foundation models less biased than classical, purely-supervised models?
2. What is the effect of the pretraining data regime on the foundation model's bias?

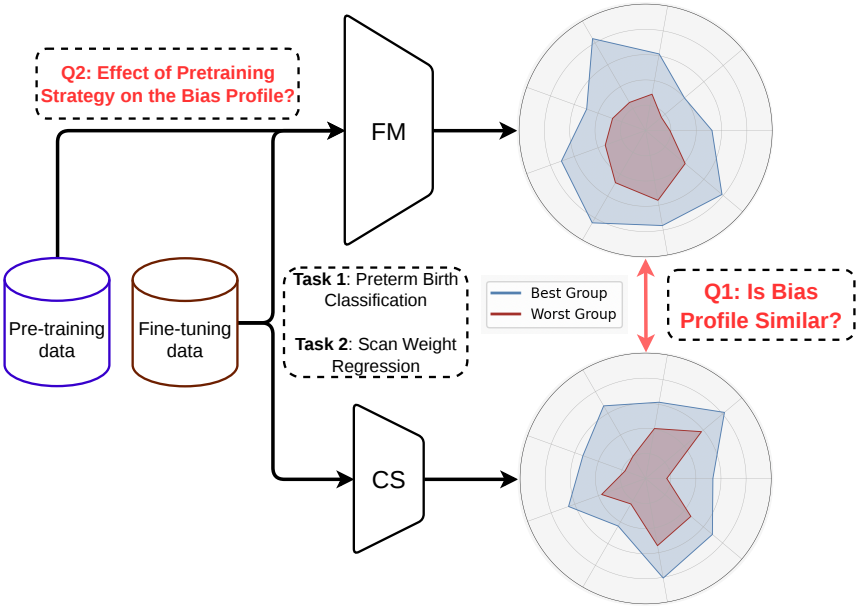

Fig. 1. Study Design. CS: classically supervised; FM: Foundation Model.

2 Method

A meaningful comparison between FMs and CS models requires: (i) a large unlabeled dataset for SSL pretraining; (ii) an identical, labeled fine-tuning set for the FM and CS models; and (iii) a CS baseline trained on that fine-tuning data alone. This work evaluates multiple FM variants with different pretraining corpora on the same test set across two tasks (Figure 1).

2.1 Bias Quantification

The test set is stratified by clinical and acquisition factors. For each factor, subgroups with a minimum sample size are retained, and the task metric is computed for each subgroup.

Visual analysis. Radar plots display the best- and worst-performing subgroup metric on each axis (one per factor), forming two polygons. Overlaying plots from two models enables direct comparison.

Quantitative metrics. Four complementary measures characterize a model's bias profile:

- *Absolute gap*: $|\text{worst} - \text{best}|$, averaged across factors.
- *Relative gap*: $|\text{worst} - \text{best}| \,/\, \text{best} \times 100\,\%$, averaged across factors.
- *Improved/worsened subgroups*: number of subgroups with a statistically significant performance gain or loss relative to a reference model, assessed by 1000-fold bootstrap resampling and Mann–Whitney test for classification and independent two-sample t-test in regression.
- *Bias similarity index*: a DSC value is computed between the two gap regions (area of the best/worse gaps) using the radar plots of two models, after normalizing so that the mean best-group metric is equal across models. This index is purely indicative to associate some familiar metric value of overlapping next to the two visuals of the bias profiles.

3 Experiment

3.1 Datasets

All dataset came from the Country X national fetal ultrasound database:

1. **FM Pretraining (PT) Data**: 13 million fetal ultrasound images from the same national cohort as described in a prior published work [Anon.].
2. **Spontaneous Preterm Birth Prediction (sPTB)**: 7862 transvaginal ultrasound images corresponding to n=3931 preterm births and n=3931 term births. The dataset was divided in 10 splits with 6290/786/786 samples for training/validation/test sets. The performance metric was the sensitivity at the threshold corresponding to a specificity of 85.

3. **Scan Weight Prediction**: 433,096 ultrasound images derived from 94,538 examinations conducted on 65,752 patients. The fetal weight ground truth at scan time was derived from the birth weight after using the Marsál [14] growth curve. The data was divided on a patient basis between training, validation, and test sets (85/5/10%). The performance metric was the Mean Relative Error (MRE) (%).

The finetuning datasets were stratified over the following studied factors: ethnicity, body-to-mass index (BMI), maternal hormonal treatment, conception method, cervical length, gestational age, birth weight, birth year, birth hospital, ultrasound device, pixel spacing. This resulted in a total of 55 and 38 subgroups for the sPTB and scan weight datasets, respectively.

3.2 Models

In order to ensure a fair comparison, we implement qualitative and validated CS baselines for each task:

1. **CS model for sPTB**: The CNN baseline, described in [18], predicts the probability of preterm birth from a cervical scan.
2. **CS model for Scan Weight**: CNN baseline as in [15] predicting the fetal scan weight from a set of head, abdomen, and femur standard planes.
3. **FM**: a ViT-B foundation model pretrained from scratch using a DINOv2 [17] self-supervised learning objective on fetal ultrasound images. All FM parts are finetuned: encoder, decoder, up to the final classification layer.

3.3 Pretraining Data Strategies

We evaluate four FM variants with the following pretraining regimes:

1. **FM2M**: Base FM pretrained on randomly selected 2 million ultrasound images out of the 13 million available.
2. **FM13M**: Base FM pretrained on all 13 million ultrasound images.
3. **FM2M E10**: Base FM pretrained on 2 million ultrasound images with a rare ultrasound device (Voluson E10) oversampled (500k).
4. **FM2M Cervix**: Base FM pretrained on 2 million ultrasound images with cervical scans oversampled, based on 120k images identified as cervical scans in the 13M PT dataset by a ResNet18 [9] trained to classify anatomical planes. These were oversampled five times (600k), and the remaining 1.4 million with a random sample from the rest of the PT data.

4 Results

We divide the analysis of our results along our two research questions: First, how different are the biases of the FM and CS model? Second, what is the impact of the pretraining data regime on bias?

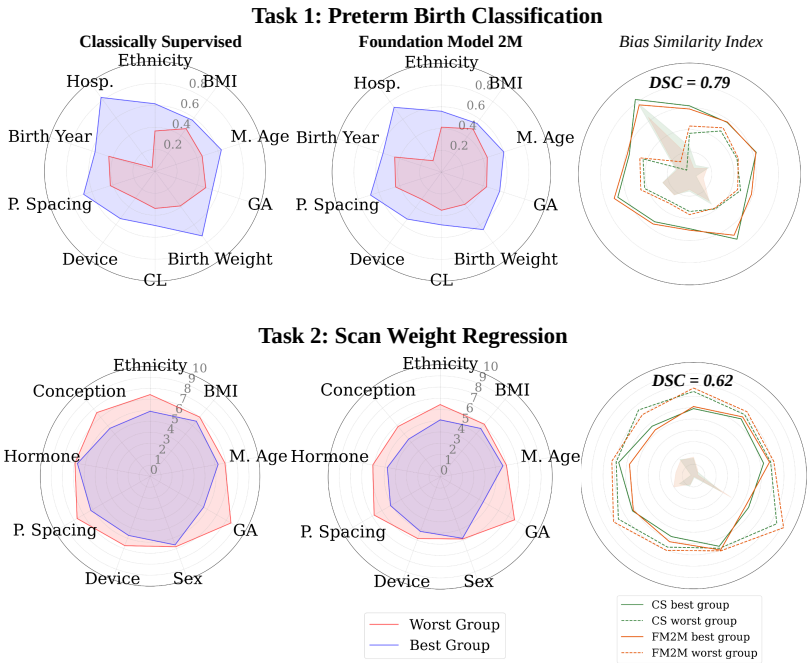

Fig. 2. Is the Foundation Model Less Biased Than Its Classically Supervised Counterpart? The blue radar gives the performance metric of the best subgroups; the red radar of the worst subgroups. Classification metric: sensitivity at 85% specificity (higher is better); regression metric: mean relative error (lower is better); S; Bias Similarity Index: DSC between the two gap areas (in the center) after scaling so that the mean metric of the best groups is the same for the two models.

4.1 Does the Bias Profile of a FM Resemble That of a CS Model?

The bias profiles of the FM2M and CS models, for both the sPTB classification and scan weight regression tasks, show strong similarity in shape (Figure 2). This is reflected in the overall relative gap between the best- and worst-performing groups across the CS and FM2M models for each task, at 37.3% and 33.8% for sPTB classification and 14.9% and 17.8% in scan weight prediction (Table 1). For the preterm birth prediction task, FM2M also failed to substantially improve overall performance compared to the CS model (0.516 sensitivity at 85% specificity for CS, 0.518 for FM2M).

Notably, the majority of the best- and worst-performing subgroups were the same across the FM and CS models (14 out of 20 in the classification task, 9 out of 18 in the regression task). There was also no consistent improvement in fine-grained subgroup performance with FM2M compared to the CS model.

Table 1. Global performance and mean subgroup performance shift (best vs. worst subgroup, averaged over all bias factors) for all methods. sPTB: sensitivity at 85% specificity; Scan Weight: mean relative error (%). Rel. Gap: $|\text{worst} - \text{best}|/\max \times 100$; Abs. Gap: $|\text{worst} - \text{best}|$. Models are highlighted **best** and second-best. In the case of a tie, both are marked as **best**. (**): different from CS baseline with p<.001.

Model	sPTB			Scan Weight		
	↑Sens.	Rel. gap	Abs. gap	↓MRE (%)	Rel. gap	Abs. gap
CS	51.6	37.3%	0.25	6.7±5.4	14.9%	1.14
FM2M	51.8	**33.8%**	**0.21**	**5.9±4.9****	17.8%	1.18
FM13M	53.1**	33.9%	0.22	6.0±5.1**	14.8%	**1.00**
FM2M E10	50.8**	34.4%	**0.21**	9.7±7.3**	**14.4%**	1.61
FM2M Cervix	**54.7****	34.5%	0.23	-	-	-

For the sPTB classification task, 17 subgroups (out of 55) improved significantly, while 17 worsened significantly (Table 2). These findings challenge the hypothesis that large-scale SSL pretraining reduces bias in downstream tasks, and suggest instead that bias is primarily a product of the fine-tuning data.

4.2 Does Pretraining Data Strategy Influence the Bias Profile of Foundation Models?

Here, we review the results associated to each pretraining strategy.

Scaling PT Data (FM13): The model pretrained with 13 million ultrasound images did result in a better bias profile in terms of improved/worsened subgroups (sPTB, Table 2) and in terms of relative gap (Scan Weight, Table 1). In sPTB, the effect of scaling was noticeable on certain ethnic and device subgroups, but subtle overall (Figure 3).

Oversampling rare device (FM2M E10): Oversampling the Voluson E10 resulted in an improvement (0.31 to 0.36) on this particular device in sPTB prediction. But it also led to a decrease in overall performance, especially on the regression task, where errors increased from 5.9% to 9.7%.

Oversampling Cervical Scans (FM2M Cervix): This strategy significantly improved both the global performance and subgroup performance of the foundation model for sPTB prediction, which uses a cervical scan as input. 39 subgroups improved, while only 7 worsened, compared to 17 and 17, respectively, for FM2M (Table 2). In particular, some worrying disparities in the CS model were mitigated (much more strongly than in FM13M), as evidenced by the noticeable drop in performance for babies born in 2016-2017 and after 2017 (Figure 4). However, in terms of the absolute and relative gaps between the best and worst subgroups, the bias profile was not improved compared to FM2M.

These results suggest that the benefit of SSL pretraining depends critically on the relevance of the pretraining data to the downstream task. Scaling and naive oversampling offer limited or even negative returns.

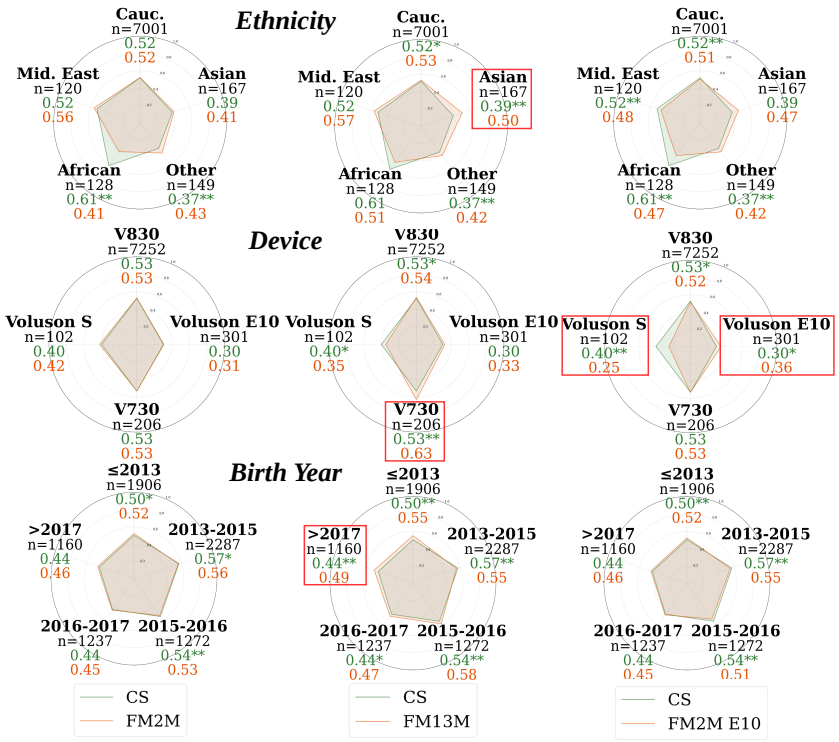

Fig. 3. Can the Pretraining Strategy Change the FM's sPTB Bias Profile? Sensitivity values are displayed; stars indicate significance of difference: (*) or (**) : p<.05 or .001.

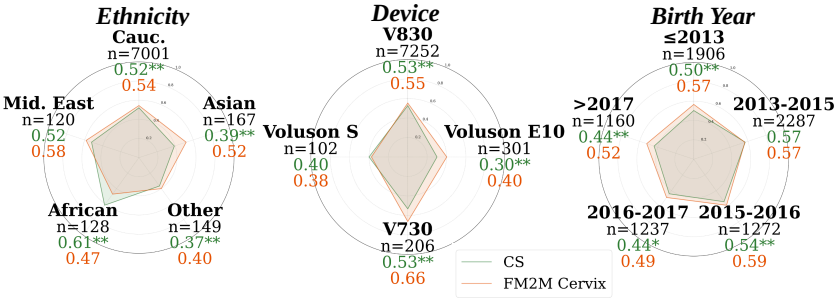

Fig. 4. Effect on sPTB Bias of Oversampling Cervical Scans in Pretraining. Sensitivity values are displayed; stars indicate significance of difference: (*) or (**) : p<.05 or .001.

Table 2. Number of subgroups (out of 55) with a statistically significant sensitivity gain or loss (at 85% specificity) relative to CS. Models are highlighted **best** and second-best.

Transition	Improved	Worsened
CS → FM2M	17	17
CS → FM13M	30	11
CS → FM2M E10	16	26
CS → FM2M Cervix	**39**	**8**

5 Discussion

Our **main findings** are:

1. Foundation models reproduce the bias profile of specialized models.
2. The pretraining data strategy can *moderately* influence the bias profile, in particular on smaller subgroups.
3. Oversampling task-related images outperformed scaling to 13 million images on sPTB prediction, both in global performance and subgroup fairness.

FMs mirrored the bias profile of specialized models. The best- and worst-performing subgroups largely coincided, and mean performance gaps were comparable in magnitude. These findings align with prior work [10, 21]. But to our knowledge, this is the first study to systematically compare foundation and CS models by a direct, subgroup-level performance analysis across a comprehensive set of factors. This contrasts with Glocker *et al.* [7], whose notion of bias mainly focused on the encoding of demographic attributes in latent representations rather than performance disparities across numerous subgroups.

Pretraining strategy did matter, and its effect depended on its alignment with the downstream task. This is supported by the differential benefit of FM2M across tasks. For example, fetal biometric scans (head, abdomen, femur), used as input for scan weight prediction, were abundant in the pretraining corpus, which may explain why FM2M improved performance on that task without any targeted curation. Cervical scans, by contrast, are rare in standard prenatal ultrasound collections, and generic pretraining provided little benefit for sPTB prediction. Oversampling cervical images during pretraining outperformed scaling to 13 million images, challenging the assumption that large-scale, diverse pretraining uniformly benefits any downstream task.

This study had limitations. We only investigated ViT-B models (90 million parameters), while some foundation models as DINOv2/v3, CLIP, are much larger (up to 1B parameters). Also, we focused on ultrasound imaging and two obstetric tasks. But such comparisons are inherently resource-intensive, requiring simultaneous access to huge unlabeled data pools, large labeled fine-tuning datasets, and the infrastructure to train multiple foundation models—making the scale achieved here (13 million PT images, 94k and 8k labeled samples) a strong baseline. We also ignored possible interaction effects between clinical/acquisition factors. Such intersectional analysis could reveal some additional

and more subtle differences between models. But given enough factors, two very similar bias profiles already tell us that the two bias behaviors correspond in their major directions, which was what we wanted to know.

Our findings have direct practical implications. FMs carry substantially higher computational costs than CS models; if their bias profiles are not demonstrably superior, the return on investment is questionable in high-data regimes. Certain configurations (e.g., rare-device oversampling) even degraded performance, underscoring that naïve pretraining data augmentation can be counterproductive.

In **conclusion**, SSL pretraining alone does not guarantee reduced bias relative to classically supervised models. Aligning pretraining data with the downstream task may be more impactful than indiscriminate scaling, and future work should rigorously compare this strategy with fine-tuning data curation.

Acknowledgments. The work was partly funded by the Novo Nordisk foundation through the project NNF0087102.

Disclosure of Interests. ANC, MT, AF, and MN hold shares in Prenaital ApS.

References

1. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al.: Flamingo: a visual language model for few-shot learning. Advances in neural information processing systems **35**, 23716–23736 (2022)
2. Balestriero, R., Ibrahim, M., Sobal, V., Morcos, A., Shekhar, S., Goldstein, T., Bordes, F., Bardes, A., Mialon, G., Tian, Y., et al.: A cookbook of self-supervised learning. arXiv preprint arXiv:2304.12210 (2023)
3. Bommasani, R., Hudson, D.A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M.S., Bohg, J., Bosselut, A., Brunskill, E., et al.: On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258 (2021)
4. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers). pp. 4171–4186 (2019)
5. FDA: Artificial intelligence-enabled device software functions: Lifecycle management and marketing submission recommendations (Jan 2025), https://www.fda.gov/regulatory-information/search-fda-guidance-documents/artificial-intelligence-enabled-device-software-functions-lifecycle-management-and-marketing
6. Geirhos, R., Jacobsen, J.H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., Wichmann, F.A.: Shortcut learning in deep neural networks. Nature Machine Intelligence **2**(11), 665–673 (2020)
7. Glocker, B., Jones, C., Roschewitz, M., Winzeck, S.: Risk of Bias in Chest Radiography Deep Learning Foundation Models. Radiology: Artificial Intelligence **5**(6), e230060 (Nov 2023). https://doi.org/10.1148/ryai.230060, https://pubs.rsna.org/doi/full/10.1148/ryai.230060

8. Goyal, P., Duval, Q., Seessel, I., Caron, M., Misra, I., Sagun, L., Joulin, A., Bojanowski, P.: Vision models are more robust and fair when pretrained on uncurated images without supervision. arXiv preprint arXiv:2202.08360 (2022)
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition (2015), https://arxiv.org/abs/1512.03385
10. Jin, R., Xu, Z., Zhong, Y., Yao, Q., QI, D., Zhou, S.K., Li, X.: Fairmedfm: fairness benchmarking for medical imaging foundation models. Advances in Neural Information Processing Systems **37**, 111318–111357 (2024)
11. Khan, M.O., Afzal, M.M., Mirza, S., Fang, Y.: How fair are medical imaging foundation models? In: Hegselmann, S., Parziale, A., Shanmugam, D., Tang, S., Asiedu, M.N., Chang, S., Hartvigsen, T., Singh, H. (eds.) Proceedings of the 3rd Machine Learning for Health Symposium. Proceedings of Machine Learning Research, vol. 225, pp. 217–231. PMLR (10 Dec 2023), https://proceedings.mlr.press/v225/khan23a.html
12. Lee, T., Puyol-Antón, E., Ruijsink, B., Shi, M., King, A.P.: A systematic study of race and sex bias in cnn-based cardiac mr segmentation. In: International Workshop on Statistical Atlases and Computational Models of the Heart. pp. 233–244. Springer (2022)
13. Lin, M., Weng, N., Mikolaj, K., Bashir, Z., Svendsen, M.B., Tolsgaard, M.G., Christensen, A.N., Feragen, A.: Shortcut learning in medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 623–633. Springer (2024)
14. Maršál, K., Persson, P.H., Larsen, T., Lilja, H., Selbing, A., Sultan, B.: Intrauterine growth curves based on ultrasonically estimated foetal weights. Acta paediatrica **85**(7), 843–848 (1996)
15. Mikołaj, K.W., Christensen, A.N., Taksøe-Vester, C.A., Feragen, A., Petersen, O.B., Lin, M., Nielsen, M., Svendsen, M.B.S., Tolsgaard, M.G.: Predicting abnormal fetal growth using deep learning. npj Digital Medicine **8**(1) (May 2025). https://doi.org/10.1038/s41746-025-01704-0, http://dx.doi.org/10.1038/s41746-025-01704-0
16. Moor, M., Banerjee, O., Abad, Z.S.H., Krumholz, H.M., Leskovec, J., Topol, E.J., Rajpurkar, P.: Foundation models for generalist medical artificial intelligence. Nature **616**(7956), 259–265 (2023)
17. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
18. Pegios, P., Sejer, E.P.F., Lin, M., Bashir, Z., Svendsen, M.B.S., Nielsen, M., Petersen, E., Christensen, A.N., Tolsgaard, M., Feragen, A.: Leveraging Shape and Spatial Information for Spontaneous Preterm Birth Prediction, p. 57–67. Springer Nature Switzerland (2023). https://doi.org/10.1007/978-3-031-44521-7_6, http://dx.doi.org/10.1007/978-3-031-44521-7$_6$
19. Puyol-Antón, E., Ruijsink, B., Mariscal Harana, J., Piechnik, S.K., Neubauer, S., Petersen, S.E., Razavi, R., Chowienczyk, P., King, A.P.: Fairness in cardiac magnetic resonance imaging: assessing sex and racial bias in deep learning-based segmentation. Frontiers in cardiovascular medicine **9**, 859310 (2022)
20. Seyyed-Kalantari, L., Zhang, H., McDermott, M.B., Chen, I.Y., Ghassemi, M.: Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. Nature medicine **27**(12), 2176–2182 (2021)
21. Yang, Y., Liu, Y., Liu, X., Gulhane, A., Mastrodicasa, D., Wu, W., Wang, E.J., Sahani, D., Patel, S.: Demographic bias of expert-level vision-language foundation models in medical imaging. Science Advances **11**(13), eadq0305 (2025)

22. Yfantidou, S., Spathis, D., Constantinides, M., Vakali, A., Quercia, D., Kawsar, F.: Using self-supervised learning can improve model fairness. In: Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. pp. 3942–3953 (2024)
23. Zhang, S., Metaxas, D.: On the challenges and perspectives of foundation models for medical image analysis. Medical image analysis **91**, 102996 (2024)