SynthFair: A Semi-Synthetic Medical Imaging Dataset to Propel Research on Bias Detection & Mitigation

Anonymous Author(s)
Affiliation
Address
email

1 Introduction

Artificial intelligence (AI) methods for analysing medical images have vast potential for facilitating more accurate and efficient clinical workflows [Topol, 2019]. In particular, in medical image computing (MIC), AI systems help tackle tasks such as pathology classification, anatomical segmentation, lesion delineation, among many others. However, rapid growth in this research domain contrasts starkly with the slow adoption rate of these technologies in clinical practice [Aristidou et al., 2022, Wenderott et al., 2025]. There are multiple reasons for this gap, but a central one is the diminished and often disparate performance of AI methods when transferred to clinical settings due to bias [Adamson and Smith, 2018, Sarkar et al., 2021, Cross et al., 2024, Lara et al., 2022, Ma et al., 2025].

Various studies have shown that AI methods often result in unequal performance across different subpopulations, which are typically defined in terms of protected attributes such as sex, gender, age, or race. This has led to the study of fairness and bias in AI [Barocas et al., 2023, Chen et al., 2023]. One of the primary factors driving performance disparities is biased data, such as the underrepresentation of specific subpopulations in training datasets [Dulaney and Virostko, 2024]. Multiple studies have shown that AI tools underperform on subgroups that were underrepresented in the training data. [Wiens et al., 2019, Larrazabal et al., 2020, Seyyed-Kalantari et al., 2021, Esmaeilzadeh, 2024].

In a similar vein, spurious correlations between variables of interest often induce shortcut learning, by which models appear effective in solving the task at hand, but do so relying on associations that are brittle or irrelevant to the task and may not be present at deployment [D'Amour et al., 2022, Banerjee et al., 2023]. For example, disease detection models have been found to rely on site-specific image acquisition features that are spuriously correlated with diagnostic labels [DeGrave et al., 2021, Zhang et al., 2023]. Furthermore, models may use features associated with protected attributes as harmful shortcuts [Brown et al., 2023, Gichoya et al., 2023, Xia et al., 2024]. Since data sets used to train AI models are often sourced from various acquisition sites with diverse demographics, complex confounders emerge, with no current systematic way to identify and disentangle their effects.

When datasets are not representative and confounding factors are omnipresent, the development of methods aimed at detecting and mitigating these forms of bias becomes paramount. However, the assessment and mitigation of such biases face significant challenges. Generating large and diverse annotated datasets in the medical domain is expensive and time-consuming, and most available datasets either lack demographic information or include a minimal fraction of images from minority subpopulations. This means that it is highly challenging to comprehensively evaluate the fairness of AI models and the impacts of bias detection and mitigation strategies. Furthermore, effectively mitigating bias also requires a deep understanding of the underlying sources of bias present in a dataset [Jones et al., 2024, 2025]. With the inherent complexity of hidden confounders in real-world data, it seems impossible to use such data to validate and benchmark bias detection and mitigation strategies. This calls for a more systematic and controlled approach that leverages synthetic data.

Recent work has proposed a framework for generating synthetic medical imaging datasets with fully traceable confounding features, which enables the objective and systematic evaluation of how

Submitted to AI for Science workshop (NeurIPS 2025).

underrepresentation biases and shortcut learning are addressed by bias detection and mitigation strategies [Stanley et al., 2023, 2024]. However, this framework is limited to a "toy" setup, as the generated datasets are not intended to represent real diseases or confounders. Expanding on this approach to model real-world subpopulations and pathologies would enable comprehensive and conclusive evaluations while benefiting from diverse representation and traceable biases. In this context, the recent development of powerful causal generative methods for creating synthetic images represents a promising way forward. Generative models have shown effectiveness both in producing high-quality images [Ho et al., 2020, Ribeiro et al., 2025, Dutt et al., 2025, Labs et al., 2025] and in improving the robustness of AI methods in MIC [Gowal et al., 2021, Roschewitz et al., 2025]. The key ability of causal generative models is that of creating counterfactual images [Pawlowski et al., 2020, Monteiro et al., 2023, Ribeiro et al., 2023], whereby starting from an observed image, one can generate images in which a particular attribute, e.g. pathology or protected group label, is modified.

To facilitate future research on bias and fairness challenges in medical AI models, we propose the SynthFair dataset. We employ state-of-the-art causal generative models to create an unprecedentedly large, diverse, annotated semi-synthetic dataset with traceable confounders, comprising over one million chest radiographs in total. We focus on chest X-rays because they are widely used for clinical assessment, publicly available, clinically relevant, and exhibit well-documented demographic and acquisition-related biases [Johnson et al., 2023], making this modality an ideal candidate to study bias and fairness in medical imaging AI. The dataset will be released with a balanced distribution across covariates, organised into a taxonomy based on their nature (c.f. Appendix A). This will enable the simulation of relevant subsets and real-world variations, inducing commonly encountered biases. The dataset will be accompanied by a comprehensive statistical analysis and quantitative evaluation of image fidelity, with concrete examples, tutorials, and ample opportunities for further extensions.

2 Dataset Requirements and Evaluation Criteria

AI task definition: What scientific question will the dataset enable? This dataset will enable researchers to develop and evaluate urgently needed bias detection and mitigation methods for medical imaging AI in a controlled yet highly realistic setting. SynthFair will facilitate research and experiments aimed at advancing our insights into the mechanics of bias amplification by AI models.

Dataset rationale: Why is this dataset the bottleneck? The lack of large annotated datasets, representative of the rich diversity of local and global populations, makes bias and fairness studies challenging and intersectional analyses practically impossible. Currently, bias detection and mitigation methods can only be comprehensively tested on toy or non-medical datasets, whose findings are known not to generalise well enough to the real-world medical domain to be clinically useful.

Acceleration potential: How will access to this dataset transform model development and downstream science? SynthFair will bridge the gap between toy datasets, where synthetic biases are known but overly simple, and real-world medical settings, where biases are hidden and often too complex to fully account for. While final validation on real-world data will always be required, SynthFair will propel research into bias detection and mitigation strategies by faithfully modelling clinical pathologies, subpopulations, and confounders in a traceable manner, contributing directly to the aim of developing robust and trustworthy AI systems for safety-critical applications.

Data-creation pathway: Where will the data come from? We employ one of the most advanced causal generative AI models pre-trained on the MIMIC-CXR database [Johnson et al., 2019]. We will expand the training data using images from multiple, geographically diverse sources, such as CheXpert [Irvin et al., 2019] (US), VinDr-CXR [Nguyen et al., 2022] (Vietnam), Padchest [Bustos et al., 2020] (Spain), BRAX [Reis et al., 2022] (Brazil), among others. These large and diverse datasets will allow us to generate images with a wide range of characteristics and scale the SynthFair dataset to an unprecedented size of over one million images across relevant demographic and clinical attributes. SynthFair will enable the generation of any desired target distribution for comprehensive analysis and better understanding of the effects of dataset bias on AI model performance.

Cost & scalability. This project will require high performance compute, using multiple high-end GPUs in parallel for training the generative models. We estimate the costs of compute to be 2,000 USD for every thousand GPU hours. Our methods scale naturally using parallel processing.

References

Adewole S Adamson and Avery Smith. Machine learning and health care disparities in dermatology. *JAMA dermatology*, 154(11):1247–1248, 2018.

Angela Aristidou, Rajesh Jena, and Eric J. Topol. Bridging the chasm between AI and clinical implementation. *The Lancet*, 399(10325):620, February 2022. ISSN 0140-6736, 1474-547X. doi: 10.1016/S0140-6736(22)00235-5. Publisher: Elsevier.

Imon Banerjee, Kamanasish Bhattacharjee, John L. Burns, Hari Trivedi, Saptarshi Purkayastha, Laleh Seyyed-Kalantari, Bhavik N. Patel, Rakesh Shiradkar, and Judy Gichoya. "Shortcuts" Causing Bias in Radiology Artificial Intelligence: Causes, Evaluation, and Mitigation. *Journal of the American College of Radiology*, 20(9):842–851, July 2023. ISSN 1546-1440. doi: 10.1016/j.jacr.2023.06.025.

Solon Barocas, Mortiz Hardt, and Arvind Narayana. *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press, 2023. URL http://www.fairmlbook.org.

Alexander Brown, Nenad Tomasev, Jan Freyberg, Yuan Liu, Alan Karthikesalingam, and Jessica Schrouff. Detecting shortcut learning for fair medical AI using shortcut testing. *Nature Communications*, 14(1):4314, July 2023. ISSN 2041-1723. doi: 10.1038/s41467-023-39902-7. Number: 1 Publisher: Nature Publishing Group.

Aurelia Bustos, Antonio Pertusa, Jose-Maria Salinas, and Maria De La Iglesia-Vaya. Padchest: A large chest x-ray image dataset with multi-label annotated reports. *Medical image analysis*, 66: 101797, 2020.

Richard J. Chen, Judy J. Wang, Drew F. K. Williamson, Tiffany Y. Chen, Jana Lipkova, Ming Y. Lu, Sharifa Sahai, and Faisal Mahmood. Algorithmic fairness in artificial intelligence for medicine and healthcare. *Nature Biomedical Engineering*, 7(6):719–742, June 2023. ISSN 2157-846X. doi: 10.1038/s41551-023-01056-8. Publisher: Nature Publishing Group.

James L Cross, Michael A Choma, and John A Onofrey. Bias in medical ai: Implications for clinical decision-making. *PLOS Digital Health*, 3(11):e0000651, 2024.

Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D Hoffman, et al. Underspecification presents challenges for credibility in modern machine learning. *Journal of Machine Learning Research*, 23(226):1–61, 2022.

Alex J. DeGrave, Joseph D. Janizek, and Su-In Lee. AI for radiographic COVID-19 detection selects shortcuts over signal. *Nature Machine Intelligence*, 3(7):610–619, July 2021. ISSN 2522-5839. doi: 10.1038/s42256-021-00338-7. Publisher: Nature Publishing Group.

Aidan Dulaney and John Virostko. Disparities in the Demographic Composition of The Cancer Imaging Archive. *Radiology: Imaging Cancer*, 6(1):e230100, January 2024. ISSN 2638-616X. doi: 10.1148/rycan.230100.

Raman Dutt, Pedro Sanchez, Yongchen Yao, Steven McDonagh, Sotirios A Tsaftaris, and Timothy Hospedales. Chexgenbench: A unified benchmark for fidelity, privacy and utility of synthetic chest radiographs. *arXiv preprint arXiv:2505.10496*, 2025.

Pouyan Esmaeilzadeh. Challenges and strategies for wide-scale artificial intelligence (ai) deployment in healthcare practices: A perspective for healthcare organizations. *Artificial Intelligence in Medicine*, 151:102861, 2024.

Judy Wawira Gichoya, Kaesha Thomas, Leo Anthony Celi, Nabile Safdar, Imon Banerjee, John D Banja, Laleh Seyyed-Kalantari, Hari Trivedi, and Saptarshi Purkayastha. Ai pitfalls and what not to do: mitigating bias in ai. *The British Journal of Radiology*, 96(1150):20230023, 2023.

Sven Gowal, Sylvestre-Alvise Rebuffi, Olivia Wiles, Florian Stimberg, Dan Andrei Calian, and Timothy A Mann. Improving robustness using generated data. *Advances in neural information processing systems*, 34:4218–4233, 2021.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison, 2019.

Alistair E. W. Johnson, Tom J. Pollard, Seth J. Berkowitz, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih ying Deng, Roger G. Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6(1):317, 2019. ISSN 2052-4463. doi: 10.1038/s41597-019-0322-0.

Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1, 2023.

Charles Jones, Daniel C. Castro, Fabio De Sousa Ribeiro, Ozan Oktay, Melissa McCradden, and Ben Glocker. A causal perspective on dataset bias in machine learning for medical imaging. *Nature Machine Intelligence*, pages 1–9, February 2024. ISSN 2522-5839. doi: 10.1038/s42256-024-00797-8. Publisher: Nature Publishing Group.

Charles Jones, Fabio De Sousa Ribeiro, Mélanie Roschewitz, Daniel C. Castro, and Ben Glocker. Rethinking fair representation learning for performance-sensitive tasks. In *The Thirteenth International Conference on Learning Representations*, 2025.

Dovile Juodelyte, Yucheng Lu, Amelia Jiménez-Sánchez, Sabrina Bottazzi, Enzo Ferrante, and Veronika Cheplygina. Source matters: Source dataset impact on model robustness in medical imaging. In *MICCAI International Workshop on Applications of Medical AI*, pages 105–115. Springer, 2024.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv preprint arXiv:2506.15742*, 2025.

María Agustina Ricci Lara, Rodrigo Echeveste, and Enzo Ferrante. Addressing fairness in artificial intelligence for medical imaging. *Nature Communications*, 13(1):4581, 2022. ISSN 2041-1723. doi: 10.1038/s41467-022-32186-3.

Agostina J. Larrazabal, Nicolás Nieto, Victoria Peterson, Diego H. Milone, and Enzo Ferrante. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proceedings of the National Academy of Sciences*, 117(23):12592–12594, 2020. doi: 10.1073/pnas.1919012117.

Haroui Ma, Francesco Quinzan, Theresa Willem, and Stefan Bauer. Ai alignment in medical imaging: Unveiling hidden biases through counterfactual analysis. *arXiv preprint arXiv:2504.19621*, 2025.

Miguel Monteiro, Fabio De Sousa Ribeiro, Nick Pawlowski, Daniel C. Castro, and Ben Glocker. Measuring axiomatic soundness of counterfactual image models. In *The Eleventh International Conference on Learning Representations*, 2023.

Ha Q Nguyen, Khanh Lam, Linh T Le, Hieu H Pham, Dat Q Tran, Dung B Nguyen, Dung D Le, Chi M Pham, Hang TT Tong, Diep H Dinh, et al. Vindr-cxr: An open dataset of chest x-rays with radiologist's annotations. *Scientific Data*, 9(1):429, 2022.

Nick Pawlowski, Daniel Coelho de Castro, and Ben Glocker. Deep structural causal models for tractable counterfactual inference. *Advances in neural information processing systems*, 33:857–869, 2020.

Eduardo P Reis, Joselisa PQ De Paiva, Maria CB Da Silva, Guilherme AS Ribeiro, Victor F Paiva, Lucas Bulgarelli, Henrique MH Lee, Paulo V Santos, Vanessa M Brito, Lucas TW Amaral, et al. Brax, brazilian labeled chest x-ray dataset. *Scientific Data*, 9(1):487, 2022.

Fabio De Sousa Ribeiro, Tian Xia, Miguel Monteiro, Nick Pawlowski, and Ben Glocker. High fidelity image counterfactuals with probabilistic causal models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 7390–7425, 23–29 Jul 2023.

Fabio De Sousa Ribeiro, Ben Glocker, et al. Demystifying variational diffusion models. *Foundations and Trends® in Computer Graphics and Vision*, 17(2):76–170, 2025.

Mélanie Roschewitz, Fabio De Sousa Ribeiro, Tian Xia, Galvin Khara, and Ben Glocker. Robust image representations with counterfactual contrastive learning. *Medical Image Analysis*, page 103668, 2025.

Rahuldeb Sarkar, Christopher Martin, Heather Mattie, Judy Wawira Gichoya, David J Stone, and Leo Anthony Celi. Performance of intensive care unit severity scoring systems across different ethnicities in the usa: a retrospective observational study. *The Lancet Digital Health*, 3(4): e241–e249, 2021.

Laleh Seyyed-Kalantari, Haoran Zhang, Matthew BA McDermott, Irene Y Chen, and Marzyeh Ghassemi. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nature medicine*, 27(12):2176–2182, 2021.

Emma A. M. Stanley, Matthias Wilms, and Nils D. Forkert. A Flexible Framework for Simulating and Evaluating Biases in Deep Learning-Based Medical Image Analysis. In Hayit Greenspan, Anant Madabhushi, Parvin Mousavi, Septimiu Salcudean, James Duncan, Tanveer Syeda-Mahmood, and Russell Taylor, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, Lecture Notes in Computer Science, pages 489–499, Cham, 2023. Springer Nature Switzerland. ISBN 978-3-031-43895-0. doi: 10.1007/978-3-031-43895-0_46.

Emma A. M. Stanley, Raissa Souza, Anthony J. Winder, Vedant Gulve, Kimberly Amador, Matthias Wilms, and Nils D. Forkert. Towards objective and systematic evaluation of bias in artificial intelligence for medical imaging. *Journal of the American Medical Informatics Association*, page ocae165, June 2024. ISSN 1527-974X. doi: 10.1093/jamia/ocae165.

Eric J. Topol. High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine*, 25(1):44–56, January 2019. ISSN 1546-170X. doi: 10.1038/s41591-018-0300-7. Publisher: Nature Publishing Group.

Katharina Wenderott, Jim Krups, Matthias Weigl, and Abigail R. Wooldridge. Facilitators and Barriers to Implementing AI in Routine Medical Imaging: Systematic Review and Qualitative Analysis. *Journal of Medical Internet Research*, 27(1):e63649, July 2025. doi: 10.2196/63649. Company: Journal of Medical Internet Research Distributor: Journal of Medical Internet Research Institution: Journal of Medical Internet Research Label: Journal of Medical Internet Research Publisher: JMIR Publications Inc., Toronto, Canada.

Jenna Wiens, Suchi Saria, Mark Sendak, Marzyeh Ghassemi, Vincent X Liu, Finale Doshi-Velez, Kenneth Jung, Katherine Heller, David Kale, Mohammed Saeed, et al. Do no harm: a roadmap for responsible machine learning for health care. *Nature medicine*, 25(9):1337–1340, 2019.

Tian Xia, Mélanie Roschewitz, Fabio De Sousa Ribeiro, Charles Jones, and Ben Glocker. Mitigating attribute amplification in counterfactual image generation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 546–556. Springer, 2024.

Ran Zhang, Dalton Griner, John W. Garrett, Zhihua Qi, and Guang-Hong Chen. Training certified detectives to track down the intrinsic shortcuts in COVID-19 chest x-ray data sets. *Scientific Reports*, 13(1):12690, August 2023. ISSN 2045-2322. doi: 10.1038/s41598-023-39855-3. Publisher: Nature Publishing Group.

Appendix

A SynthFair: Taxonomy of Covariates

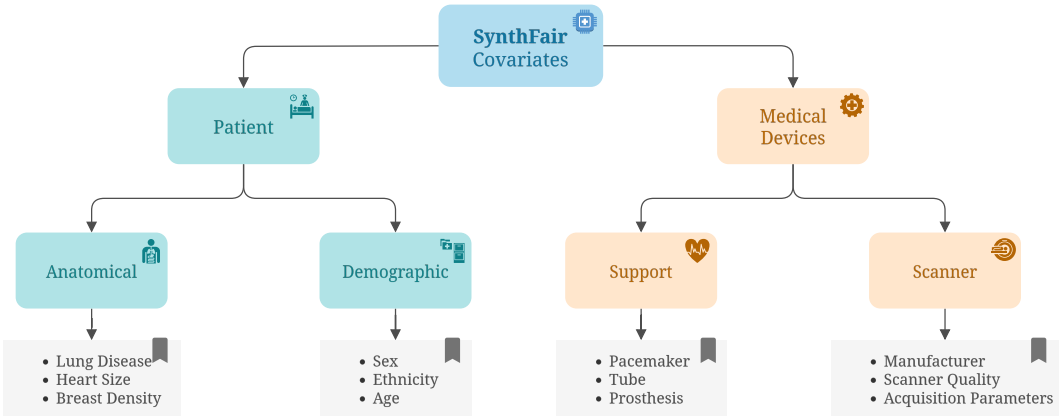

Figure 1: SynthFair taxonomy of different covariates from which counterfactual examples can be generated (inspired by the Medical Imaging Contextualized Confounder Taxonomy [Juodelyte et al., 2024]). These are organised based on their nature, mainly being covariates related to the patient itself or to external factors, such as the medical devices used to acquire the medical scans.

B Quantitative Evaluation: Counterfactual Image Quality

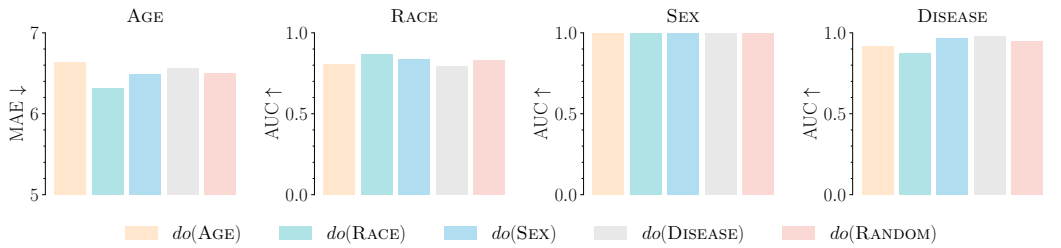

Figure 2: Quantitative evaluation of the quality of generated counterfactual images. Following Monteiro et al. [2023], we assess the axiomatic soundness of generated counterfactuals by measuring the *effectiveness* of different interventions (denoted by $do(\cdot)$) using external attribute classifiers/regressors, which help determine whether the interventions produce the expected changes in the target attributes. In all cases, we observe sufficiently good performance, confirming that the counterfactuals are faithful and can be useful for identifying and measuring counterfactual fairness and bias in downstream medical AI models. Further improvements are anticipated with an increase in model and dataset size.

C Qualitative Evaluation: Examples of Generated Counterfactuals

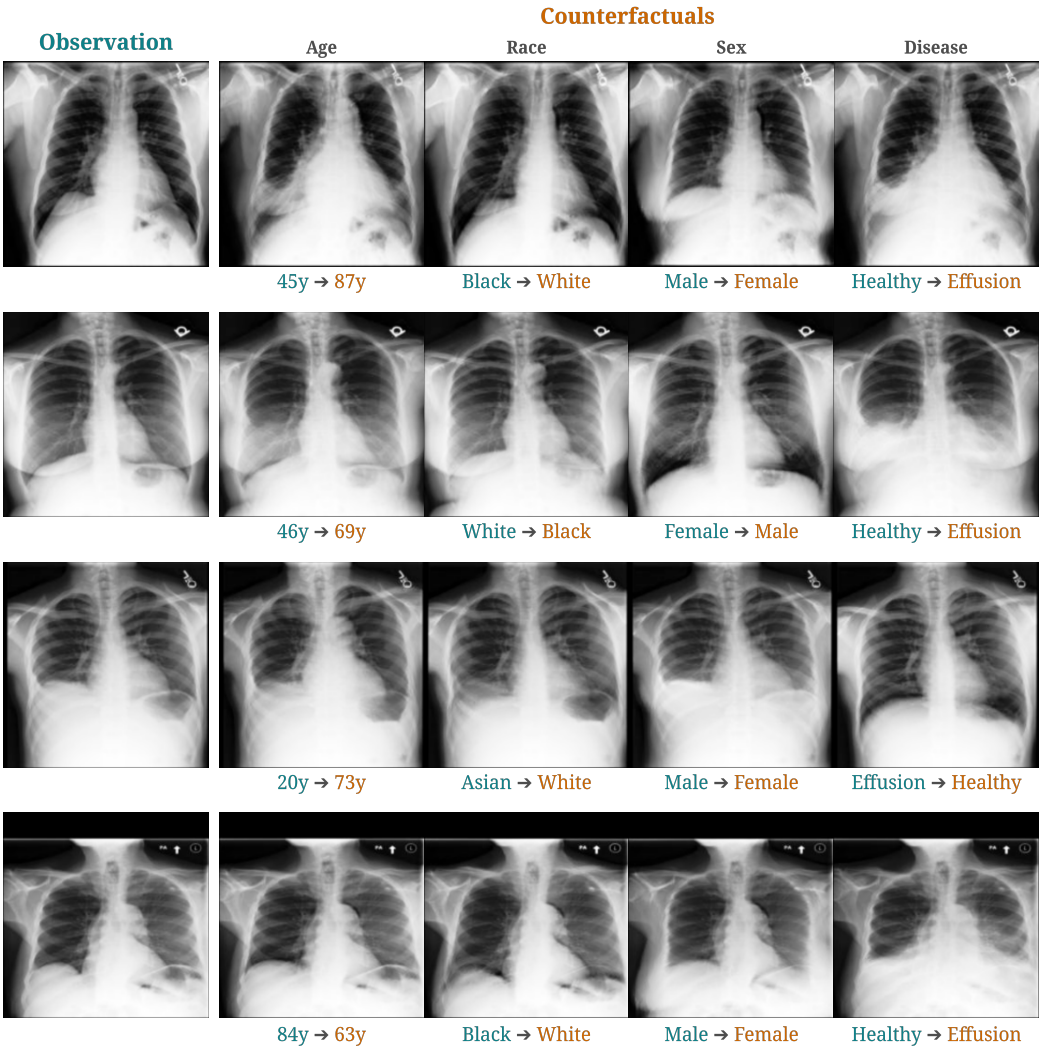

Figure 3: Examples of generated counterfactuals of different observations (left) using our model trained on MIMIC-CXR. In this case, *age*, *race*, *sex* and *disease* counterfactuals are shown. We observed faithful, localised changes while preserving the identity of the initial subject well.