# CF-Seg: Counterfactuals meet Segmentation

Raghav Mehta[1], Fabio De Sousa Ribeiro[1], Tian Xia[1],
Mélanie Roschewitz[1], Ainkaran Santhirasekaram[2],
Dominic C. Marshall[1], Ben Glocker[1]

[1] Imperial College London, UK.
[2] Imperial College Healthcare NHS Trust, UK.
raghav.mehta@imperial.ac.uk

**Abstract.** Segmenting anatomical structures in medical images plays an important role in the quantitative assessment of various diseases. However, accurate segmentation becomes significantly more challenging in the presence of disease. Disease patterns can alter the appearance of surrounding healthy tissues, introduce ambiguous boundaries, or even obscure critical anatomical structures. As such, segmentation models trained on real-world datasets may struggle to provide good anatomical segmentation, leading to potential misdiagnosis. In this paper, we generate counterfactual (CF) images to simulate how the same anatomy would appear in the absence of disease without altering the underlying structure. We then use these CF images to segment structures of interest, without requiring any changes to the underlying segmentation model. Our experiments on two real-world clinical chest X-ray datasets show that the use of counterfactual images improves anatomical segmentation, thereby aiding downstream clinical decision-making.

**Keywords:** Counterfactual Images · Anatomical Segmentation.

## 1 Introduction

Anatomical segmentation plays an important role in medical imaging, as it enables precise identification and delineation of organs and tissues, which in turn plays a pivotal role in treatment planning. For example, accurate lung segmentation in chest X-ray (CXR) helps in identifying the progression of lung disease and monitoring response to therapy [11,14]. Recent advances in deep learning have enabled precise and accurate segmentation of anatomical structures [28,27,18]. However, these methods often struggle in the presence of disease, where abnormalities can obscure or alter these structures [27]. For example, pathologic states, such as pleural effusion, edema, tumour or pneumonia, can alter the appearance of the lungs, increasing their opacity and making lung segmentation more difficult.

The ability to remove disease patterns from medical images while preserving anatomical structures is crucial for improving anatomical segmentation accuracy. In this work, we explore this possibility using recent advancements in counterfactual image generation models. Fundamentally, counterfactual (CF) images

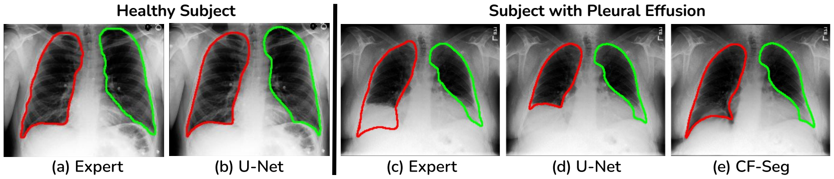

Fig. 1: Effect of disease on lung segmentation (red - right lung, green - left lung). **(a-b)** Healthy example: segmentation is relatively easy with high similarity between (a) expert segmentation and (b) automatic segmentation. **(c-d)** For a subject with pleural effusion, lungs are partially obscured, and segmentation is difficult with major differences between (c) expert segmentation and (d) automatic segmentation. Generating the counterfactual 'pseudo-healthy' image removes effusion without altering the underlying anatomy, on which (e) automatic segmentation becomes more similar to (c) expert segmentation.

represent 'what-if' scenarios, such as: *What would the patient's chest X-ray look like if there was no pleural effusion?* In this case, the underlying lung structure should remain unchanged, while the effusion would be removed. This would lead to a clearer anatomical representation, making lung segmentation easier for machine learning models (see Fig. 1).

Motivated by this, we propose a framework that leverages counterfactual image generation models to produce pseudo-healthy images from diseased images and then applies pre-trained segmentors to these counterfactual images. Notably, our approach does not require re-training the segmentation model, making it directly compatible with any pretrained segmentator. Our main contributions can be summarized as: (i) We propose CF-Seg, a novel framework leveraging healthy CF images, to improve anatomical segmentation in the presence of disease during inference, without requiring any modifications to the underlying segmentation model. (ii) We conduct a user study with two expert radiologists – reviewing 300 images each from MIMIC-CXR and PadChest [13,3] – finding that experts prefer lung segmentation generated with CF-Seg in comparison to publicly available (silver standard) segmentation masks [7]. (iii) We collect "ground-truth" expert segmentation for 140 images from healthy subjects and subjects with pleural effusion and find that using counterfactual images improves the lung segmentation performance substantially in the presence of effusion.

## 2 Related Work

To improve anatomical segmentation, one popular research avenue consists of leveraging anatomical knowledge by incorporating shape priors [22,17] or atlas-based segmentation [21,9,34]. However, these studies mainly focus on improving anatomical segmentation for healthy subjects without considering the effect of disease pathology on segmentation performance. On the other hand, in the context of pathology segmentation, methods have been proposed to either directly

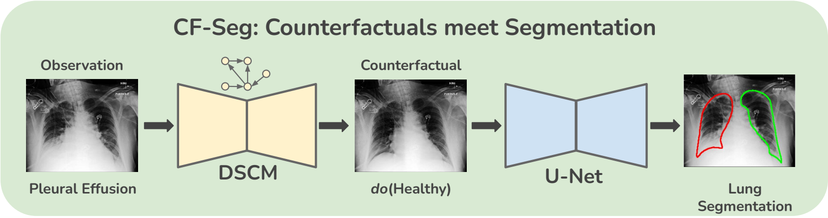

Fig. 2: Overview of the proposed CF-Seg framework. Instead of directly segmenting the diseased image, we first obtain a pseudo-healthy CF using a DSCM, and then use the U-net to obtain the segmentation from this counterfactual.

incorporate anatomical priors in neural networks [12,20]; or use generative models to synthesize normal pseudo-healthy anatomy and use difference maps to localize and segment anomalies or pathology [32,36,2]. The latter methods are closest to our work, as we also use pseudo-healthy images. However, in contrast to prior work, we specifically focus on *anatomy structure* segmentation in the presence of disease rather than on the segmentation of *pathology*.

Various approaches have been proposed for pseudo-healthy image generation, including GANs [15,32], VAEs [33,8], and diffusion models [30,1]. However, they do not explicitly model the underlying causal structure of the generative process. In this work, we utilize deep structural causal models (DSCMs) that integrate causal structures with deep generative models [23,19,5]. Specifically, we use a hierarchical variational auto-encoder (HVAE) based generative model proposed by Riberio et al. [5] to generate pseudo-healthy counterfactual images, which we then utilize for downstream segmentation of anatomical structure. Note that in this paper, we do not focus on proposing a new generative model, but rather focus on using existing generative models to improve anatomical segmentation.

Counterfactual images have been employed for various medical image analysis tasks and applications, such as data augmentation [10,35], contrastive learning [29], bias mitigation [16], explainability [6,4], and disease progression modeling [25,26]. However, their application for anatomical segmentation in the presence of disease pathology remains unexplored.

# 3 Methodology

## 3.1 Background on Deep Structural Causal Models

A Structural Causal Model (SCM) [24] consists of a set of endogenous variables $\mathbf{X} = \{X_i\}_{i=1}^n$ (e.g. medical scan, patient age, disease status etc), exogenous variables $\mathbf{U} = \{U_i\}_{i=1}^n$ (unobserved influences), and a set of functions $\mathbf{F} = \{f_i\}_{i=1}^n$ which define causal relationships (e.g. mechanism of disease). Each $X_i$ is determined by its parents $\mathbf{pa}_i$ (direct causes) and an exogenous variable $U_i$ via a structural equation: $X_i := f_i(\mathbf{pa}_i, U_i)$. SCMs enable the estimation of *counterfactuals*, which represent hypothetical scenarios given observed evidence. For example, one may query *"What would this patient's scan look like if there was no disease?"*. Counterfactual inference involves three steps: (i) Abduction: infer the posterior exogenous noise distribution given observed evidence $P_{\mathbf{U}|\mathbf{X}}$; (ii) Action: intervene on one or more of the endogenous variables $do(X_i \coloneqq x)$, such as disease status, to obtain a modified SCM $\mathcal{M}_x$; (iii) Prediction: use $\mathcal{M}_x$ and $P_{\mathbf{U}|\mathbf{X}}$ to generate a counterfactual. Deep SCMs [23,5] (DSCMs) and Neural Causal Models (NCMs) [31] provide a principled framework for using deep learning components in SCMs, thereby enabling tractable counterfactual inference of high-dimensional variables such as images. To estimate counterfactuals of chest X-rays, we use the Hierarchical Variational Autoencoder (HVAE) based causal mechanism from [5].

### 3.2 CF-Seg: Pseudo-healthy counterfactuals for segmentation

We propose a novel anatomical segmentation framework, designed to improve segmentation quality in the presence of disease pathology. An overview of our proposed CF-Seg framework is given in Fig. 2. In contrast to the standard framework, which directly uses U-net [28] for anatomical segmentation from input images, we incorporate a DSCM-based counterfactual image generation network [5] prior to anatomical segmentation. Specifically, irrespective of the subject's disease status (healthy or diseased), we first generate a pseudo-healthy counterfactual by intervening on the disease attribute. We then generate the final anatomical segmentation from this CF using standard U-Net. No changes are required for the employed segmentation model.

## 4 Experiments and Results

### 4.1 Datasets

We use two publicly available CXR datasets, namely, PadChest [3] and MIMIC-CXR [13]. These are large-scale datasets (more than 100k images) where associated image-level labels, such as disease pathology, are derived using a natural language processing toolbox from the associated reports. No associated anatomical segmentation marking is provided for these datasets. In such cases, we use the recently published CheXMask [7] database, which provides "silver-standard" anatomical segmentation boundaries derived using a pre-trained segmentation neural network, for both these datasets. Note that as these labels are *not* generated by an expert radiologist, they might not be reliable (See Sec. 4.3).

### 4.2 Implementation Details

In this work, we specifically focus on pleural effusion (PE), a disease inducing visible lung opacities in CXR. We primarily applied our method to lung segmentation in the presence of PE, the disease of interest for our clinical collaborators. However, any of the many other diseases associated with lung opacities could

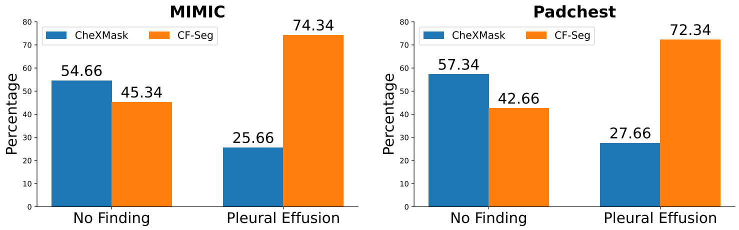

Fig. 3: Preference study results. Experts were shown CheXMask and our CF-Seg output masks side-by-side and asked to select their preferred segmentation mask. We here report the percentage of each segmentation method among the preferred segmentation masks on a total of 300 test images for each dataset.

have been considered, e.g., edema, ARDS, and pneumonia. Here, we restrict the study to images labeled as ‘no finding’ (NF) – as in healthy – or labeled as *only* ‘pleural effusion’ (PE). For PadChest, this leads to a total of 37,612 images (34,289 NF and 3,323 PE), respectively 62,620 images for MIMIC (56,615 NF and 6,005 PE). Datasets were split into train/test/valid subsets with a ratio of 70/20/10 for PadChest and 60/30/10 for MIMIC. All images were resized to 224×224 for PadChest and 256×256 for MIMIC.

For the HVAE, we use the same network architecture as proposed by Riberio et al. [5]. We found that involving multiple variables in the causal graph leads to better generation performance than only using disease status as the parent for image generation. As such, for MIMIC, we follow the recommended causal graph in Riberio et al. [5]; for PadChest, we consider three variables *scanner*, *sex*, and *disease*, assuming they are independent. In terms of the segmentation network, we follow the U-Net architecture [28] and utilize the segmentation masks provided by CheXMask to train the network. During inference time, we utilize this trained U-Net as a part of our proposed CF-Seg framework (See Fig. 2).

### 4.3 Experiment-1: Preference study

We first conducted a preference study with two different expert doctors (one for each dataset) with 7 years of experience. Specifically, they were asked to choose their preferred segmentation for a specific image. Two available options were (i) the segmentation from CheXMask (‘silver standard’), and (ii) the segmentation generated by our CF-Seg method. For this experiment, we randomly selected 150 healthy (NF) and 150 PE images from each dataset. Experts were shown the same image with two different segmentations side-by-side (without revealing which one is which), and were asked to select their preference. We also randomly changed the order (left or right) of these segmentations to make sure that the selection was not biased towards one or another.

From the results plotted in Fig. 3, we can observe that for both datasets, for healthy patient images (No-Finding), experts only marginally prefer the segmen-

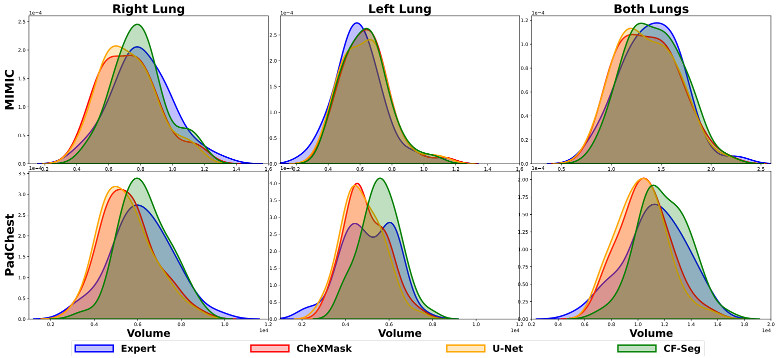

Fig. 4: For (Top) MIMIC and (Bottom) PadChest, density plots of lung volumes for left-, right- and both-lungs, measured by different segmentation methods.

tation provided by CheXMask over CF-Seg. This is expected as the underlying U-Net network was trained using CheXMask. On the contrary, for more than 70% of Pleural Effusion images, experts prefer segmentation provided by CF-Seg in comparison to CheXMask. This validates the usefulness of our proposed framework in clinical applications.

### 4.4 Experiment-2: Comparison against expert segmentations

Next, we obtain "ground-truth" lung segmentations manually drawn by expert doctors, henceforth known as expert segmentation (Expert). Specifically, we randomly chose 50 PE images and 20 healthy (NF) images from each dataset. PE test images include images with multiple findings (e.g., PE + cardiomegaly), to verify the robustness of CF-Seg in the presence of other findings. For each dataset, one expert marked lung boundaries. For healthy images, the Dice score between CheXMask labels and the Expert was more than 0.95, while for PE images, it was around 0.87. These results corroborate our findings from Sec. 4.3, reiterating that CheXMask labels are clinically useful for healthy subjects only, while for subjects with PE, they might not be reliable.

To assess the clinical value of generated segmentation maps, we first compare lung volumes extracted from different segmentation maps. Specifically, (i) Expert annotations, (ii) CheXMask (silver standard), (iii) predictions from a U-Net trained on CheXMask data, (iv) our CF-Seg framework, where we first generate the pseudo-healthy CF, then use it to generate the segmentation maps (using the same U-Net as in (iii)). In Fig. 4, we visualize kernel density estimate (KDE) plots of lung volume distribution on PE images, for both datasets. This figure reveals that for both datasets, CF-Seg follows the distribution of Expert more closely compared to the U-Net or CheXMask labels. Both CheXMask and U-Net undersegment lungs in comparison to the Expert, while that is not the case for the CF-Seg. We also observe that the volume difference between Expert and CheXMask is more prominent for the right lung compared to the left lung. We

Table 1: Performance measured as mean Dice coefficient (%) between Expert and CF-Seg/U-Net for right lung, left lung, and both lungs together. P-values between U-Net and CF-Seg are $\leq 0.05$ for the right lung (column 1) and both lungs (column 3), while this is not the case for the left lung (column 2). All reported results are for 50 PE images only.

| Dataset | Method | Right Lung | | Left Lung | | Both Lungs | |
|---|---|---|---|---|---|---|---|
| | | All | $\Delta V^+$ | All | $\Delta V^+$ | All | $\Delta V^+$ |
| MIMIC | U-Net | 89.16 | 85.86 | 91.47 | 91.11 | 90.30 | 87.45 |
| | CF-Seg | **90.57** | **88.52** | **91.49** | **92.13** | **91.05** | **90.03** |
| PadChest | U-Net | 89.53 | 84.71 | 91.43 | 91.96 | 90.52 | 87.87 |
| | CF-Seg | **90.64** | **88.83** | **91.62** | **92.37** | **91.20** | **90.31** |

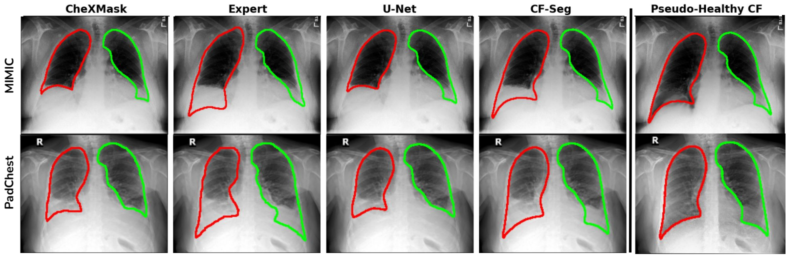

Fig. 5: Qualitative comparison of lung segmentation masks generated by different methods for two examples from (Top) MIMIC and (Bottom) PadChest. Left-to-Right: CheXMask, Expert, U-Net, CF-Seg (on original image), and CF-Seg overlaid on pseudo-healthy CF. It is clearly visible that in the case of pathology, CheXMask undersegments lungs (especially the right lung shown in red) in comparison to Expert. The U-Net generates masks similar to CheXMask, while CF-Seg outputs are much closer to the masks of Expert.

hypothesize that this might be because the heart obscures a larger part of the left lung compared to the right lung, and as such, the effect of pleural effusion might be easier to observe in the right lung.

Next, we measure Dice scores between Expert and CF-Seg as well as U-Net. For each dataset, we report the average Dice score on the corresponding test set (50 images). In addition, we report Dice scores for images with a positive volume difference between Expert and CheXMask labels ($\Delta V^+$), as these represent undersegmented cases (more than 35) which have been corrected by experts. This enables us to assess the benefit of CF-Seg in cases where experts disagree the most with CheXMask. In Table 1, we can see that CF-Seg improves performance in all cases. We find that the difference in performance is more prominent for right lung segmentation, especially on $\Delta V^+$ images. This is expected as there was a higher volume difference (and as such, more correction) between CheXMask and Expert for the right lung compared to the left lung. This is also clearly visible in

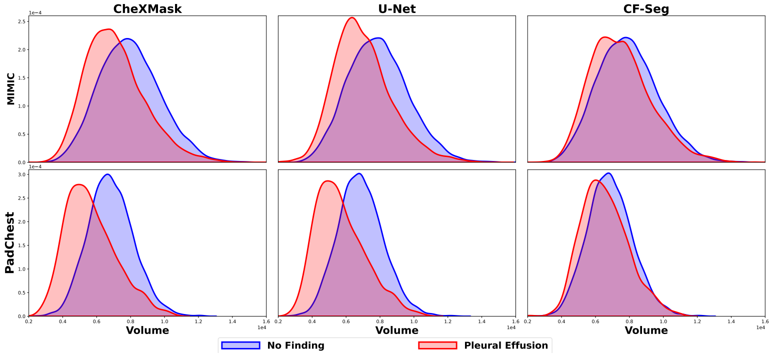

Fig. 6: Comparison of right lung volume density plots measured by different segmentation methods between healthy subjects (no finding) and subjects with pleural effusion. (Top) MIMIC and (Bottom) PadChest dataset.

the qualitative results presented in Fig. 5. Crucially, comparing the performance of U-Net and CF-Seg, these results demonstrate that applying pseudo-healthy counterfactuals prior to segmentation improves anatomical segmentation substantially, without making any changes to the underlying segmentation model.

### 4.5 Experiment-3: CF-Seg evaluation on large-scale datasets

In this section, we aim to evaluate the proposed CF-Seg on the full test sets from MIMIC and PadChest (see Sec. 4.1). However, no ground truth annotations are available for these sets, and obtaining manual annotations for such large sets (>5k) is infeasible in practice due to the burden on experts. Hence, we utilize a proxy method to evaluate segmentations by comparing volume density plots between healthy (NF) and disease images (PE). We hypothesize that irrespective of the disease stage (NF or PE), lung volume density across the whole population should remain similar. From Fig. 6, we observe that right lung volume density plots between NF and PE patient images overlap considerably for CF-Seg, while that is not the case for CheXMask and U-Net. The difference in mean volume between PE and NF, for CheXMask (MIMIC: 859, PadChest: 1177) and U-Net (MIMIC: 893, PadChest: 1189) is comparatively higher than CF-Seg (MIMIC: 398, PadChest: 306). This shows that CF-Seg is better able to segment underlying true anatomy compared to either CheXMask or U-Net.

## 5 Conclusion

We proposed CF-Seg, a general counterfactual segmentation framework for improving anatomical segmentation in the presence of disease pathology. This framework was motivated by the fact that obtaining accurate anatomical segmentations in the presence of disease is challenging, as abnormalities can obscure or alter anatomical structures. Our two-stage approach employed counterfactual generative modeling to first infer pseudo-healthy counterfactuals of medical scans, which are then utilized to more easily segment the structure of interest. We conducted extensive experiments using two large publicly available CXR datasets, namely MIMIC and PadChest, and found that, on images with pleural effusion, our counterfactual lung segmentations are more accurate and consistently preferred by experts in a user study, despite being trained only on "silver-standard" undersegmenting segmentation masks. Our results underscore the potential of counterfactual inference for broader clinical applications. We make our code and segmentation masks publicly available: https://github.com/biomedia-mira/CF-Seg.

## Acknowledgment

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement 101080302. M.R. is funded by an Imperial College London President's PhD Scholarship and a Google PhD Fellowship. D. M. is supported by an MRC clinical research training fellowship (award MR/Y000404/1) and the Mittal Fund at Cleveland Clinic Philanthropy. B.G. received support from the Royal Academy of Engineering as part of his Kheiron/RAEng Research Chair. B.G. and F.d.S.R. acknowledge the support of the UKRI AI programme, and the EPSRC, for CHAI - EPSRC Causality in Healthcare AI Hub (grant no. EP/Y028856/1).

### Disclosure of interests

B.G. is part-time employee of DeepHealth. No other competing interests.

## References

1. Baugh, M., Reynaud, H., Marimont, S.N., Cechnicka, S., Müller, J.P., Tarroni, G., Kainz, B.: Image-conditioned diffusion models for medical anomaly detection. In: International Workshop on Uncertainty for Safe Utilization of Machine Learning in Medical Imaging. pp. 117–127. Springer (2024)
2. Baugh, M., Tan, J., Müller, J.P., Dombrowski, M., Batten, J., Kainz, B.: Many tasks make light work: Learning to localise medical anomalies from multiple synthetic tasks. In: MICCAI 2023. pp. 162–172. Springer (2023)
3. Bustos, A., Pertusa, A., Salinas, J.M., De La Iglesia-Vaya, M.: Padchest: A large chest x-ray image dataset with multi-label annotated reports. Medical image analysis **66**, 101797 (2020)
4. Cohen, J.P., Brooks, R., En, S., Zucker, E., Pareek, A., Lungren, M.P., Chaudhari, A.: Gifsplanation via latent shift: a simple autoencoder approach to counterfactual generation for chest x-rays. In: MIDL. pp. 74–104 (2021)
5. De Sousa Ribeiro, F., Xia, T., Monteiro, M., Pawlowski, N., Glocker, B.: High fidelity image counterfactuals with probabilistic causal models. ICML (2023)

6. Fathi, N., Kumar, A., Nichyporuk, B., Havaei, M., Arbel, T.: Decodex: Confounder detector guidance for improved diffusion-based counterfactual explanations. MIDL (2024)
7. Gaggion, N., Mosquera, C., Mansilla, L., Saidman, J.M., Aineseder, M., Milone, D.H., Ferrante, E.: Chexmask: a large-scale dataset of anatomical segmentation masks for multi-center chest x-ray images. Scientific Data **11**(1), 511 (2024)
8. Hassanaly, R., Brianceau, C., Solal, M., Colliot, O., Burgos, N.: Evaluation of pseudo-healthy image reconstruction for anomaly detection with deep generative models: Application to brain fdg pet. Machine Learning for Biomedical Imaging **2**, 611–656 (2024)
9. Huang, H., Zheng, H., Lin, L., Cai, M., Hu, H., Zhang, Q., Chen, Q., Iwamoto, Y., Han, X., Chen, Y.W., et al.: Medical image segmentation with deep atlas prior. IEEE Transactions on Medical Imaging **40**(12), 3519–3530 (2021)
10. Ilse, M., Tomczak, J.M., Forré, P.: Selecting data augmentation for simulating interventions. In: ICML. pp. 4555–4562. PMLR (2021)
11. Jaeger, S., Karargyris, A., Candemir, S., Folio, L., Siegelman, J., Callaghan, F., Xue, Z., Palaniappan, K., Singh, R.K., Antani, S., et al.: Automatic tuberculosis screening using chest radiographs. IEEE transactions on medical imaging **33**(2), 233–245 (2013)
12. Jaus, A., Seibold, C., Reiß, S., Heine, L., Schily, A., Kim, M., Bahnsen, F.H., Herrmann, K., Stiefelhagen, R., Kleesiek, J.: Anatomy-guided pathology segmentation. In: MICCAI 2024. pp. 3–13. Springer (2024)
13. Johnson, A.E., Bulgarelli, L., Shen, L., Gayles, A., Shammout, A., Horng, S., Pollard, T.J., Hao, S., Moody, B., Gow, B., et al.: Mimic-iv, a freely accessible electronic health record dataset. Scientific data **10**(1), 1 (2023)
14. Khomduean, P., Phuaudomcharoen, P., Boonchu, T., Taetragool, U., et al.: Segmentation of lung lobes and lesions in chest ct for the classification of covid-19 severity. Scientific Reports **13**(1), 20899 (2023)
15. Kocaoglu, M., Snyder, C., Dimakis, A.G., Vishwanath, S.: Causalgan: Learning causal implicit generative models with adversarial training. ICLR (2017)
16. Kumar, A., Fathi, N., Mehta, R., Nichyporuk, B., Falet, J.P.R., Tsaftaris, S., Arbel, T.: Debiasing counterfactuals in the presence of spurious correlations. In: MICCAI Workshop on Clinical Image-Based Procedures. pp. 276–286. Springer (2023)
17. Larrazabal, A.J., Martínez, C., Glocker, B., Ferrante, E.: Post-dae: anatomically plausible segmentation via post-processing with denoising autoencoders. IEEE Transactions on Medical Imaging **39**(12), 3813–3820 (2020)
18. Liu, W., Luo, J., Yang, Y., Wang, W., Deng, J., Yu, L.: Automatic lung segmentation in chest x-ray images using improved u-net. Scientific Reports **12**(1), 8649 (2022)
19. Monteiro, M., Ribeiro, F.D.S., Pawlowski, N., Castro, D.C., Glocker, B.: Measuring axiomatic soundness of counterfactual image models. In: The Eleventh International Conference on Learning Representations (2023)
20. Müller, P., Meissen, F., Brandt, J., Kaissis, G., Rueckert, D.: Anatomy-driven pathology detection on chest x-rays. In: MICCAI. pp. 57–66. Springer (2023)
21. Navarro, F., Shit, S., Ezhov, I., Paetzold, J., Gafita, A., Peeken, J.C., Combs, S.E., Menze, B.H.: Shape-aware complementary-task learning for multi-organ segmentation. In: Machine Learning in Medical Imaging. pp. 620–627. Springer (2019)
22. Oktay, O., Ferrante, E., Kamnitsas, K., Heinrich, M., Bai, W., et al.: Anatomically constrained neural networks (acnns): application to cardiac image enhancement and segmentation. IEEE Transactions on Medical Imaging **37**(2), 384–395 (2017)

23. Pawlowski, N., Coelho de Castro, D., Glocker, B.: Deep structural causal models for tractable counterfactual inference. NeurIPS **33**, 857–869 (2020)
24. Pearl, J.: Causality. Cambridge university press (2009)
25. Pombo, G., Gray, R., Cardoso, M.J., Ourselin, S., Rees, G., Ashburner, J., Nachev, P.: Equitable modelling of brain imaging by counterfactual augmentation with morphologically constrained 3d deep generative models. Medical Image Analysis **84**, 102723 (2023)
26. Puglisi, L., Alexander, D.C., Ravì, D.: Enhancing spatiotemporal disease progression models via latent diffusion and prior knowledge. In: MICCAI. pp. 173–183. Springer (2024)
27. Reamaroon, N., Sjoding, M.W., Derksen, H., Sabeti, E., Gryak, J., Barbaro, R.P., Athey, B.D., Najarian, K.: Robust segmentation of lung in chest x-ray: applications in analysis of acute respiratory distress syndrome. BMC medical imaging **20**, 1–13 (2020)
28. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: MICCAI. pp. 234–241. Springer (2015)
29. Roschewitz, M., de Sousa Ribeiro, F., Xia, T., Khara, G., Glocker, B.: Counterfactual contrastive learning: robust representations via causal image synthesis. In: MICCAI Workshop on Data Engineering in Medical Imaging. pp. 22–32 (2024)
30. Sanchez, P., Tsaftaris, S.A.: Diffusion causal models for counterfactual estimation. Conference on Causal Learning and Reasoning (2022)
31. Xia, K.M., Pan, Y., Bareinboim, E.: Neural causal models for counterfactual identification and estimation. In: ICLR 2023 (2023)
32. Xia, T., Chartsias, A., Tsaftaris, S.A.: Pseudo-healthy synthesis with pathology disentanglement and adversarial learning. Medical Image Analysis **64**, 101719 (2020)
33. Yang, M., Liu, F., Chen, Z., Shen, X., Hao, J., Wang, J.: Causalvae: Disentangled representation learning via neural structural causal models. In: CVPR. pp. 9593–9602 (2021)
34. Yao, J., Cai, J., Yang, D., Xu, D., Huang, J.: Integrating 3d geometry of organ for improving medical image segmentation. In: MICCAI 2019. pp. 318–326. Springer (2019)
35. Zhou, X., Wu, O., Ng, M.K.: Implicit counterfactual data augmentation for robust learning. arXiv:2304.13431 (2023)
36. Zimmerer, D., Full, P.M., Isensee, F., Jäger, P., Adler, T., Petersen, J., Köhler, G., Ross, T., Reinke, A., Kascenas, A., et al.: Mood 2020: A public benchmark for out-of-distribution detection and localization on medical images. IEEE Transactions on Medical Imaging **41**(10), 2728–2738 (2022)