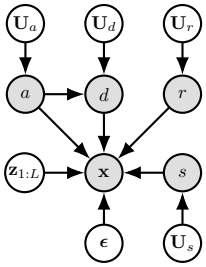
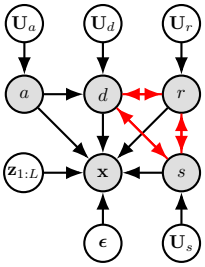
A Supplementary Material

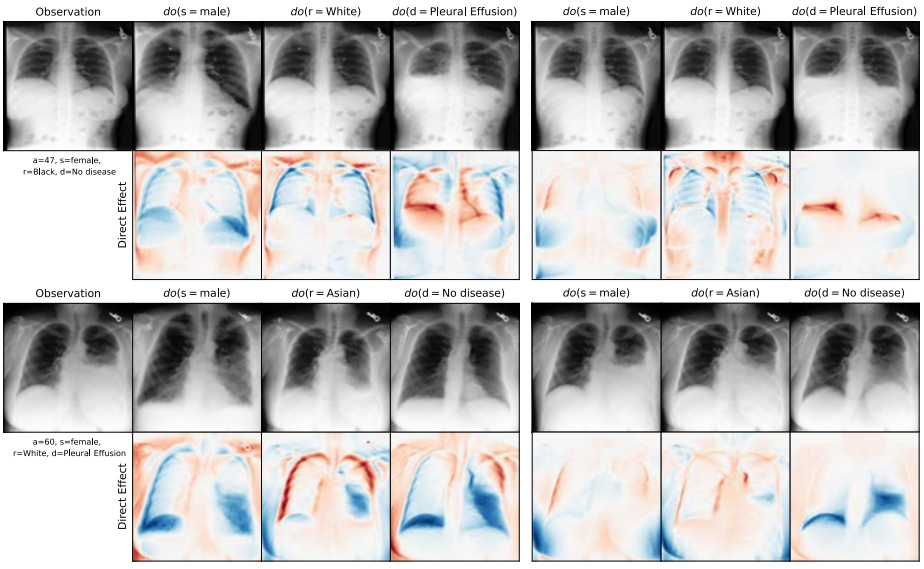


(a) Assumed causal graph for MIMIC-CXR. Variables in the causal graph are: age (a), sex (s), race (r), disease (d) (pleural effusion) and chest X-ray ($\mathbf{x}$).



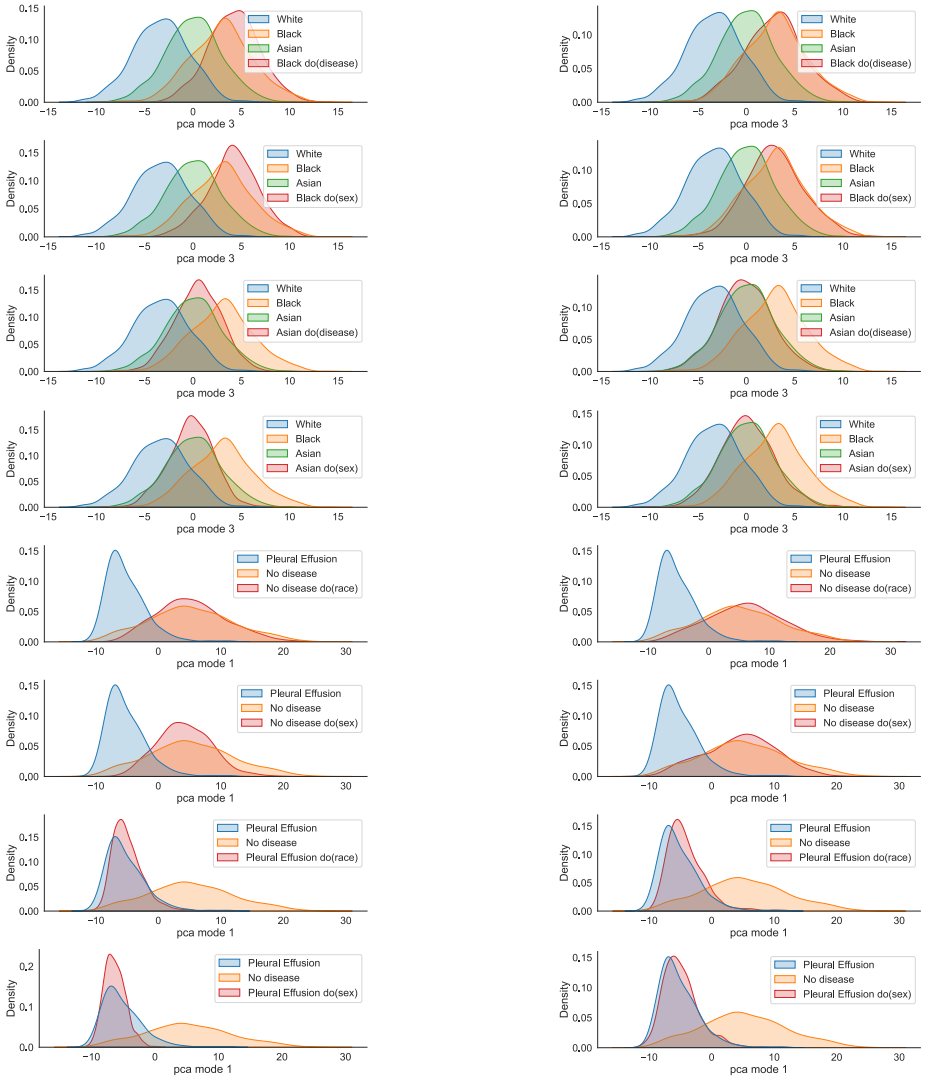
(b) With attribute amplification, the assumed causal graph is violated. For instance, $do(s)$ affecting d could make s a parent of d and vice versa.

Fig. A1: Illustration of how attribute amplification may violate the causal graph pre-defined for the DSCM which may lead to spurious correlations between protected characteristics and disease status encoded in the counterfactual images.



(a) Hard-CFT (b) Soft-CFT

Fig. A2: Generated CFs with (a) Hard-CFT and (b) Soft-CFT. Top rows show original image $\mathbf{x}$ and CFs $\widetilde{\mathbf{x}}$; bottom rows show direct effect of CFs, i.e. $\widetilde{\mathbf{x}} - \mathbf{x}$.



(a) Hard-CFT (b) Soft-CFT

Fig. A3: Marginal distribution of PCA modes of pretrained embeddings from a multi-task model predicted all attributes. We plot embeddings of real data along side with generated counterfactuals of various subgroups. We can see that when training with Hard-CFT (left) there is a distribution shift between real images and images after intervention (red). Conversely, this shift is mitigated when using our proposed soft counterfactual fine-tuning (Soft-CFT, right).