High Fidelity Image Counterfactuals with Probabilistic Causal Models

Imperial College London

BioMedIA

Fabio De Sousa Ribeiro Tian Xia Miguel Monteiro Nick Pawlowski[†] Ben Glocker

[†]Microsoft Research Cambridge, UK

ICML International Conference On Machine Learning

1. Overview

Research Question: How can we generate **plausible** high-fidelity **counterfactuals** of real data, and how do we **evaluate** them?

Motivation:

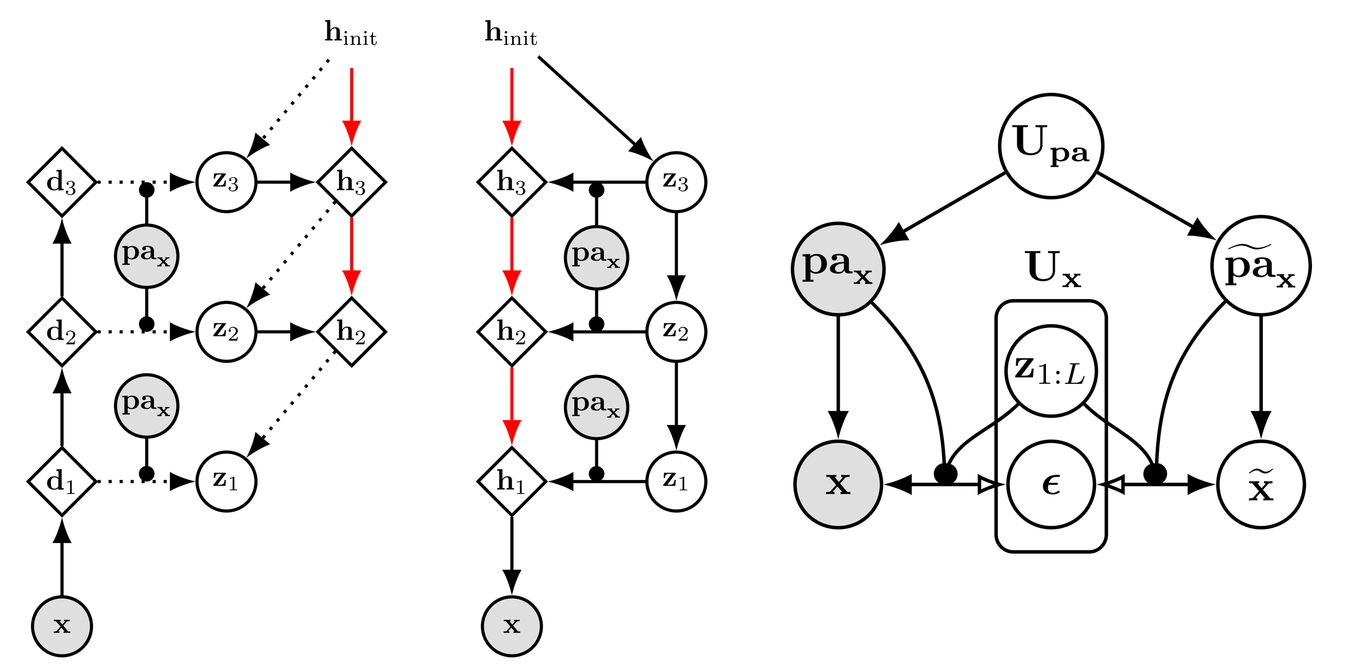
- Counterfactuals are useful for **explainability**, **data aug.**, **fairness** etc
- High fidelity counterfactuals of **structured variables** are challenging
- **Identifiability** guarantees are **absent** in the general case
- Prior work is mostly theoretical, we take a pragmatic approach

Contribution:

- **Hierarchical causal mechanisms** for high-dim **structured variables**
- **Latent mediator** model for **direct**, **indirect** and **total** effect estimation
- **Counterfactual training** to mitigate ignored counterfactual conditioning
- Demonstrate the **axiomatic soundness** of our inferred counterfactuals

2. Causal Mechanisms: Hierarchical VAEs

○ : observed ○ : stochastic ◇ : deterministic → : conv block → : shared block → : conditioning ⇢ : inference



(a) Inference Model **(b)** Generative Model **(c) Exogenous Prior:** $p_\theta(\mathbf{z}_{1:L})$

Figure: Twin network of our deep SCM (c) induced by an exogenous prior (b), where $\mathbf{z}_{1:L}$ is part of $\mathbf{x}$'s exogenous noise $\mathbf{U}_\mathbf{x}$. Directly compatible with Pawlowski et al. (2020)'s framework.

- **Abduction:**

$$\mathbf{z}_{1:L} \sim q_\phi(\mathbf{z}_{1:L} \mid \mathbf{x}, \mathbf{pa_x})$$

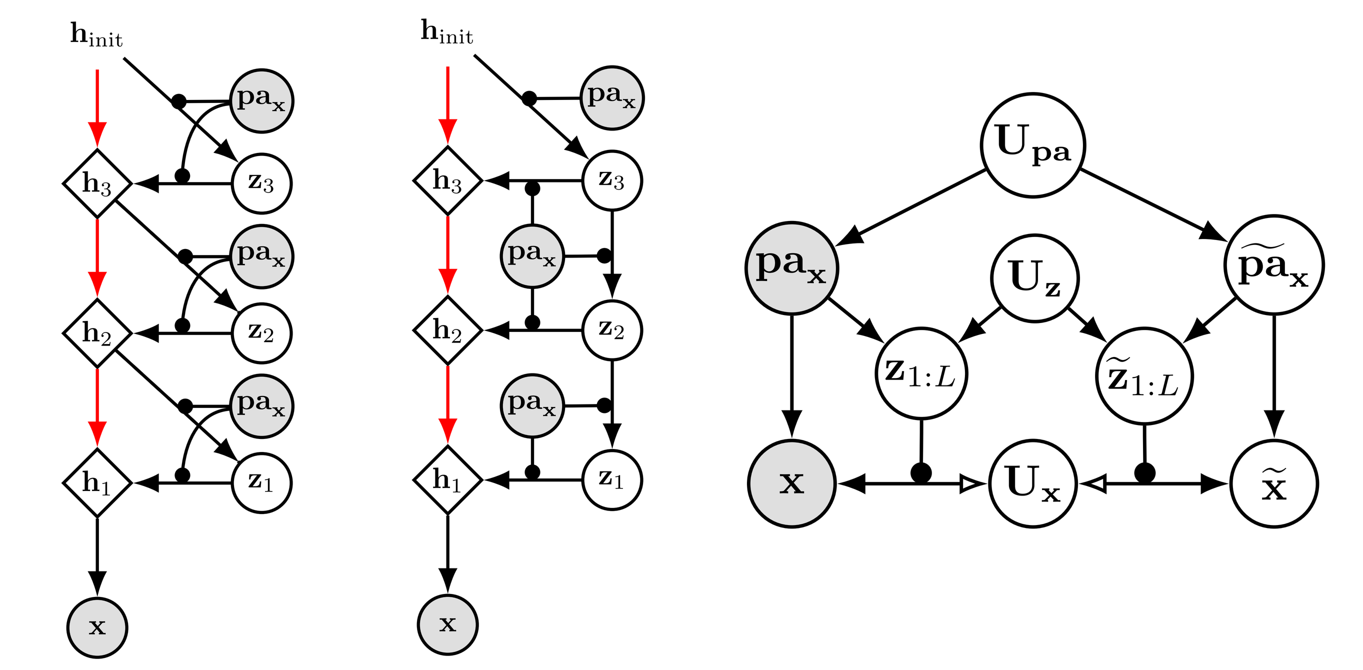
$$\epsilon = h^{-1}(\mathbf{x}; g_\theta(\mathbf{z}_{1:L}, \mathbf{pa_x})) = \frac{\mathbf{x} - \boldsymbol{\mu}(\mathbf{z}_{1:L}, \mathbf{pa_x})}{\boldsymbol{\sigma}(\mathbf{z}_{1:L}, \mathbf{pa_x})}$$

- **Action[†]:** $do(\mathbf{pa_x} \coloneqq \widetilde{\mathbf{pa}}_\mathbf{x})$
- **Prediction:**

$$\widetilde{\mathbf{x}} \sim p_\theta(\widetilde{\mathbf{x}} \mid \mathbf{z}_{1:L}, \widetilde{\mathbf{pa}}_\mathbf{x})$$

$$= h(\epsilon; g_\theta(\mathbf{z}_{1:L}, \widetilde{\mathbf{pa}}_\mathbf{x})) = \boldsymbol{\mu}(\mathbf{z}_{1:L}, \widetilde{\mathbf{pa}}_\mathbf{x}) + \boldsymbol{\sigma}(\mathbf{z}_{1:L}, \widetilde{\mathbf{pa}}_\mathbf{x}) \odot \epsilon$$

[†]Counterfactual parents $\widetilde{\mathbf{pa}}_\mathbf{x}$ are a result of upstream interventions on the associated causal graph.



(a) Generative #2 **(b)** No $q_\phi(\cdot)$ correction **(c) Latent Mediator:** $p_\theta(\mathbf{z}_{1:L} \mid \mathbf{pa_x})$

Figure: Twin network of our latent mediator SCM (c), induced by a conditional prior (a-b). Here, $\mathbf{z}_{1:L}$ becomes a **latent mediator** due to its dependence on endogenous variables $\mathbf{pa_x}$.

- **Abduction:**

$$\forall i : \mathbf{z}_i \sim q_\phi(\mathbf{z}_i \mid \mathbf{z}_{>i}, \mathbf{x}, \mathbf{pa_x}) = \boldsymbol{\mu}_i^q(\mathbf{z}_{>i}, \mathbf{pa_x}) + \boldsymbol{\sigma}_i^q(\mathbf{z}_{>i}, \mathbf{pa_x}) \odot \mathbf{U}_{\mathbf{z}_i}$$

$$\mathbf{U_x} = h^{-1}(\mathbf{x}; g_\theta(\mathbf{z}_{1:L}, \mathbf{pa_x})) = (\mathbf{x} - \boldsymbol{\mu}(\mathbf{z}_{1:L}, \mathbf{pa_x})) \oslash \boldsymbol{\sigma}(\mathbf{z}_{1:L}, \mathbf{pa_x})$$

- **Action:** $do(\mathbf{pa_x} \coloneqq \widetilde{\mathbf{pa}}_\mathbf{x})$
- **Prediction:**

$$\forall i : \widetilde{\mathbf{z}}_i \sim \pi \cdot p_\theta(\mathbf{z}_i \mid \mathbf{z}_{>i}, \widetilde{\mathbf{pa}}_\mathbf{x}) + (1-\pi) \cdot q_\phi(\mathbf{z}_i \mid \mathbf{z}_{>i}, \mathbf{x}, \mathbf{pa_x}) \qquad \text{(approx.)}$$

$$= \boldsymbol{\mu}_i^r(\widetilde{\mathbf{z}}_{>i}, \widetilde{\mathbf{pa}}_\mathbf{x}) + \boldsymbol{\sigma}_i^r(\widetilde{\mathbf{z}}_{>i}, \widetilde{\mathbf{pa}}_\mathbf{x}) \odot \mathbf{U}_{\mathbf{z}_i}$$

$$\widetilde{\mathbf{x}} = h(\mathbf{U_x}; g_\theta(\widetilde{\mathbf{z}}_{1:L}, \widetilde{\mathbf{pa}}_\mathbf{x})) = \boldsymbol{\mu}(\widetilde{\mathbf{z}}_{1:L}, \widetilde{\mathbf{pa}}_\mathbf{x}) + \boldsymbol{\sigma}(\widetilde{\mathbf{z}}_{1:L}, \widetilde{\mathbf{pa}}_\mathbf{x}) \odot \mathbf{U_x}$$

3. Causal Mediation Analysis

- The study of how a treatment effect is **mediated** by another variable, to help explain **why** or **how** an individual may respond to certain stimulus.

Estimating **Direct** (DE), **Indirect** (IE) and **Total** (TE) causal effects[†]:

$$\mathrm{DE}_\mathbf{x}(\widetilde{\mathbf{pa}}) \coloneqq \mathbb{E}\left[g_\theta(\widetilde{\mathbf{pa}}_\mathbf{x}, \mathbf{z}_{1:L}) - g_\theta(\mathbf{pa_x}, \mathbf{z}_{1:L})\right]$$

$$\mathrm{IE}_\mathbf{x}(\widetilde{\mathbf{z}}_{1:L}) \coloneqq \mathbb{E}\left[g_\theta(\mathbf{pa_x}, \widetilde{\mathbf{z}}_{1:L}) - g_\theta(\mathbf{pa_x}, \mathbf{z}_{1:L})\right]$$

$$\mathrm{TE}_\mathbf{x}(\widetilde{\mathbf{pa}}, \widetilde{\mathbf{z}}_{1:L}) \coloneqq \mathbb{E}\left[g_\theta(\widetilde{\mathbf{pa}}_\mathbf{x}, \widetilde{\mathbf{z}}_{1:L}) - g_\theta(\mathbf{pa_x}, \mathbf{z}_{1:L})\right]$$

[†]Identifiability assumptions: sequential ignorability (Imai et al., 2010); iVAE setup (Khemakhem et al., 2020).

4. Counterfactual Training

- A technique for improving **axiomatic effectiveness** of counterfactuals.

$$I(\widetilde{\mathbf{pa}}_k; \widetilde{\mathbf{x}}) \geq \mathbb{E}_{p(\widetilde{\mathbf{pa}}_k, \widetilde{\mathbf{x}})}\left[\log q_\psi(\widetilde{\mathbf{pa}}_k \mid \widetilde{\mathbf{x}})\right] + H(\widetilde{\mathbf{pa}}_k)$$

Our **constrained optimization** objective:

$$\arg\min_{\theta,\phi} \mathbb{E}_{p_{\mathrm{data}}(\mathbf{x}, \mathbf{pa_x})}\left[\mathcal{L}_{\mathrm{CT}}(\mathcal{M}; \mathbf{x}, \mathbf{pa_x})\right] \qquad \text{s.t. } \mathcal{F}_{\mathrm{FE}}(\theta, \phi; \mathbf{x}, \mathbf{pa_x}) \leq c,$$

$$\mathcal{L}_{\mathrm{CT}}(\mathcal{M}; \mathbf{x}, \mathbf{pa_x}) = -\sum_k \mathbb{E}_{\widetilde{\mathbf{pa}}_k \sim p(\mathbf{pa}_k), \widetilde{\mathbf{x}} \sim P_\mathcal{M}(\widetilde{\mathbf{x}} | do(\widetilde{\mathbf{pa}}_k), \mathbf{x})}\left[\log q_{\psi_k}(\widetilde{\mathbf{pa}}_k \mid \widetilde{\mathbf{x}})\right]$$

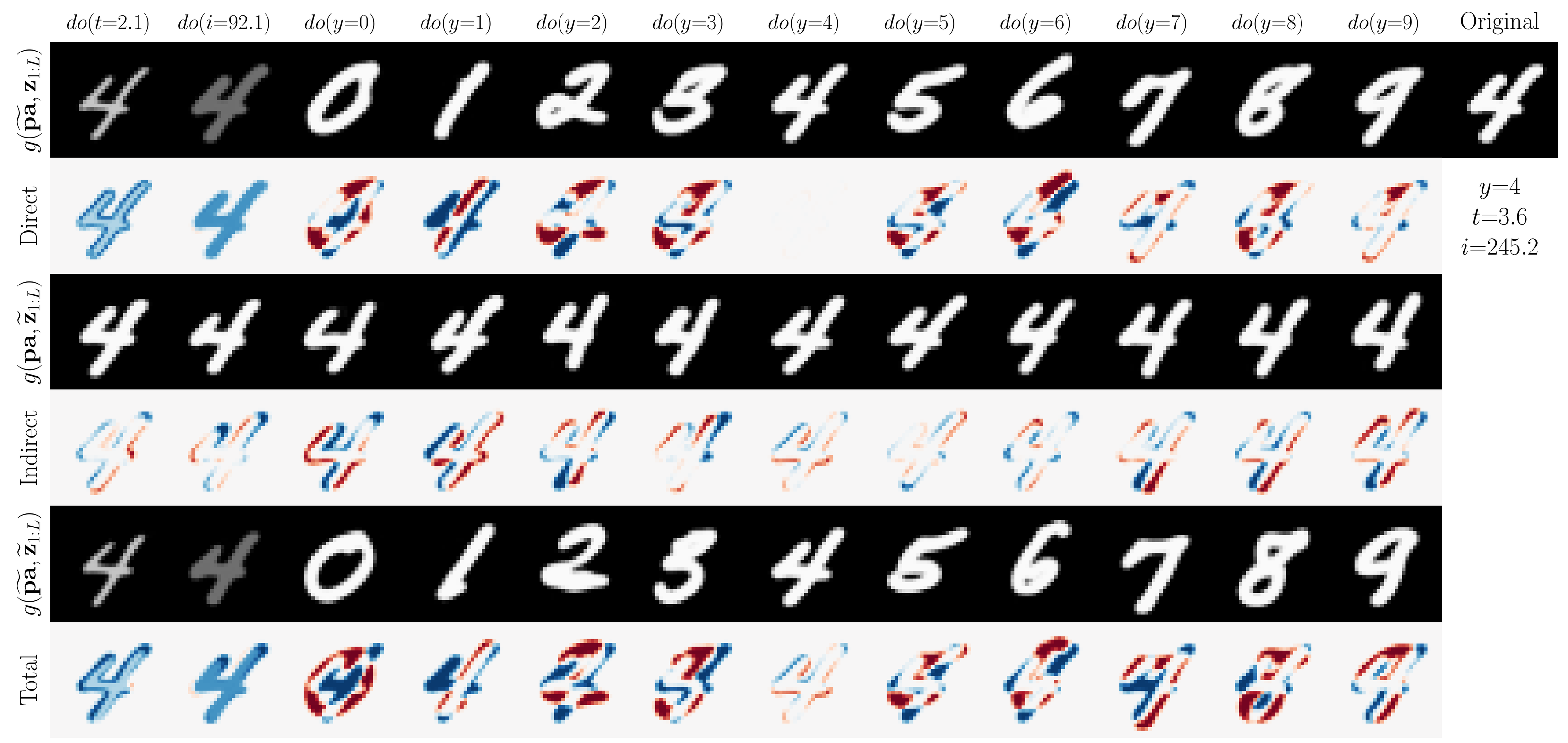
Paper, Demo and Code:

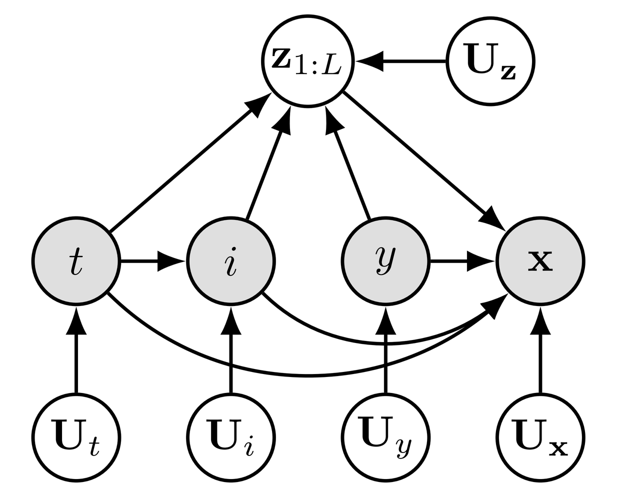
github.com/biomedia-mira/causal-gen

References:

{Pawlowski*, Castro*} et al. Deep Structural Causal Models for Tractable Counterfactual Inference. NeurIPS 2020.
{Monteiro*, De Sousa Ribeiro*} et al. Measuring Axiomatic Soundness of Counterfactual Image Models. ICLR 2023.
Khemakhem et al. Variational Autoencoders and Nonlinear ICA: A Unifying Framework. AISTATS 2020.
Imai et al. Identification, Inference and Sensitivity Analysis for Causal Mediation Effects. *Statistical Science*, 2010.

5. Morpho-MNIST

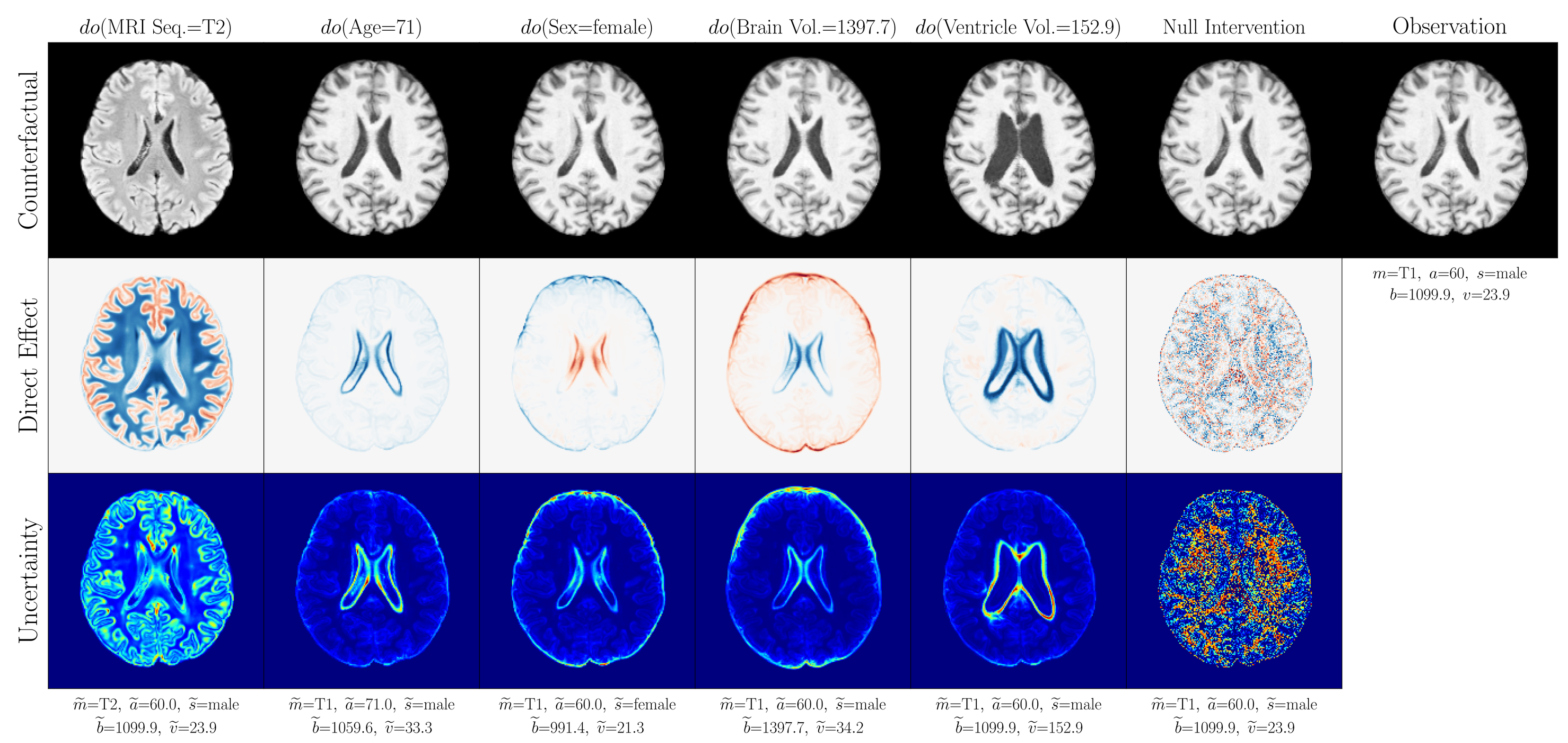


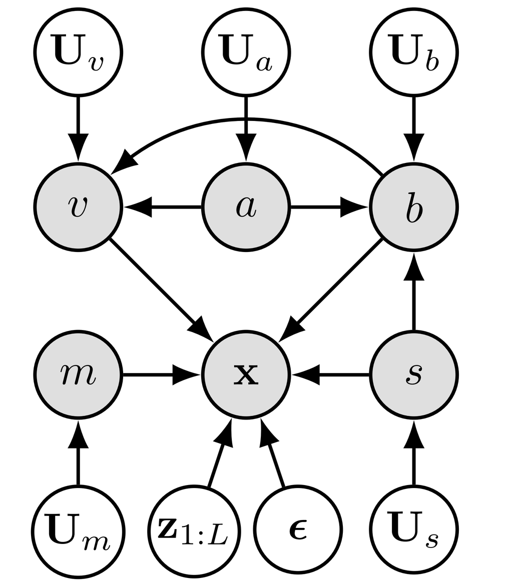


(a) Latent mediator SCM used for the Morpho-MNIST dataset. The observed variables are: image ($\mathbf{x}$), digit (y), stroke thickness (t) and pixel intensity (i).

Figure: Morpho-MNIST **counterfactuals** from our latent mediator SCM. **Direct**, **indirect** and **total** causal effects are shown (red: increase, blue: decrease). **Cross-world counterfactuals** (row 3) are the potential outcome of $\mathbf{x}$ given $\mathbf{pa_x}$ and the counterfactual mediator $\widetilde{\mathbf{z}}_{1:L}$ we could've observed had $\mathbf{pa_x} \coloneqq \widetilde{\mathbf{pa}}_\mathbf{x}$.

6. Brain Imaging (UK Biobank)

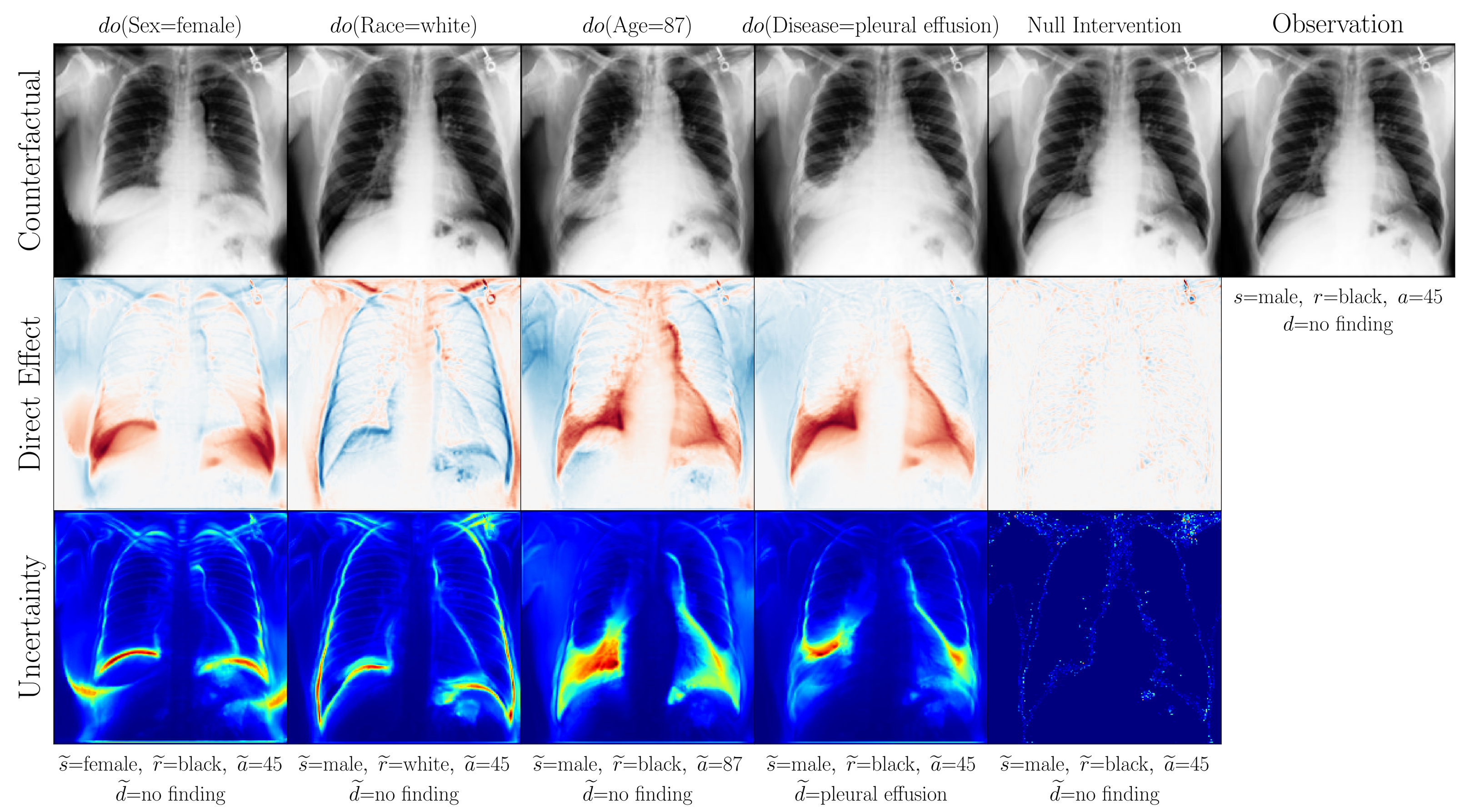


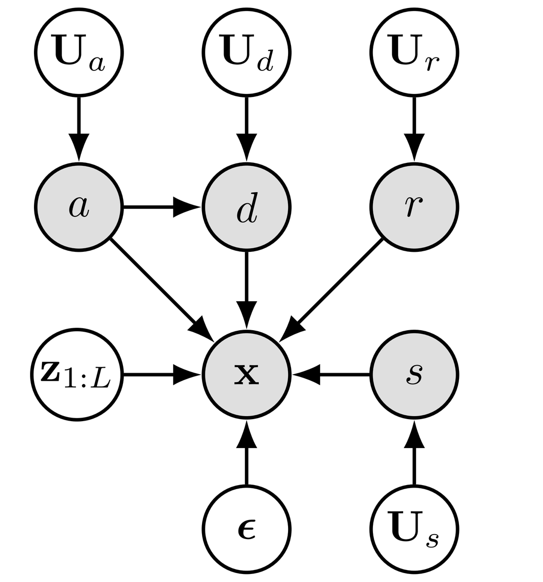


(a) Deep SCM used for UK Biobank. Variables: MRI Seq. (m), age (a), sex (s), brain (b) and ventricle (v) volume.

Figure: Brain MRI **counterfactuals** from our SCM (exogenous prior). Subject identity is preserved. **Counterfactual uncertainty** from stochastic abduction.

7. Chest X-ray (MIMIC)





(a) Deep SCM used for MIMIC-CXR. The observed variables in the causal graph are: age (a), sex (s), race (r), disease (d) and chest x-ray ($\mathbf{x}$). The disease studied is pleural effusion.

Figure: Chest X-ray **counterfactuals** from our SCM. **Localized interventional changes** which respect the causal graph (a) and preserve subject identity.