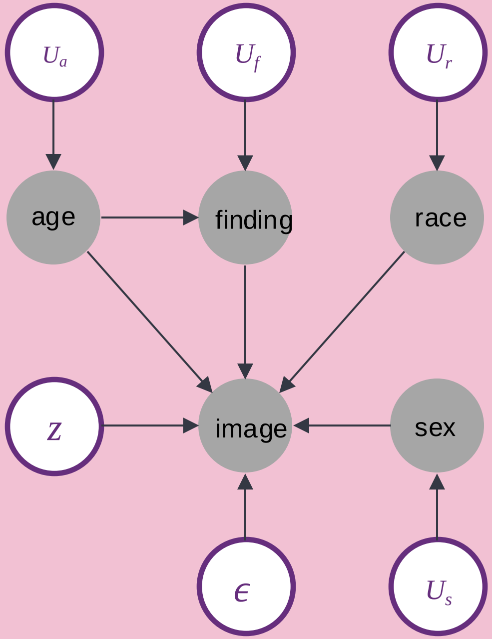
Comparing Causally-Informed and Uninformed Models in Scenarios with Biases and Mediating Effects

08/10/2024

McGill

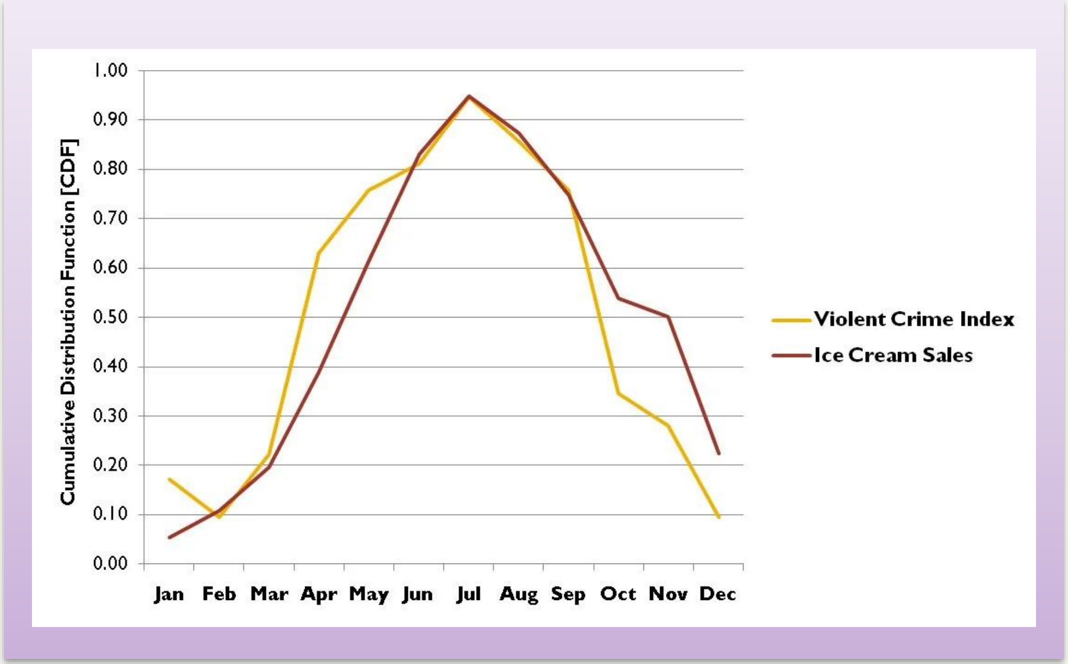
Mila

Table of Contents

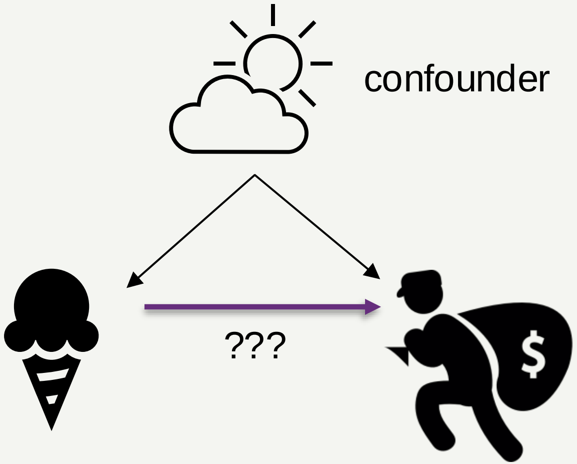


Why do we need causality?

Motivating Example



Should we intervene and shutdown ice cream sales to reduce crime?



First Instinct: Condition!

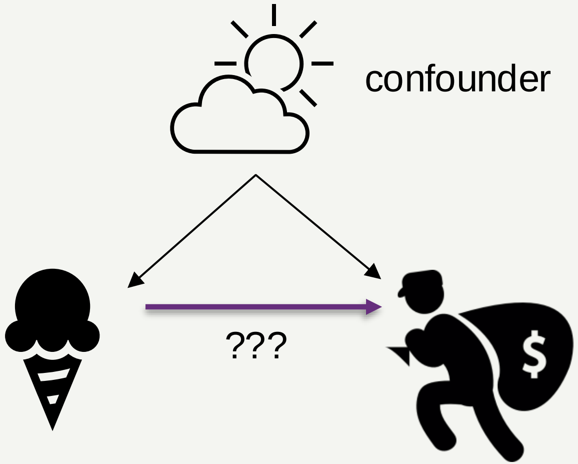
What happens if we condition on ***low*** *ice cream sales?*

Crime Rate \ Ice Cream Sales	low	high
low	5	1
high	2	4

World

World: All year

Should we intervene and shutdown ice cream sales to reduce crime?



First Instinct: Condition!

*What happens if we condition on **low** ice cream sales?*

Ice Cream Sales

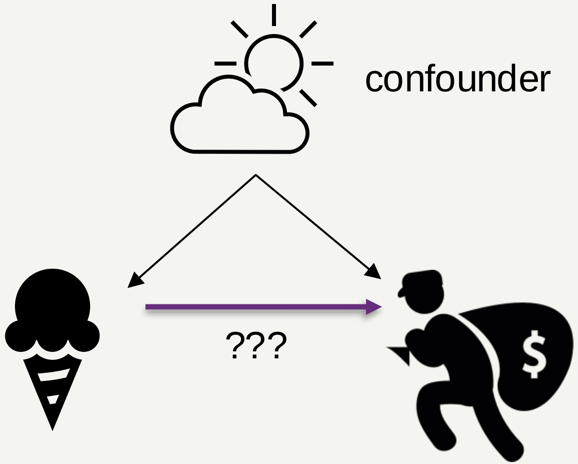
Crime Rate		low	high
	low	5	1
	high	2	4

World

$P(low\ crime \mid low\ ice\ cream\ sales) = 71\%$ ☺

World: Months where ice cream sales were low

Should we intervene and shutdown ice cream sales to reduce crime?



First Instinct: Condition!

What happens if we condition on ***high*** *ice cream sales?*

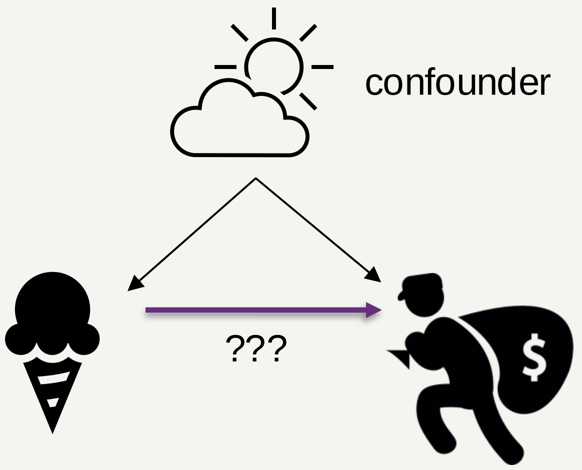
Crime Rate \ Ice Cream Sales	*low*	*high*
low	5	1
high	2	4

World

$P(low\ crime \mid high\ ice\ cream\ sales) = 20\%$ ☹

World: Months where ice cream sales were high

Should we intervene and shutdown ice cream sales to reduce crime?



Condition DOES NOT EQUAL Intervention

Problem: Conditioning constricts our perception to **certain months**, cannot extrapolate to **all months**

Ice Cream Sales

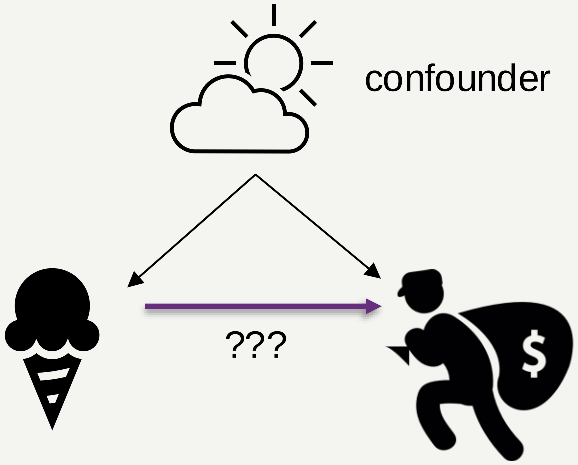
Crime Rate		*low*	*high*
	low	5	1
	high	2	4

World

$P(low\ crime \mid high\ ice\ cream\ sales) = 20\%$ ☹

World: Months where ice cream sales were high

Should we intervene and shutdown ice cream sales to reduce crime?



Condition DOES NOT EQUAL Intervention

This analysis gives us the crime rate given that ice cream sales were low in **certain months**:

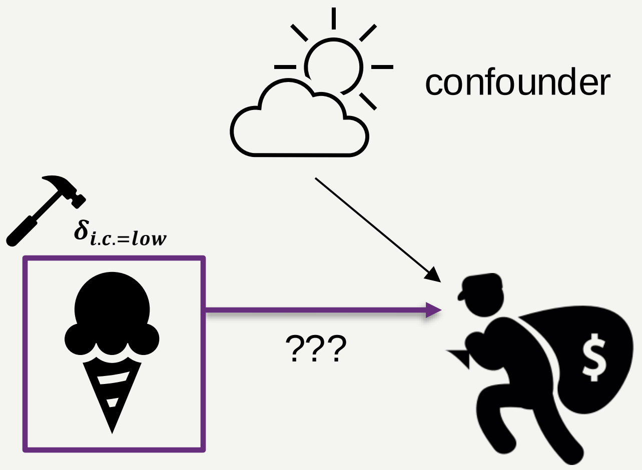
$$P(crime|ice\ cream = low)$$

We want to know the crime rate in a **world** where we intervene and set ice cream sales to be low **in all months**:

$$P(crime|do(ice\ cream = low))$$

Q: How can we do this?
A: Find the interventional distribution via manipulation of a structural causal model (SCM)!

Should we intervene and shutdown ice cream sales to reduce crime?



Structural Causal Model

A structural causal model (SCM) is the triple: $M := \langle V, U, F \rangle$

Where there are two sets of variables:

Endogenous:
the attributes within our model
e.g. ice cream sales

$V = \{v_1, \ldots, v_N\}$

Exogeneous:
the attributes external to our model

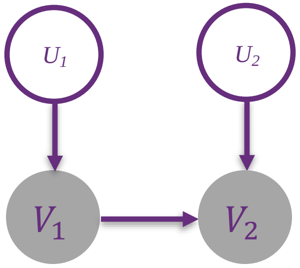
$U = \{u_1, \ldots, u_N\}$

& a set of **functions**:

$F = \{f_1, \ldots, f_N\}$

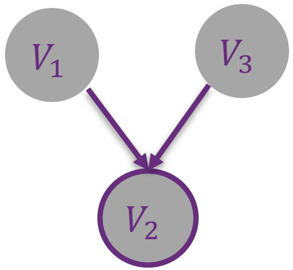
Each **endogenous** variable is a **function** of it's direct causes and respective exogeneous noise variable:

$$v_k = f(\text{pa}_k; u_k).$$



$$V_1 := f_1(U_1), \quad U_1 \sim N(0,1)$$
$$V_2 := f_2(V_1, U_2), \quad U_2 \sim N(0,1)$$

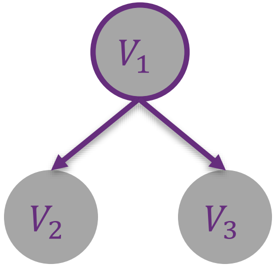
Structural Causal Model



Collider

$V_1 \perp V_3$

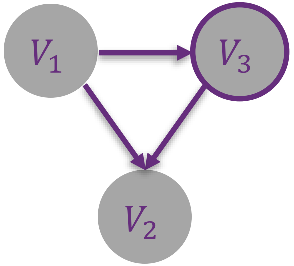
$V_1 \not\perp V_3 \,|V_2$



Confounder

$V_2 \not\perp V_3$

$V_2 \perp V_3 \,|V_1$



Mediator

$V_2 \not\perp V_3$

$V_1 \not\perp V_2$

$V_1 \not\perp V_3$

Structural Causal Model

A structural causal model (SCM) is the triple: $M := \langle V, U, F \rangle$

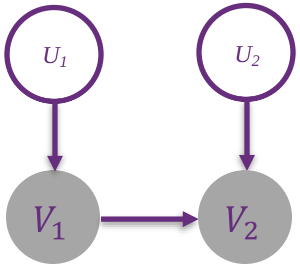
Assumptions!

$$P(U) = \prod_{k=1}^{N} P(u_k)$$

[no unobserved confounders]

$$P(V) = \prod_{k=1}^{N} p(v_k | pa_k)$$

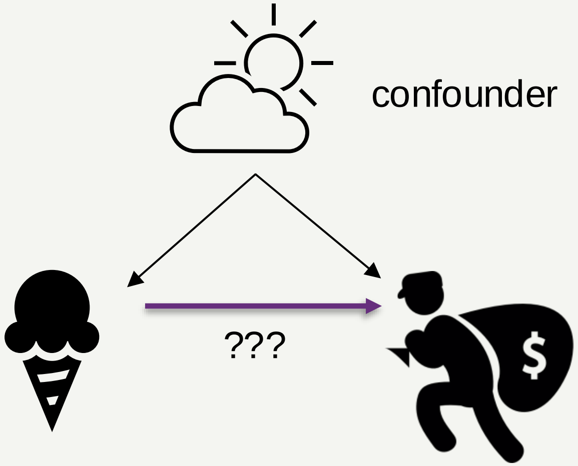
[Each var is independent of its non-desc. given its parents ➔ Markovian]



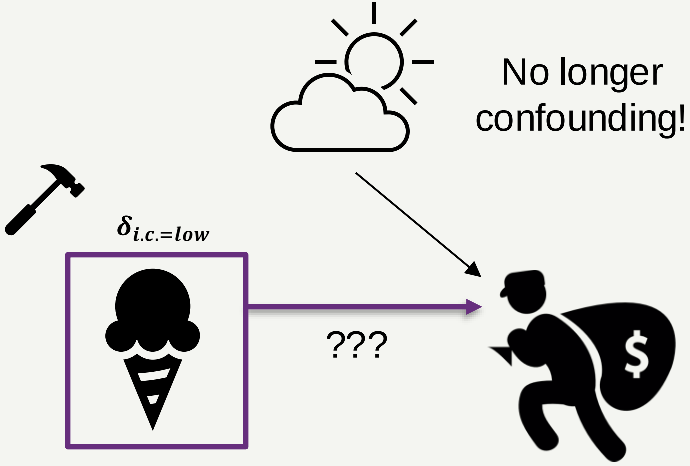
$$V_1 := f_1(U_1), \quad U_1 \sim N(0,1)$$
$$V_2 := f_2(V_1, U_2), \quad U_2 \sim N(0,1)$$

Motivating Example

Joint distribution:



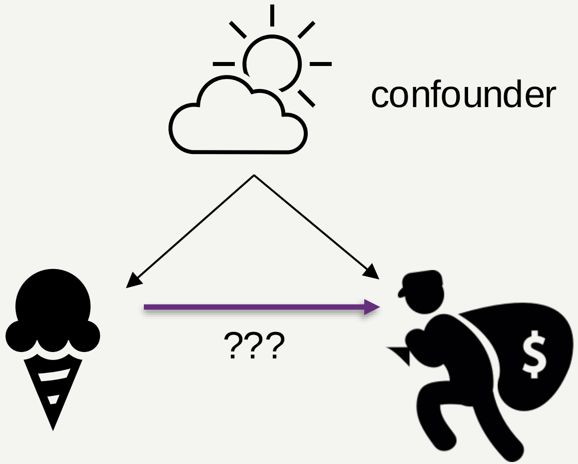
Interventional distribution



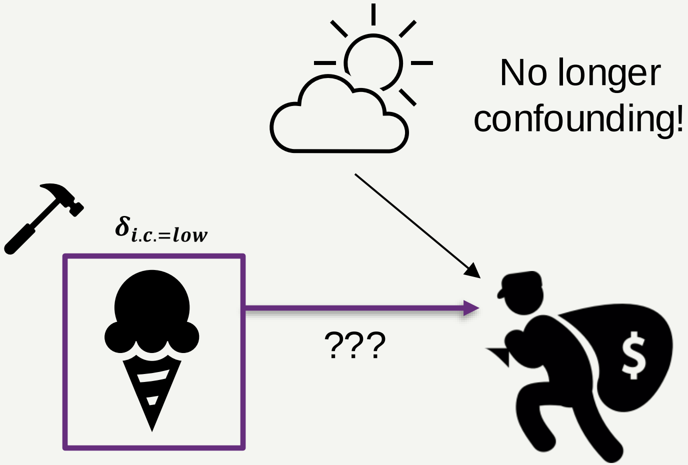
If we had an interventional distribution, we could directly sample from it to get our answer…

Motivating Example

Joint distribution:



Interventional distribution:



$$P(crime|do(ice\ cream = low)) = P_{int.}(crime|\ ice\ cream = low)\ [\,Def.]$$

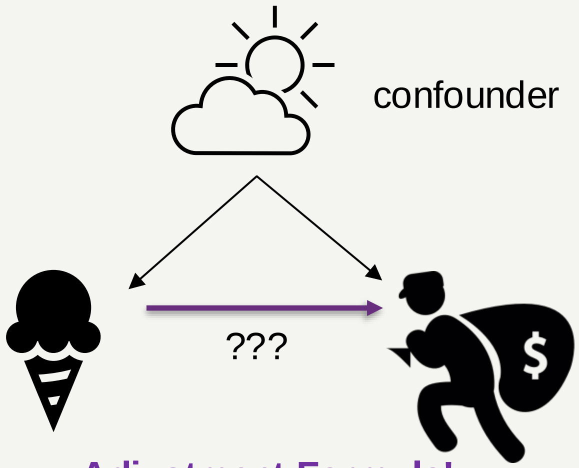
$$= \sum_{z} P_{int.}(crime\ |\ ice\ cream = low, Month\ = z)\ * P_{int}(Month = z\,|)\ [Marg.\ out\ z]$$

$$= \sum_{z} P\,(crime|ice\ cream = low, Month = z) * P(Month = z)$$

Adjustment Formula!

Motivating Example

Joint distribution:



Adjustment Formula!

$$P(crime|do(ice\ cream = low)) = \sum_z P\,(crime\ |ice\ cream = low, Month = z) * P(Month = z)$$

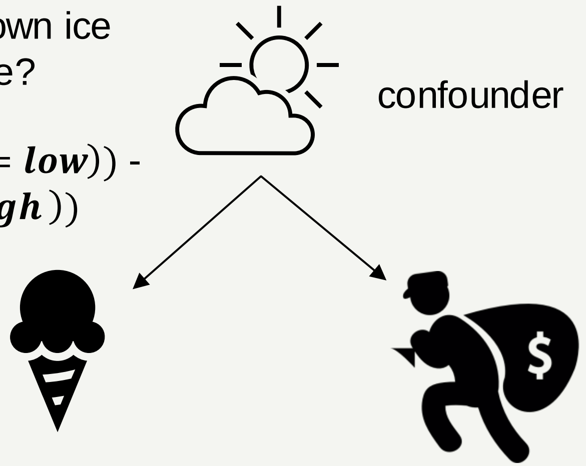
We can use pre-intervention probabilities to correctly adjust for time of year!

Condition DOES NOT EQUAL Intervention

Joint distribution:

Should we intervene and shutdown ice cream sales to reduce crime?

$$ATE = P(crime|do(ice\ cream = low)) - P(crime|do(ice\ cream = high)) = \text{negligable (hopefully)}$$



No!
National crisis averted.

Take-home message:
In some cases, conditioning alone is not sufficient. Need knowledge of the causal graph to obtain correct distribution!

What else is possible with SCMs?

Can we reason about individuals? Can we reason in the past?

Intervention

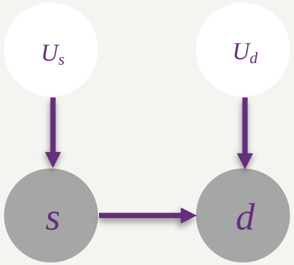
$P(y|do(x))$

*If we force **everyone** to stop smoking, what will be the prevalence of cancer?*

Association

$P(y|x)$

***Of the patients that smoke**, what is the prevalence of cancer?*



Take-home message:
In some cases, conditioning alone is not sufficient. Need knowledge of the causal graph to obtain correct distribution!

Judea Pearls “rungs of causality”

Association

$P(y|x)$

Of the patients that smoke, what is the prevalence of cancer?

Intervention

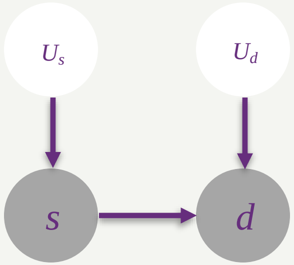
$P(y|do(x))$

If we force everyone to stop smoking, what will be the prevalence of cancer?

Counterfactual

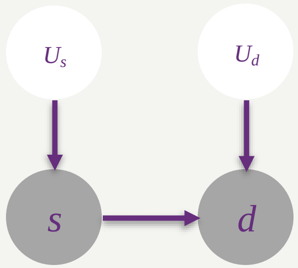
$P(y_x|y',x')$

If patient A had not smoked in the last two years, would they have cancer?



Take-home message:
In some cases, conditioning alone is not sufficient. Need knowledge of the causal graph to obtain correct distribution!

Counterfactual Inference



Observed World
$d = f(s, U_d)$

1. Abduction →

$U_s = f^{-1}(s)$
$U_d = f^{-1}(d, s)$

2. Action →

$\tilde{s} := 0$

3. Prediction →

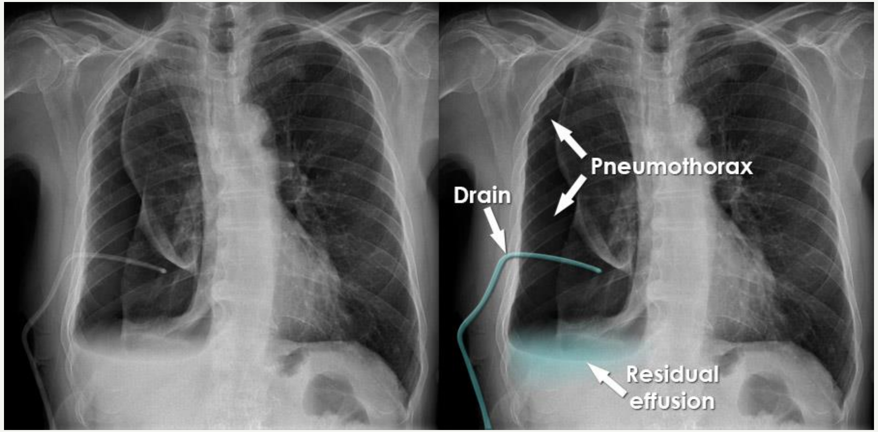
Counterfactual World
$\tilde{d} = f(\tilde{s}, U_d)$

Take-home message:
In some cases, conditioning alone is not sufficient. Need knowledge of the causal graph to obtain correct distribution!

Why should we care in medical imaging?

Take-home message:
In some cases, conditioning alone is not sufficient. Need knowledge of the causal graph to obtain correct distribution!

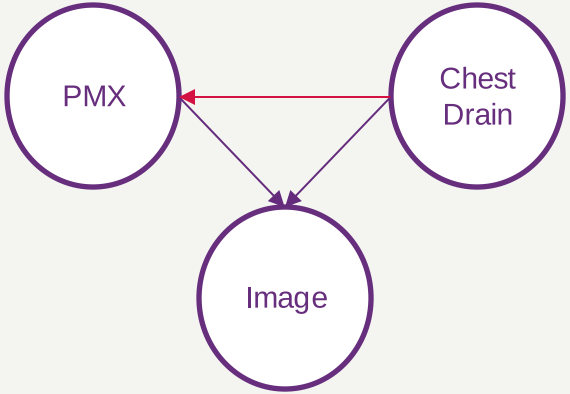
Classification



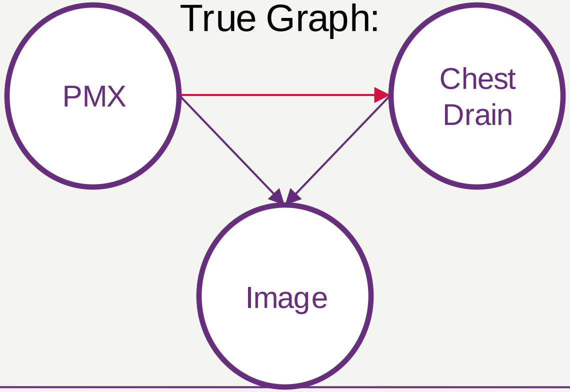
Classification:

a model that achieves "expert-level" performance in classifying pneumothorax (PMX) from chest X-rays was found to depend on the presence of chest tubes

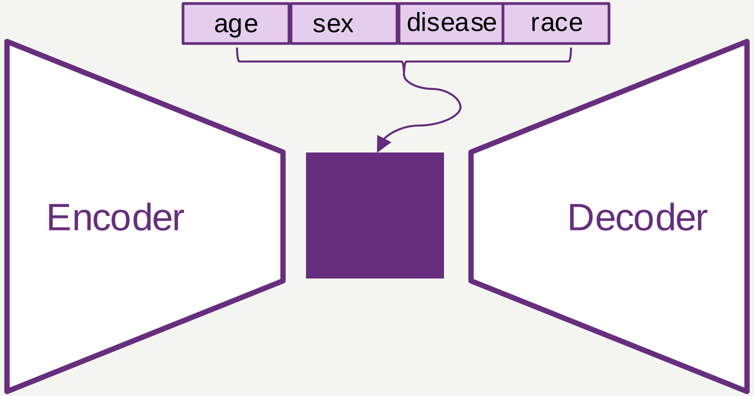
What the model learned:

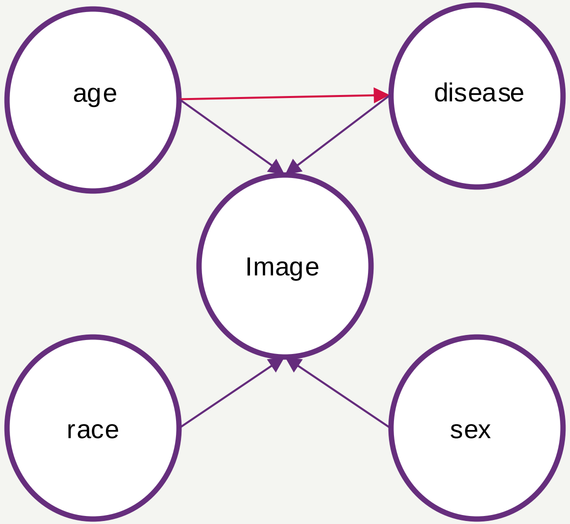


True Graph:



Conditional Generation





Potential Failure Modes

Sampling Bias/Spurious Correlations:

- Biases in the dataset lead to biased samples. *e.g. model generates disease for all older patients*

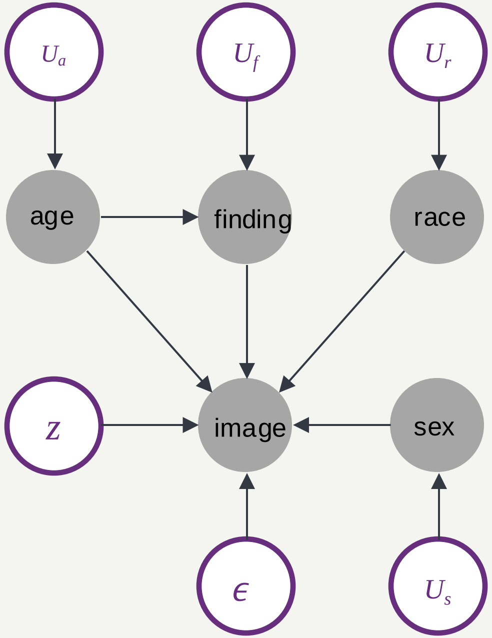
Incorrect Interventions:

- The model cannot learn how changes in one attribute affect others. *e.g. changing age doesn't show its impact on disease progression*

Mediating Effects:

- The model can't determine the relative impact of each attribute. *e.g. separating the effects of age from disease on the generated image.*

Generating Counterfactuals



What would this patient look like had they not had the disease?
What will they look like in 10 years?

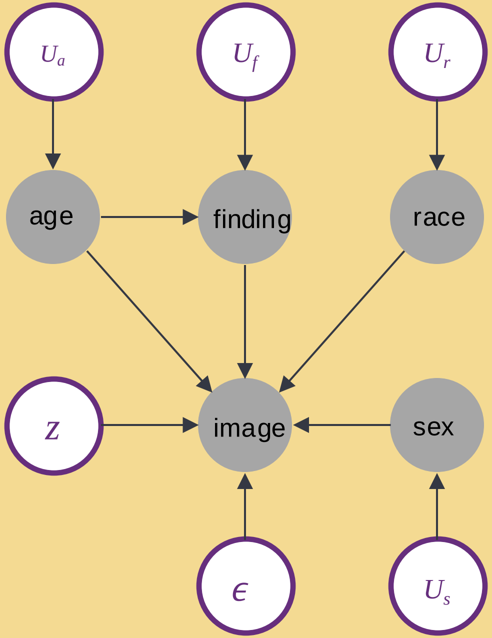
Research Questions:

1. Are SCMs necessary for correct generation compared to when using state-of-the-art (SOTA) conditional models?
2. In what scenarios (if any) do SCMs outperform SOTA conditional methods?
3. How do SCMs impact the interpretability of generated counterfactuals compared to SOTA conditional methods?

Hypothesis

SCMs will improve data generation in the presence of significant biases and mediating effects by better capturing causal relationships, resulting in more realistic and plausible samples compared to models without causal understanding.

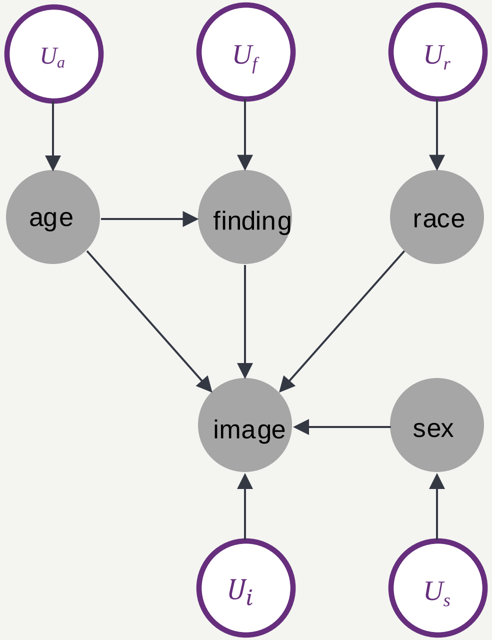
What we want:



How do attributes such as age, disease, race, sex, etc. impact each other and, in turn, the MRI that we see?

High Fidelity Image Counterfactuals with Probabilistic Causal Models

Fabio De Sousa Ribeiro [1] Tian Xia [1] Miguel Monteiro [1] Nick Pawlowski [2] Ben Glocker [1]



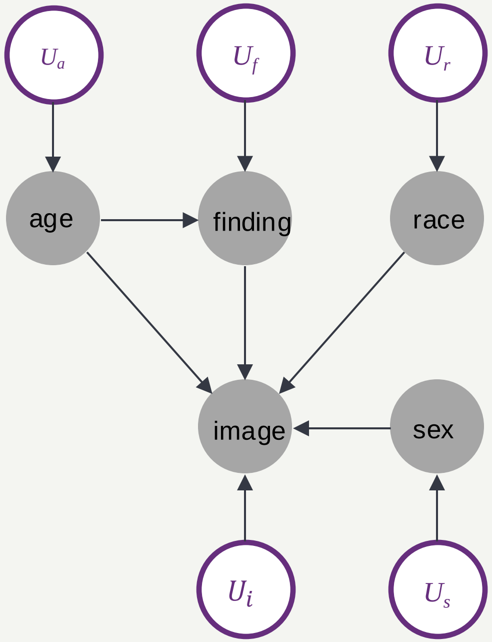
Their research question: can we generate plausible high-fidelity counterfactuals using deep mechanisms?

- Challenge: image is high-dimensional!
- Becomes a trade-off:
  - (a) learn flexible, *invertible, complex causal mechanisms
  - (b) computationally **tractable

*So we can abduct the exogeneous variables!
**So we can generate high-dimensional data!

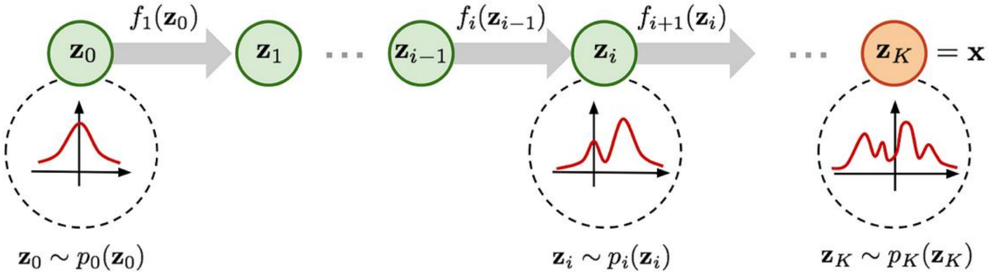
High Fidelity Image Counterfactuals with Probabilistic Causal Models

Fabio De Sousa Ribeiro [1] Tian Xia [1] Miguel Monteiro [1] Nick Pawlowski [2] Ben Glocker [1]



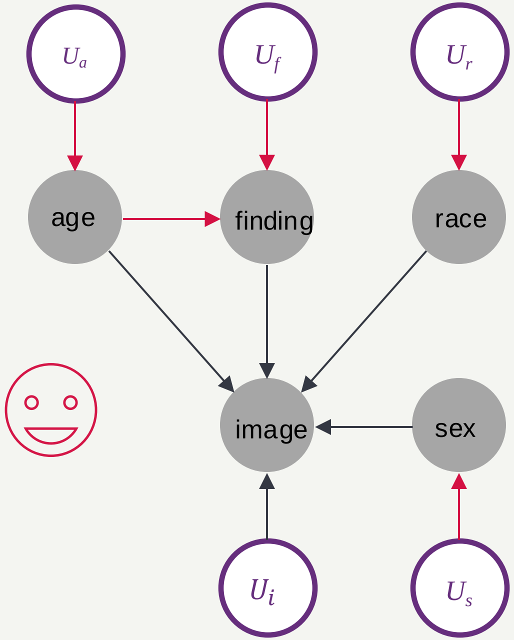
- Challenge: image is high-dimensional!
- Becomes a trade-off:
  - (a) learn flexible, invertible, complex causal mechanisms
  - (b) computationally tractable

For (a): Normalizing Flows! Use successive invertible transformations from a simple dist. to learn complex dist.



High Fidelity Image Counterfactuals with Probabilistic Causal Models

Fabio De Sousa Ribeiro [1] Tian Xia [1] Miguel Monteiro [1] Nick Pawlowski [2] Ben Glocker [1]



- Challenge: image is high-dimensional!
- Becomes a trade-off:
  - (a) learn flexible, invertible, complex causal mechanisms
  - (b) computationally tractable

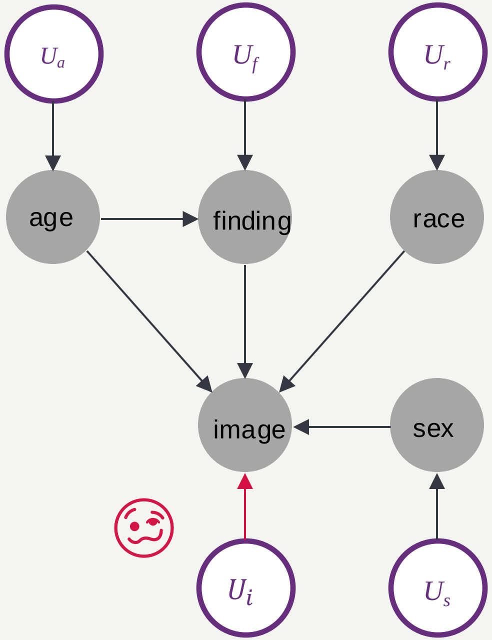
For (a): Normalizing Flows! Use successive invertible transformations from a simple dist. to learn complex dist.

Great for between non-image attributes… we can have invertible mechanisms and thus perform deterministic abduction

$$v_k = f(\mathrm{pa}_k; u_k). \qquad u_k = f^{-1}(pa_k, v_k)$$

High Fidelity Image Counterfactuals with Probabilistic Causal Models

Fabio De Sousa Ribeiro [1] Tian Xia [1] Miguel Monteiro [1] Nick Pawlowski [2] Ben Glocker [1]



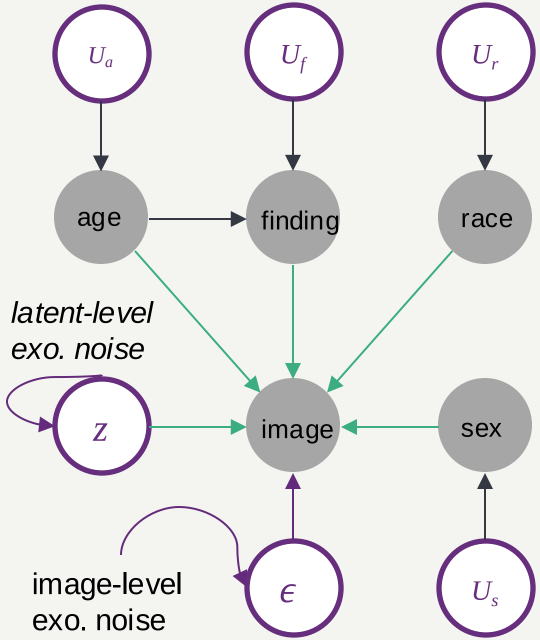
- Challenge: image is high-dimensional!
- Becomes a trade-off:
  - (a) learn flexible, invertible, complex causal mechanisms
  - (b) computationally tractable

For (a): Normalizing Flows! Use successive invertible transformations from a simple dist. to learn complex dist.

…Not so great when transformation needs to be at image-level

High Fidelity Image Counterfactuals with Probabilistic Causal Models

Fabio De Sousa Ribeiro [1] Tian Xia [1] Miguel Monteiro [1] Nick Pawlowski [2] Ben Glocker [1]



We want $x = f(pa_x, u_x)$ but f cannot be *completely* invertible due to tractability…

Can we at least make it *partially* invertible (at the cost of deterministic abduction) ?

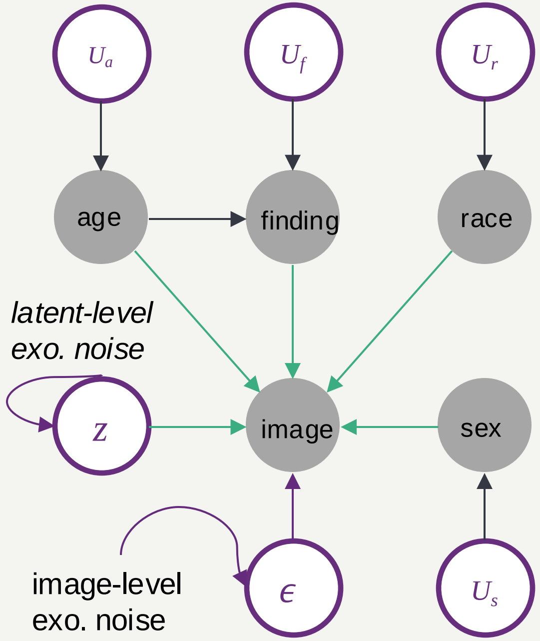
1. Encode the image into a smaller dim. latent space, z

   This can be part of the image's exogeneous noise, i.e. causes of the image that are not encapsulated by our other attributes

2. Anything not encapsulated in the image space, or the attributes, can be considered image-level noise!

High Fidelity Image Counterfactuals with Probabilistic Causal Models

Fabio De Sousa Ribeiro [1] Tian Xia [1] Miguel Monteiro [1] Nick Pawlowski [2] Ben Glocker [1]



1. Encode the image into a smaller dim. latent space, z

   This can be part of the image's exogeneous noise, i.e. causes of the image that are not encapsulated by our other attributes

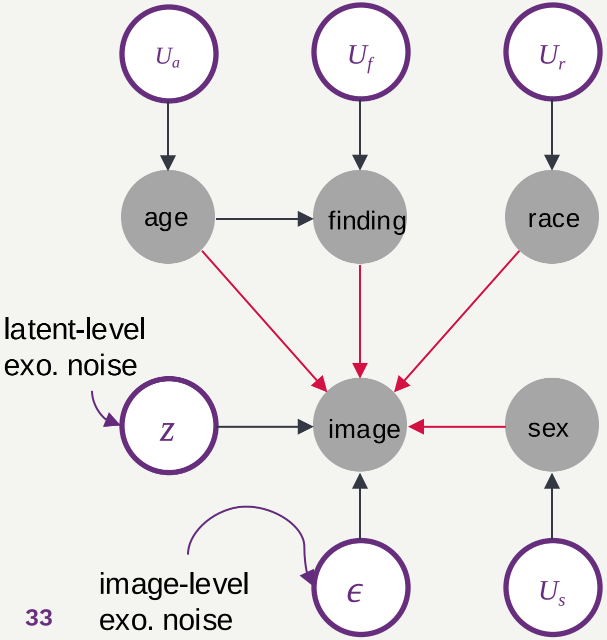
2. Decode that information along with the parents into image-space: $g_\theta(z, pa_x)$
3. Anything not encapsulated in z or the attributes, can be considered image-level noise!

$$x = f(pa_x, u_x) = h(\epsilon, g_\theta(z, pa_x))$$
$$= \mu(z, pa_x) + \sigma(z, pa_x) \circ \epsilon, \epsilon \sim N(0, I)$$

$*$non-invertible

High Fidelity Image Counterfactuals with Probabilistic Causal Models

Fabio De Sousa Ribeiro [1] Tian Xia [1] Miguel Monteiro [1] Nick Pawlowski [2] Ben Glocker [1]



Thus noise factorizes as: $p(u_x) = p_\theta(z)p(\epsilon)$.

Giving us the following steps for CF inference!

(1) Abduction: $\mathbf{z} \sim q_\phi(\mathbf{z} \mid \mathbf{x}, \mathbf{pa_x})$ $\quad \boldsymbol{\epsilon} = h^{-1}(\mathbf{x}; g_\theta(\mathbf{z}, \mathbf{pa_x})) = \dfrac{\mathbf{x} - \boldsymbol{\mu}(\mathbf{z}, \mathbf{pa_x})}{\boldsymbol{\sigma}(\mathbf{z}, \mathbf{pa_x})}$

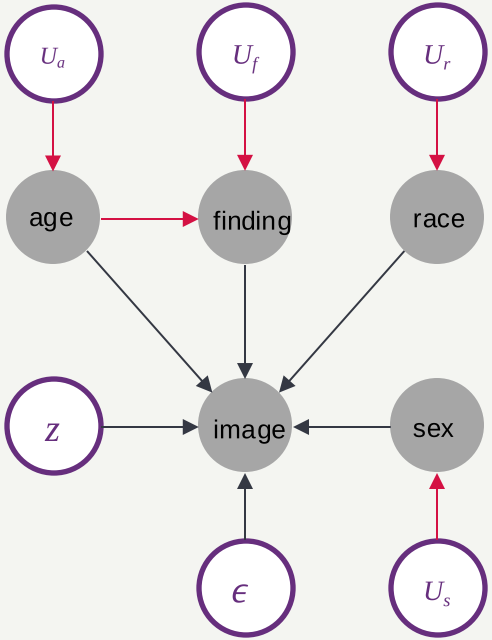
(2) Action: $do(pa_x \coloneqq \widetilde{pa}_x)$

(3) Predict: $\tilde{x} \sim p_\theta(\tilde{x} | z, \widetilde{pa}_x)$

$$= h\big(\epsilon, g_\theta(z, \widetilde{pa}_x)\big) = \mu(z, \widetilde{pa}_x) + \sigma(z, \widetilde{pa}_x) \circ \epsilon$$

High Fidelity Image Counterfactuals with Probabilistic Causal Models

Fabio De Sousa Ribeiro [1] Tian Xia [1] Miguel Monteiro [1] Nick Pawlowski [2] Ben Glocker [1]



Key takeaways:

1. To model mechanisms between parents of the image:
   Normalizing Flows
2. To model the image's mechanism:
   1. *Encode into latent space to get z*
   2. *Send through decoder $g_\theta(z, pa_x)$*
   3. *Add in image-level exogeneous noise ϵ*

Pipeline

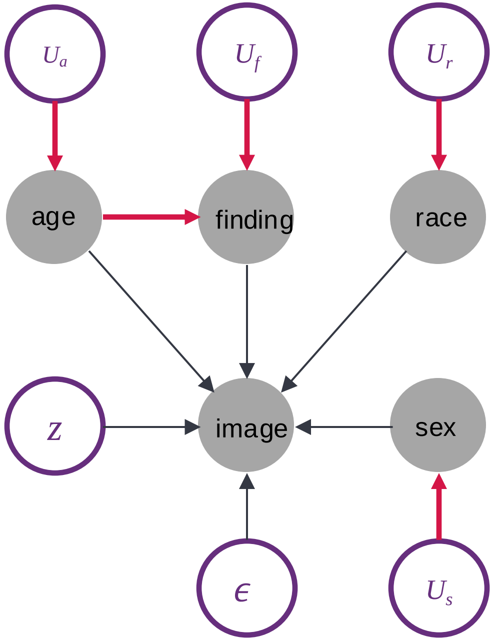
1. Train & validate SCM between attributes
2. Develop conditional SOTA models
3. Generate (pseudo) counterfactuals
4. Define scenarios & compare performances

Pipeline

Train & validate SCM between attributes → Develop conditional SOTA models → Generate (pseudo) counterfactuals → Define scenarios & compare performances

Pipeline: Phase 1

Train & validate SCM between attributes

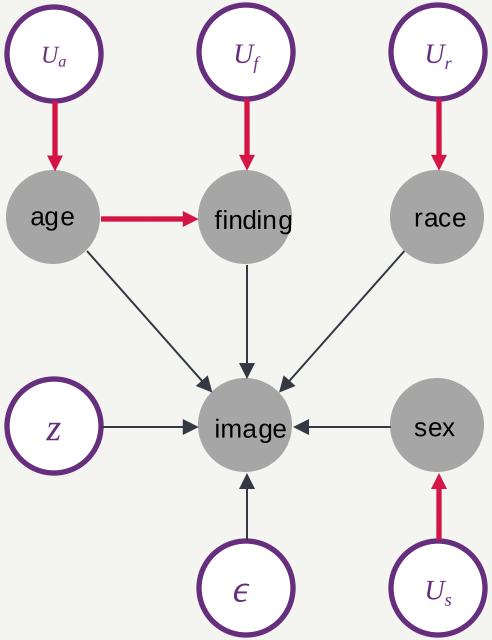


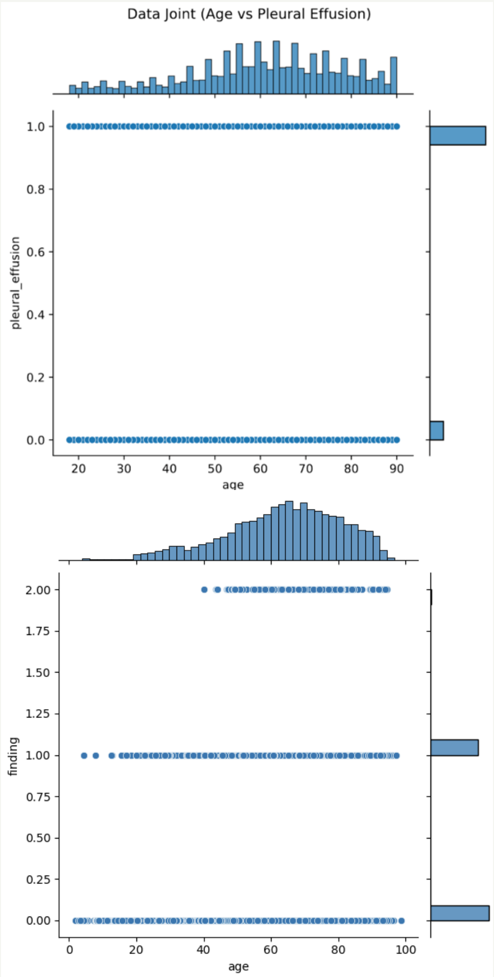
- Sample exogeneous noise from parameterized distributions
- Use normalizing flows* to build complex probability distributions
- Training via min. KL-divergence between observed and generated joint dist.

Enables **invertible mechanisms ➔ **deterministic abduction** for attributes*

Pipeline: Phase 1

Train & **validate** SCM between attributes





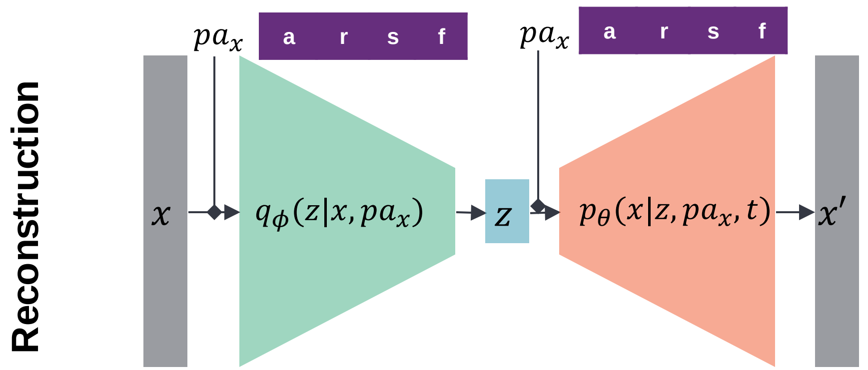
McGill

Mila

Pipeline

1. Train & validate SCM between attributes
2. Develop conditional SOTA models
3. Generate (pseudo) counterfactuals
4. Define scenarios & compare performances

Develop conditional models

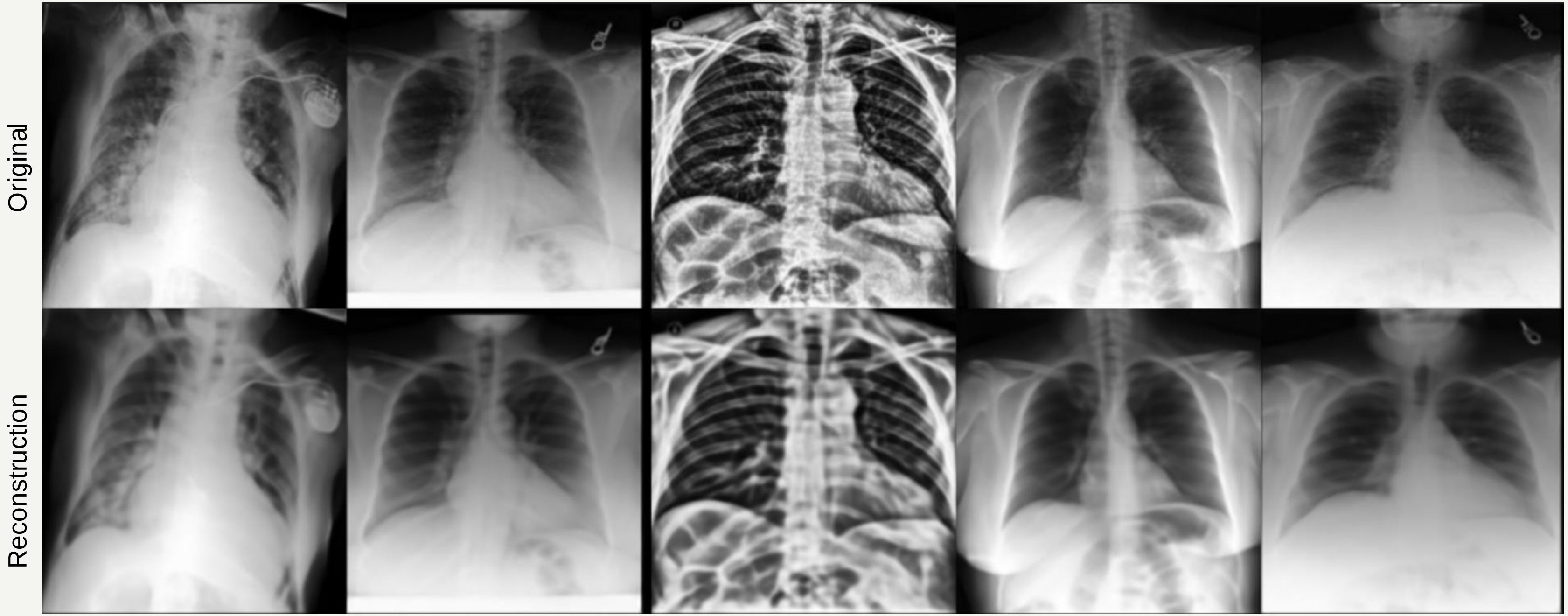


$$\text{ELBO} = \mathbb{E}_{q_\phi(z_{1:L}|x)}\left[\log p_\theta(x|z_{1:L})\right] - \text{KL}(q_\phi(z_{1:L}|x) \parallel p_\theta(z_{1:L}))$$

How likely is the data to be observed?
Can we reconstruct?

How regularized is the latent space?
Can we sample?

Can we reconstruct images?

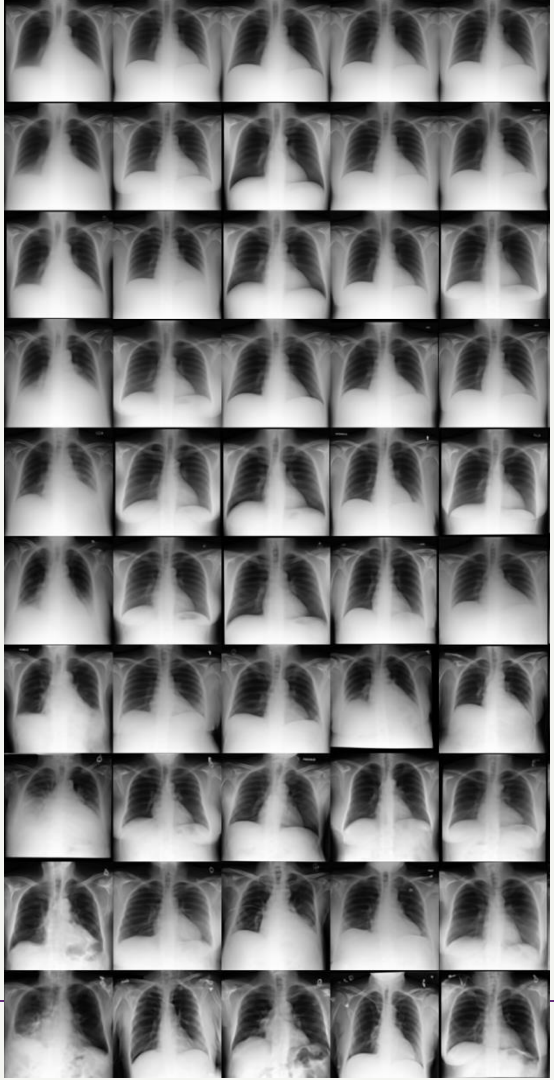


McGill

Mila

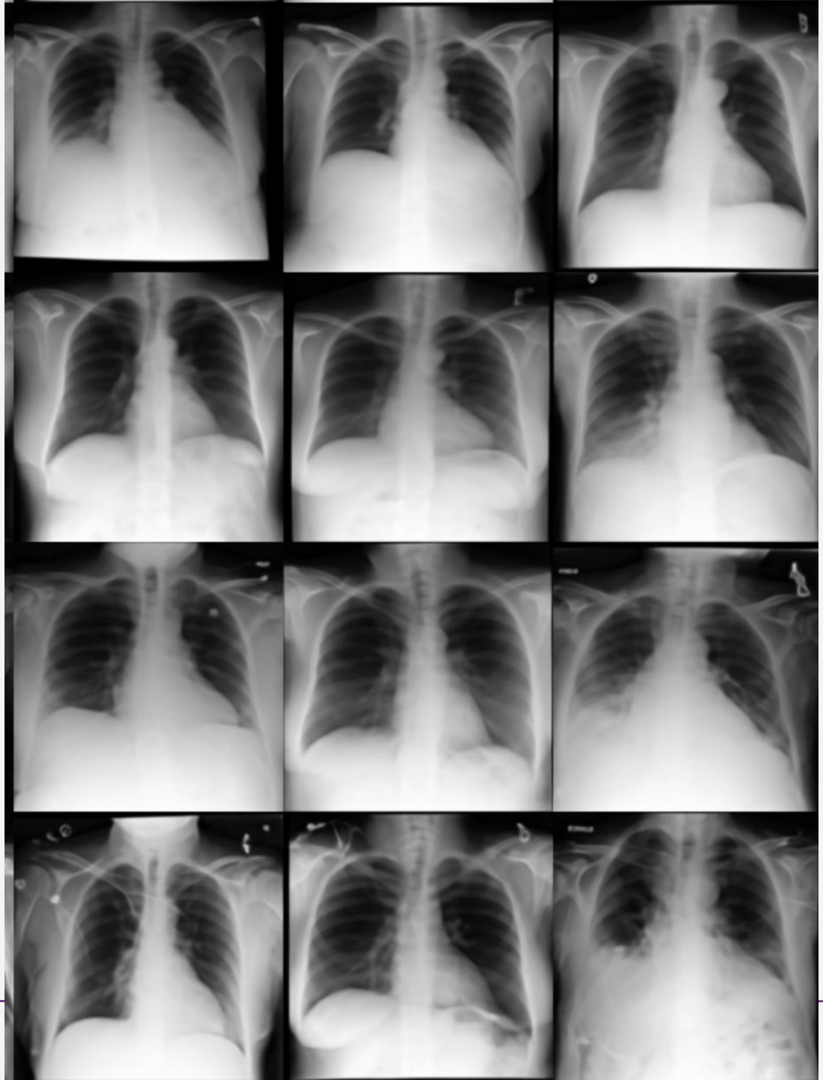
Can we sample?

(increasing temperature)



Can we sample?

(increasing temperature)
↓



McGill

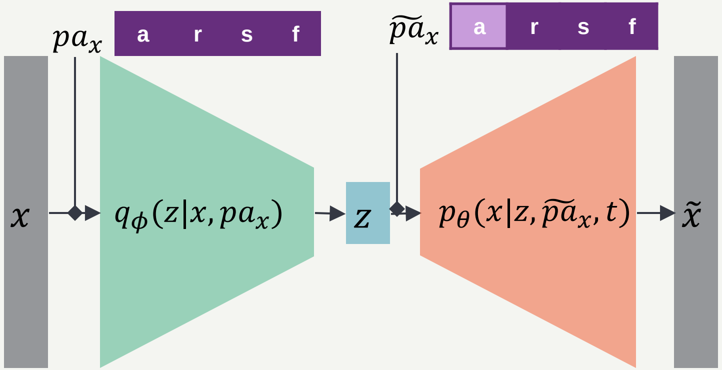
Mila

Pipeline

1. Train & validate SCM between attributes
2. Develop conditional SOTA models
3. **Generate (pseudo) counterfactuals**
4. Define scenarios & compare performances

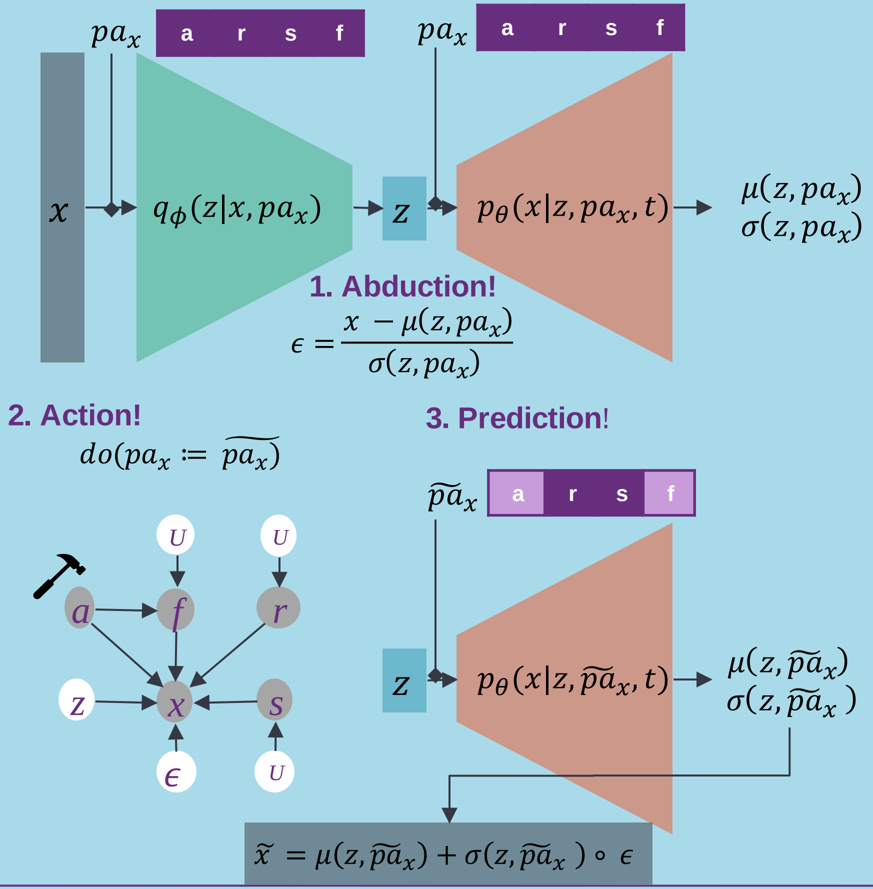
How can we compare architectures with and without SCM?

no-SCM



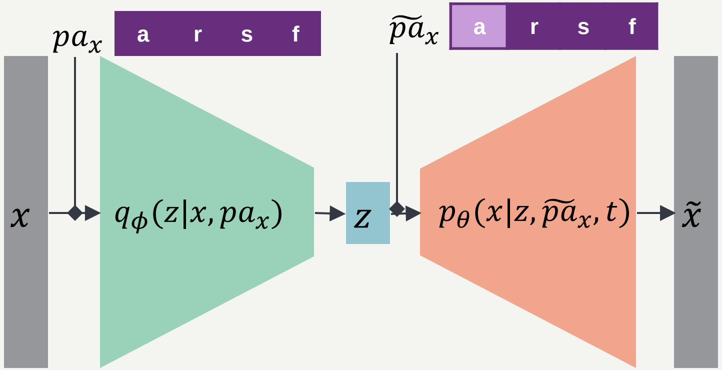
pa_x
a r s f
$\widetilde{pa}_x$
a r s f
x
$q_\phi(z|x, pa_x)$
z
$p_\theta(x|z, \widetilde{pa}_x, t)$
$\tilde{x}$

SCM



pa_x
a r s f
pa_x
a r s f
x
$q_\phi(z|x, pa_x)$
z
$p_\theta(x|z, pa_x, t)$
$\mu(z, pa_x)$
$\sigma(z, pa_x)$
1. Abduction!
$\epsilon = \frac{x - \mu(z, pa_x)}{\sigma(z, pa_x)}$
2. Action!
$do(pa_x \coloneqq \widetilde{pa_x})$
3. Prediction!
$\widetilde{pa}_x$
a r s f
U
U
a
f
r
z
x
s
ϵ
U
z
$p_\theta(x|z, \widetilde{pa}_x, t)$
$\mu(z, \widetilde{pa}_x)$
$\sigma(z, \widetilde{pa}_x)$
$\tilde{x} = \mu(z, \widetilde{pa}_x) + \sigma(z, \widetilde{pa}_x) \circ \epsilon$

no-SCM



How to train?

The authors use typical ELBO loss here.

**note that this does not allow for rigorous analysis compared with next slide*

$$\text{ELBO} = \mathbb{E}_{q_\phi(z_{1:L}|x)}\left[\log p_\theta(x|z_{1:L})\right] - \text{KL}(q_\phi(z_{1:L}|x) \parallel p_\theta(z_{1:L}))$$

What's the loss?

Problem with conditional models, they can learn:

$$p_\theta(x|c) = p_\theta(x).$$

in other words, they can learn to ignore the condition

This is a problem for downstream CF training!

$$p_\theta(\tilde{x}|\tilde{c}) = p_\theta(x).$$

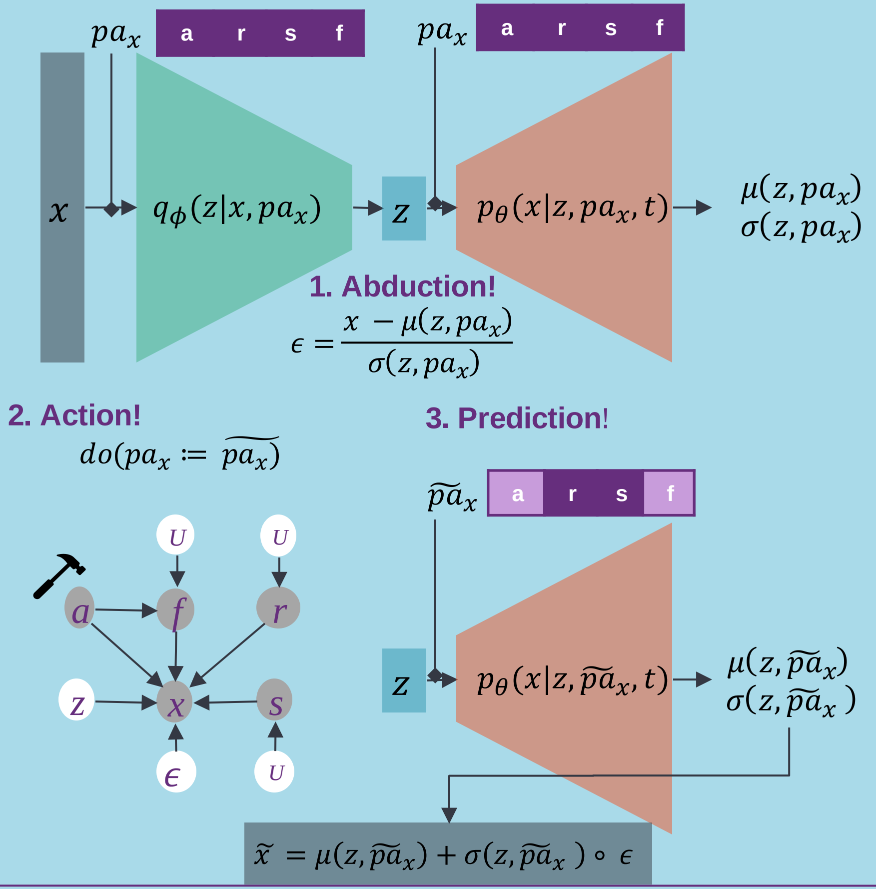
Solution used by Riberio et al: We expect that there exists mutual information (MI) between $\tilde{x}$ and $\widetilde{pa}$

$$\begin{aligned} I(\tilde{pa}_k; x) &= \mathbb{E}_{p(\tilde{pa}_k, \tilde{x})}\left[\log \frac{p(\tilde{pa}_k|\tilde{x})}{\tilde{pa}_k} \cdot \frac{q_\psi(\tilde{pa}_k|\tilde{x})}{q_\psi(\tilde{pa}_k|\tilde{x})}\right] \\ &= \mathbb{E}_{p(\tilde{pa}_k, \tilde{x})}\left[\log \frac{q_\psi(\tilde{pa}_k|\tilde{x})}{p(\tilde{pa}_k)}\right] + \mathbb{E}_{p(x)} D_{KL}(p(\tilde{pa}_k|\tilde{x}) \| q_\psi(\tilde{pa}_k|x)) \\ &\geq \mathbb{E}_{p(\tilde{pa}_k, x)}\left[\log q_\psi(\tilde{pa}_k|\tilde{x})\right] + H(\tilde{pa}_k) \end{aligned}$$

$$L_{\mathrm{CT}}(M; x, pa_x) = -\sum_{k=1}^{K} \mathbb{E}_{\substack{\tilde{pa}_k \sim p(pa_k) \\ \tilde{x} \sim P_M(x|do(\tilde{pa}_k), x)}} \log q_{\psi_k}(\tilde{pa}_k|x).$$

$$\max_{P_M, q_\psi} \mathbb{E}_{p_{data}}(x, pa_x)\left[-L_{\mathrm{CT}}(M; x, pa_x)\right]$$

SCM



How to train?

1. Train and fix parameters for SCM ω and parent predictor ψ

2. Pretrain ϕ, θ of the HVAE, but allow gradients from CF training

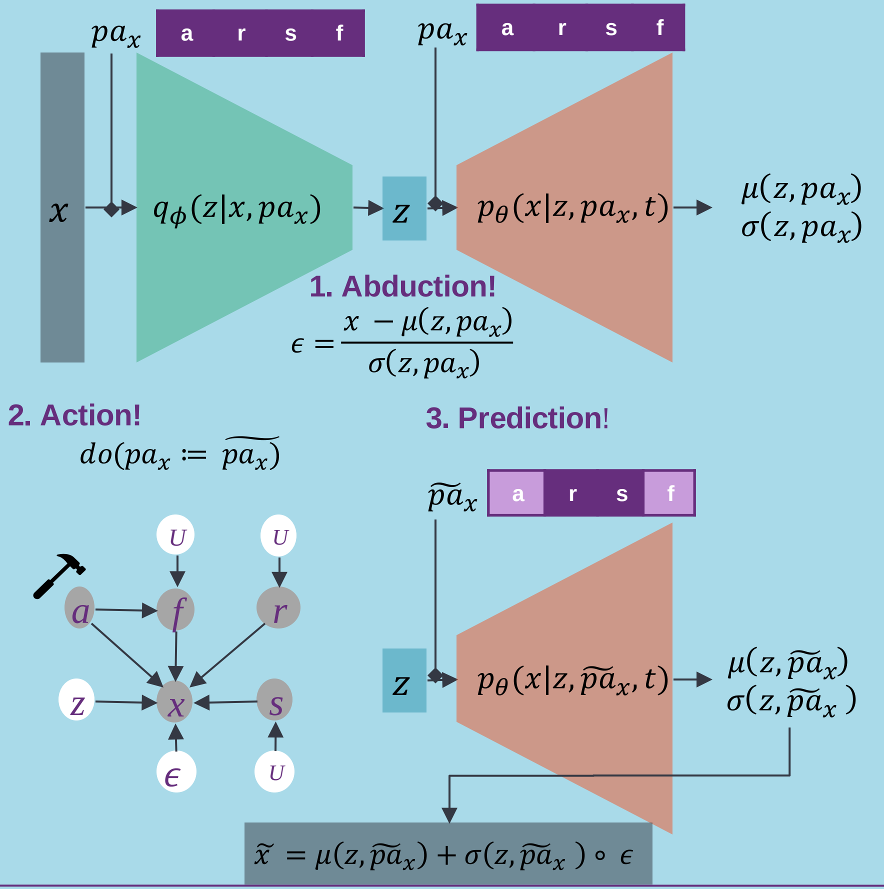
3. Train CF $(\omega, \psi, \phi, \theta)$ with *Lagrangian Optimization* to avoid degrading quality on observed data

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

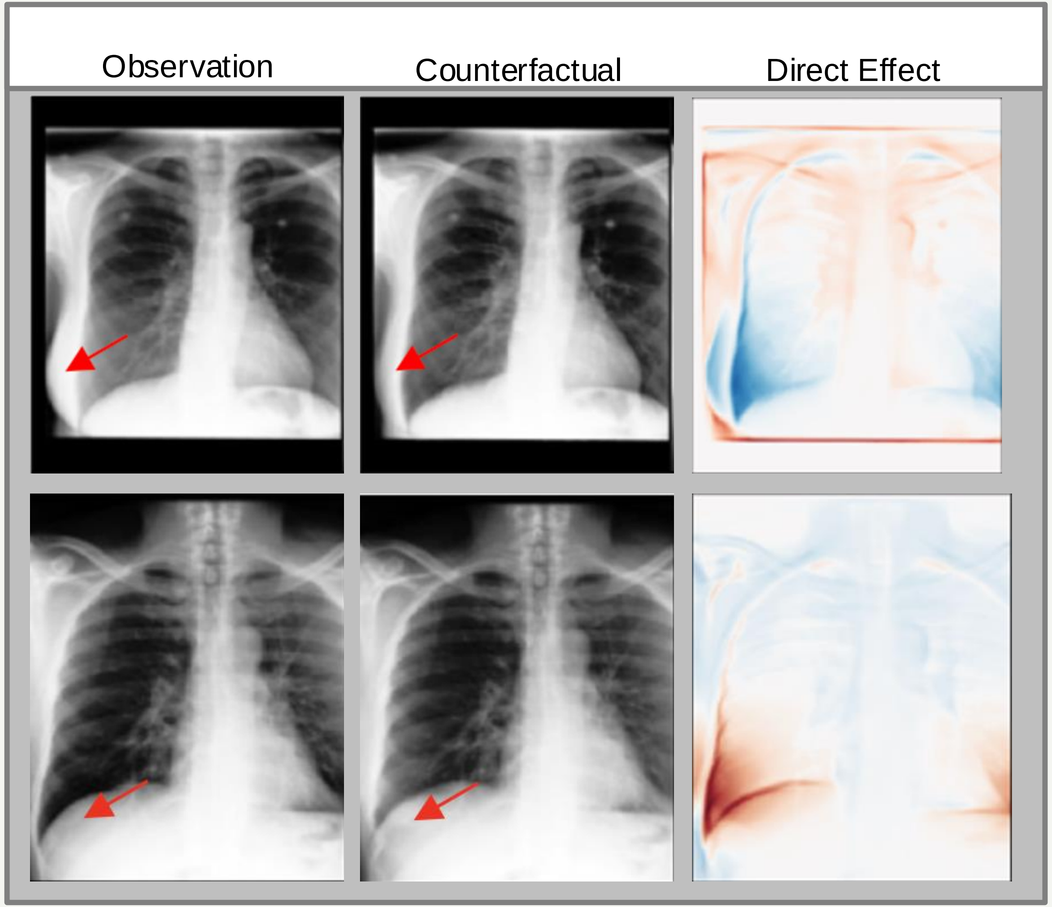
Make sure the external parent predictor can correctly classify the parents…

…But don't let the ELBO term increase on the observed data

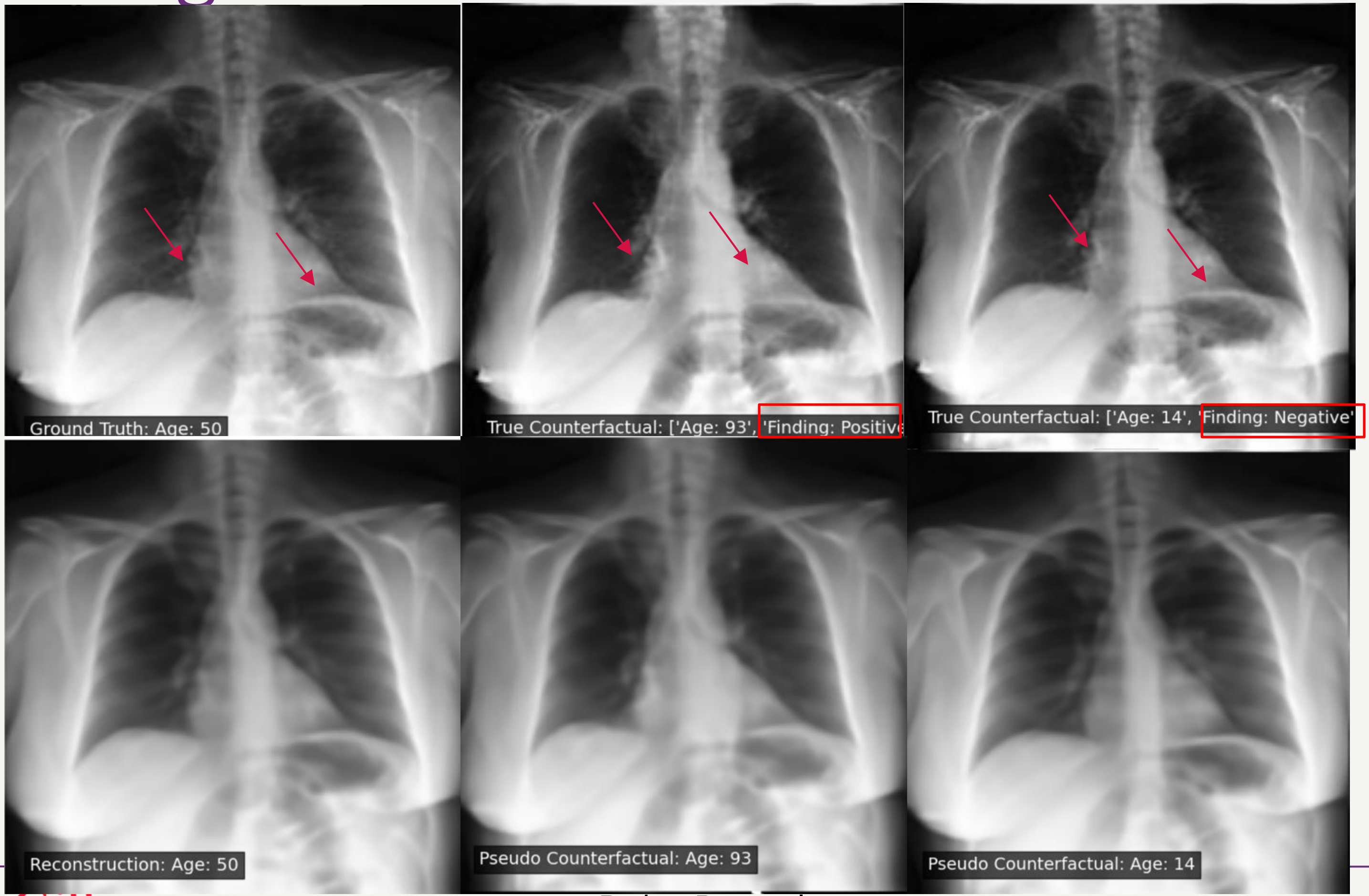
SCM



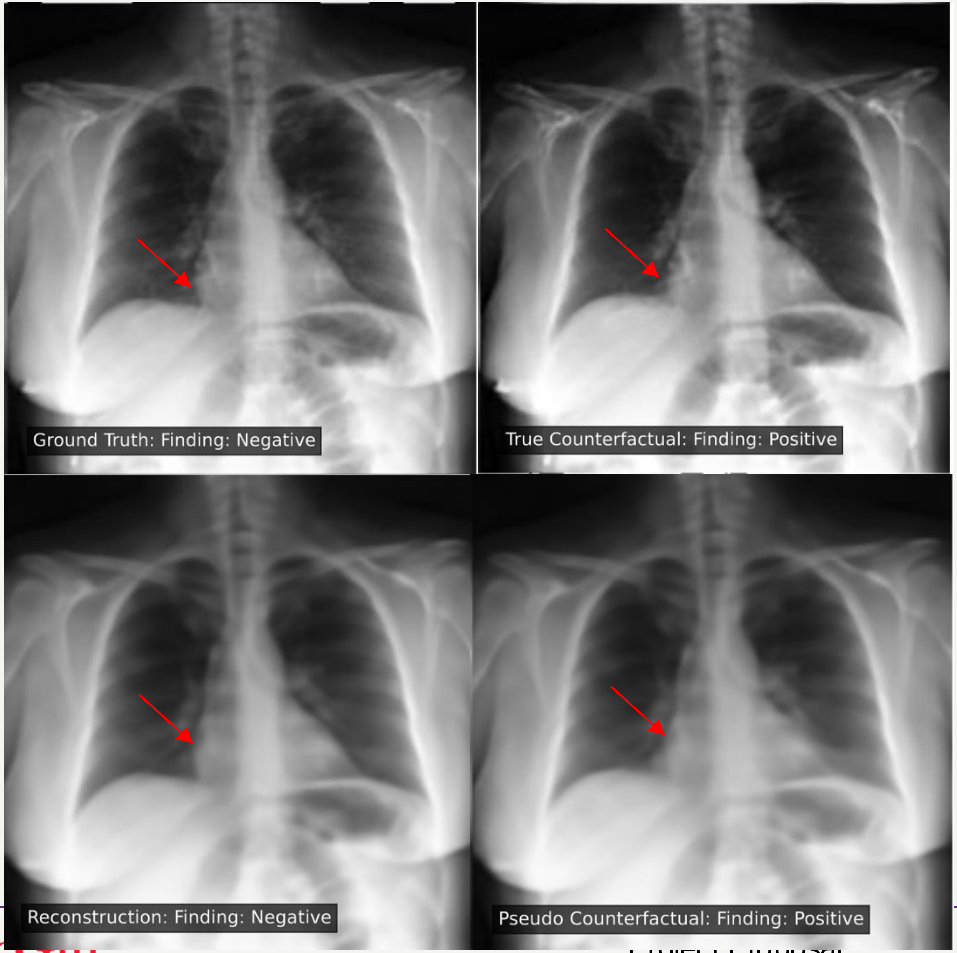
Do: Sex



Do: Age


Ground Truth: Age: 50
True Counterfactual: ['Age: 93', 'Finding: Positive
True Counterfactual: ['Age: 14', 'Finding: Negative'
Reconstruction: Age: 50
Pseudo Counterfactual: Age: 93
Pseudo Counterfactual: Age: 14

Do: Pleural Effusion



Quantitative Metrics of CFs

By the soundness theorem developed by Galles & Pearl, the following properties are necessary in all causal models. The completeness theorem states that these are sufficient

1. Composition

   Definition: Intervening on a variable to have the value it would have without the intervention (again and again) should not affect the other variables
2. Reversibility

   Definition: How well can we go back and forth between changing a specific attribute of the image
3. Effectiveness

   Definition: Intervening on a variable to have a specific value will cause the variable to take on that value

Others to consider…

- Realism - How realistic are the images produced?
- Minimality – does the model change other non-child attributes?

Potential Improvements to CF gen.

- Refine learning processes for SCM mechanisms
  - E.g., I used Gumbel softmax trick to learn sampling from cond. Categorical dist,
  - Are there other losses to use other than KL-div to ensure proper learning
- Refine the parent predictor
- Enhance conditioning
- HVAE tuning
  - Latent diffusion ?
- CF training
  - Add terms to loss, come up with terms that can be added to enhance the generation

Pipeline

Train & validate SCM between attributes → Develop conditional SOTA models with & without SCM → Generate (pseudo) counterfactuals → Define scenarios & compare performances

Define scenarios & compare performances

- **Generalization to underrepresented/unseen data:**
  - Scenario: How do SCM vs non-SCM conditional models perform on generation of underrepresented subgroups?
  - *Rationale: SCMs might handle OOD data better than non-SCM conditional models due to causal understanding. Non-SCM models might struggle without having seen similar data during training.*
  - Metrics:
    - **Uncertainty estimation –** does having an SCM make more/less certain?
    - Diversity Score (variety of generated outputs within underrepresented groups)
    - Realism between subgroups – does it looking in-dist.?
- **Estimating effect with Mediator Present:**
  - Scenario: How do SCM vs non-SCM conditional models perform in generation when mediators are present?
  - *Rationale: Non-SCM models may struggle to disentangle the contributions of each attribute to the image due to the lack of causal understanding, potentially leading to biased or incorrect generation*
  - *Metrics:*
    - *Direct/indirect effect estimation*
- **(Pseudo) - Counterfactual Generation**:
  - Scenario: How do SCM vs non-SCM conditional models perform in (pseudo) counterfactual generation?
  - *Rationale: Same as above.*
  - *Metrics:*
    - *Effectiveness*
    - *Composition*