# Semi-Supervised Learning for Deep Causal Generative Models

Yasin Ibrahim[1], Hermione Warr[1], and Konstantinos Kamnitsas[1,2,3]

[1] Department of Engineering Science, University of Oxford, Oxford, UK
{firstname.lastname}@eng.ox.ac.uk
[2] Department of Computing, Imperial College London, London, UK
[3] School of Computer Science, University of Birmingham, Birmingham, UK

**Abstract.** Developing models that are capable of answering questions of the form "How would $x$ change if $y$ had been $z$?" is fundamental to advancing medical image analysis. Training causal generative models that address such counterfactual questions, though, currently requires that all relevant variables have been observed and that the corresponding labels are available in the training data. However, clinical data may not have complete records for all patients and state of the art causal generative models are unable to take full advantage of this. We thus develop, for the first time, a semi-supervised deep causal generative model that exploits the causal relationships between variables to maximise the use of all available data. We explore this in the setting where each sample is either fully labelled or fully unlabelled, as well as the more clinically realistic case of having different labels missing for each sample. We leverage techniques from causal inference to infer missing values and subsequently generate realistic counterfactuals, even for samples with incomplete labels. Code is available at: https://github.com/yi249/ssl-causal

**Keywords:** Causal Inference · Semi-Supervised · Generative Models

## 1 Introduction

The deployment of deep learning models to real-world applications faces a variety of challenges [16], with many arguing that this is due to lack of causal considerations [4,18]. A growing research area is the generation of counterfactuals (CFs), the manifestation of a sample in an alternative world where an upstream variable has been changed [10,17,21]. Such techniques are particularly useful in medical image analysis, where models are often hampered by lack of diversity in training data [5], so methods to generate realistic synthetic samples from underrepresented classes are critical [13]. Incorporating structural causal equations into a deep learning framework has been shown to provide a powerful tool for counterfactual generation [17]. These ideas were extended by the development of a hierarchical VAE structure for greater image fidelity [7]. This method consists, however, of a separately trained generative model and structural causal model, represented by a directed acyclic graph (DAG), and hence, the two components

are unable to leverage information from one another during training. Moreover, these methods rely on fully labelled samples, so are unable to use additional data where true values are unavailable for some (or all) variables of the causal graph.

Data with limited labels are ubiquitous, so semi-supervised methods are of particular interest. A common approach to semi-supervised learning is consistency regularisation under transformations of the input [1, 11, 22]. Alongside our generative model, we present this approach from a causal perspective and demonstrate how it fits naturally into our framework. Semi-supervised methods also have a causal motivation [19] due to the principle of independence of cause and mechanism (ICM) [18], which suggests that possessing information on the effect (image) alone is beneficial for learning the joint distribution of cause (labels) and effect. In summary, we make the following contributions:

- Introduce a semi-supervised deep causal generative model,

- Generate and evaluate counterfactuals with missing causal variables,

- Provide a causal perspective on the consistency regularisation technique for semi-supervised learning,

- Inspired by the ICM, investigate the performance of our method when parent variables are missing versus when child variables are missing.

To illustrate this, we first use a semi-synthetic dataset based on Morpho-MNIST [6] which allows us to explicitly measure performance given the known underlying causal relationships. We then assess our method on the MIMIC-CXR dataset [9, 12] to demonstrate its capabilities on real, more complex, medical data.

## 2   Background

A (Markovian) Structural Causal Model (SCM) is a 4-tuple [21]:

$$\mathcal{M} = \langle V, U, \mathcal{F}, P(U) \rangle$$

where, $V = \{v_1, \ldots, v_n\}$ is the set of endogenous variables of interest, $U = \{u_1, \ldots, u_n\}$ is the set of exogenous (noise) variables, $P(U)$ is the prior distribution over them, and $\mathcal{F} = \{f_1, \ldots, f_n\}$ is a set of functions assigning values to the endogenous variables. Moreover, we assume that each endogenous variable, $x_i$, is assigned deterministically by its direct causes, i.e. its parents $\mathrm{pa}_i \subseteq V \setminus \{v_i\}$, and the corresponding noise variable, $u_i$ via the structural assignments,

$$v_i := f_i(\mathrm{pa}_i, u_i). \tag{1}$$

This supersedes conventional Bayesian approaches as it allows for greater control by explicitly considering the structural relationships between variables. We achieve this through the do-operation [3], which makes assignments of the form

$\mathrm{do}(x_i = a)$. This disconnects $x_i$ from its parents, and we obtain an intervened distribution over the endogenous variables,

$$P(V|\mathrm{do}(v_i = a)) = \prod_{j \neq i} p(v_j|pa_j) \cdot \mathbb{1}_{\{v_i=a\}}$$

However, such interventions provide only population-level effect estimations [3]. To narrow this down to unit-level and generate counterfactuals for individual samples, the following procedure is carried out [18]:

1. **Abduction**: Use the data to update the prior probability on the exogenous noise $p(U)$ to obtain $p(U|V)$

2. **Action**: Perform an intervention $\mathrm{do}(V = A)$ to obtain a modified SCM, denoted by $\tilde{\mathcal{M}}_{\mathrm{do}(V=A)}$.

3. **Prediction**: Use $\tilde{\mathcal{M}}_{\mathrm{do}(V=A)}$ to estimate the values of the desired variables.
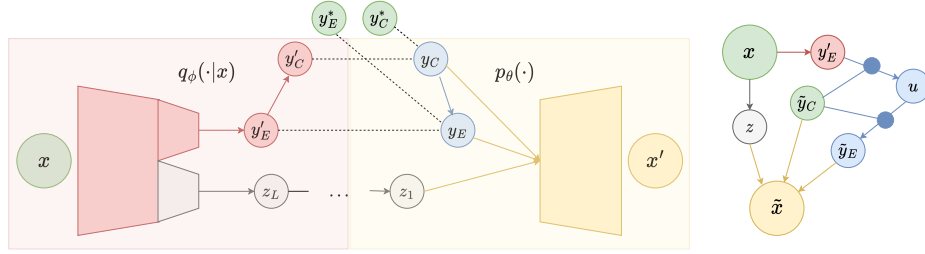
## 3   Methodology



Fig. 1: Model outline. Green: observed, Grey: latent, Red: predicted, Blue: causal generative, Yellow: decoding. (left) Training; we use the $y$ predictions for decoding unless they are observed, (right) CF generation.

An overview of our method is shown in Fig. 1. Herein, the endogenous variables consist of image, $x$, and variables, $y$; denoted by $y^*$ when observed and by $y'$ when predicted. Latent variables $z = z_{1:L}$ make up part of the exogenous noise for $x$, modelled using a hierarchical latent structure, following [7, 20]. Our model extends this structure with a predictive part that infers the endogenous variables, $y$, enabling counterfactual generation in the case of missing labels. For clarity, we limit derivations to the case with a single cause variable, $y_C$, and effect variable, $y_E$, but this can be extended to any finite number of variables with any causal structure. The ELBO loss for **fully labelled samples** drawn from $\mathcal{D}_L$ is $\mathcal{S}(x, y)$:

$$\log p_\theta(x) \geq \mathbb{E}_{q_\phi(z|x,y_E,y_C)} \left[\log \frac{p_\theta(x|z,y_E,y_C)p_\theta(z)p_\theta(y_E|y_C)p_\theta(y_C)}{q_\phi(z|x,y_E,y_C)}\right]$$
$$\Rightarrow \mathcal{S}(x,y) := -\mathcal{L}(x,y) - \log p_\theta(y_E|y_C) - \log p_\theta(y_C)$$

where $\mathcal{L}(x, y)$ is the ELBO for a conditional VAE and $p_\theta(\cdot)$ are (Gaussian) priors. For the **unlabelled samples**, drawn from $\mathcal{D}_U$, we minimise the loss:

$$
\begin{aligned}
\mathcal{U}(x) := &-\mathbb{E}_{q_\phi(y_E|x)}\big[\mathbb{E}_{q_\phi(y_C|x,y_E)}(\mathcal{L}(x,y) - D_{\mathrm{KL}}\{q_\phi(y_C|x,y_E)||p_\theta(y_C)\})\big] \\
&+ D_{\mathrm{KL}}\{q_\phi(y_E|x)||p_\theta(y_E|y_C)\}.
\end{aligned}
\tag{2}
$$

Here, we predict the labels using $y' \sim q_\phi(y|x)$ and regularise them via the KL-divergence with their respective prior $p_\phi(y)$. When **only cause $\mathbf{y_C}$ is labelled**, for samples $(x, y_C) \in \mathcal{D}_C$, we minimise the loss:

$$
\mathcal{C}(x, y_C) := -\mathbb{E}_{q_\phi(y_E|x)}[\mathcal{L}(x,y)] - \log p_\theta(y_C) + D_{\mathrm{KL}}\{q_\phi(y_E|x)||p_\theta(y_E|y_C)\}. \tag{3}
$$

When **only effect $\mathbf{y_E}$ is labelled**, we minimise the loss:

$$
\mathcal{E}(x, y_E) := -\mathbb{E}_{q_\phi(y_C|x,y_E)}[\mathcal{L}(x,y) + \log p_\theta(y_E|y_C)] + D_{\mathrm{KL}}\{q_\phi(y_C|x,y_E)||p_\theta(y_C)\},
$$

for samples $(x, y_E) \in \mathcal{D}_E$. In the case of discrete variables, when the true labels, $y*$, are not provided, we supplement these losses by inversely weighting samples with missing labels by the entropy of the labels predicted by the encoder, $H_\phi(y'|x)$. For example, when $y_E$ is missing, we multiply the expectation in (3) by $1 - H_\phi(y'_E|x)$. We use entropy here as an indicator for predictive uncertainty, to inform us how much to 'trust' the predicted label.

Under the current construction, the parent predictors, $q_\phi(y_E|x)$ and $q_\phi(y_C|x, y_E)$, are only trained when the parent variables are unobserved. To ensure the model is able to learn a good representation of $y$, we include additional classification terms in the supervised loss [14], giving the **total loss** to train our model:

$$
\begin{aligned}
\mathcal{T}(x, y) := &\sum_{(x,y)\in\mathcal{D}_L} \mathcal{S}(x,y) + \sum_{x\in\mathcal{D}_U} \mathcal{U}(x) + \sum_{(x,y_C)\in\mathcal{D}_C} \mathcal{C}(x,y_C) \\
&+ \sum_{(x,y_E)\in\mathcal{D}_E} \mathcal{E}(x,y_E) - \mathbb{E}_{(x,y)\in\mathcal{D}_L}\left[\log q_\phi(y_i|y_{<i}, x)\right]
\end{aligned}
\tag{4}
$$

In the last term, labeled variables $y_i$ are placed in a topological ordering [8] starting with the root nodes. Thus for all $i$, the ancestors of $y_i$ are a subset of $y_{<i}$ and its descendants are a subset of $y_{>i}$. In (2), we see that in the unlabelled case, we require expectations over the labels. For a single discrete label $y$, this can be achieved by summing over all possible values of $y$ [14],

$$
\mathbb{E}_{q_\phi(y|x)}[f(y,\cdot)] = \sum_y q_\phi(y|x) \cdot f(y,\cdot).
$$

However, this quickly becomes computationally expensive as the number of variables or classes grows, and is intractable for continuous variables. We avoid using a Monte-Carlo sampler, where taking more samples leads to similar computational costs. Instead, we propose lowering this variance by beginning training using only the labelled data. By doing so, predictors $q_\phi(\cdot|x)$ reach a sufficiently high accuracy before being used to estimate missing labels for the rest of the data.

**Counterfactual Regularisation** To further improve our model, we turn to a causal treatment of consistency regularisation [22]. For this, we restrict perturbations of the input image to interventions on the DAG governing the causal variables. For example, for input image $x$ with variables $(y_C, y_E)$, we alter effect variable $y_E$ via $\text{do}(y_E = \tilde{e})$, to obtain new image $\tilde{x}$ with causal variables $(\tilde{y}_C, \tilde{y}_E)$. If the DAG is obeyed, the cause variable should remain invariant to this perturbation and we can thus impose a loss that penalises divergence between $y_C$ and $\tilde{y}_C$. In the context of our model, we predict $\tilde{y}_C \sim q_\phi(\cdot|\tilde{x})$ and minimise $D(y_C, \tilde{y}_C)$, where $D(\cdot, \cdot)$ is an appropriate distance metric. If $y_C$ is unknown, we predict its value using $y_C \sim q_\phi(\cdot|x)$. Suppose instead we alter the cause variable; in this case, the effect should change given the DAG. As such, we predict $\tilde{y_E} \sim q_\phi(\cdot|\tilde{x})$ and then compute the counterfactual of $y_E$ under the proposed intervention on $y_C$. When $y_E$ is unlabelled, we first predict it using $y_E \sim q_\phi(\cdot|x)$. This causal interpretation improves the robustness not only of the generative component of the model, but also the causal inference elements, $p_\theta(y_i|y_{<i})$.

**Counterfactual Generation** Once the generative model is trained, we use the predictive component $p_\theta(y_i|y_{<i})$ to generate counterfactuals, $\tilde{x}$. For abduction we require the structural assignments (1) to be invertible in $u$, so we encode $p_\theta(y_i|y_{<i})$ as an invertible $g_{y_{<i}}(u)$, $u \sim p(u)$, parameterised by the causal parents $y_{<i}$ [17]. Counterfactual generation can then be expressed as $\tilde{y}_i = g_{\tilde{y}_{<i}}\left(g_{y_{<i}}^{-1}(y_i)\right)$ so that each counterfactual value can be calculated in sequence. If a label $y_i$ is missing, we use our trained predictor $q_\phi(y_i|x, y_{>i})$ to impute it, e.g Fig. 1 (right) for when $y_E$ is unobserved and we intervene on $y_C$.
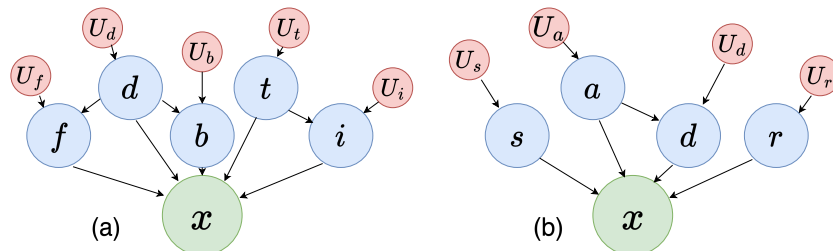
## 4 Experiments



Fig. 2: (a) DAG for Colour MorphoMNIST, $d$: digit, $f$: foreground (digit) color, $b$: background color, $t$: thickness, $i$: intensity. (b) DAG for MIMIC-CXR, $s$: sex, $a$: age, $d$: disease status, $r$: race. $U$: respective exogenous noise variables.

### 4.1 Causal Analysis on Colour Morpho-MNIST

True causal generative processes for medical imaging data are unknown. Therefore, to evaluate the causal aspects of our method in depth, we first use data
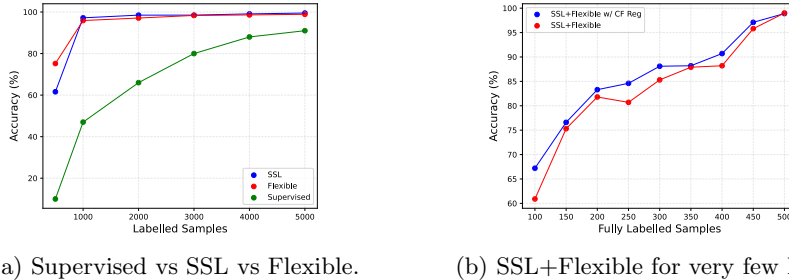
(a) Supervised vs SSL vs Flexible.



(b) SSL+Flexible for very few labels.

Fig. 3: Colour Morpho-MNIST: Accuracy of $\mathrm{do}(d = k)$ on random test images for uniformly random $k \in \{0, \dots, 9\} \setminus d$ where $d$ is the digit of the original image. For SSL, the $x$-axis represents to the number of fully labelled samples; for Flexible it represents the number of labels for each variable across all the samples. For SSL+Flexible we use 600 randomly allocated labels for each variable in addition to the number of fully labelled samples denoted by the $x$-axis.

adapted from Morpho-MNIST (60k training, 10k test samples) [6], considering the thickness ($t$) and intensity ($i$) of each digit. We increase the complexity of the causal structure by colouring the foreground ($f$) and background ($b$), with each colour drawn from a distribution conditional on the digit ($d$), as in Fig. 2a.

**Counterfactual Effectiveness**  As baseline, we use the state of the art supervised method for counterfactual image generation [7] trained only on the labelled samples of each experiment. We compare this against our method for a labelled-unlabelled split (SSL in figures) and for labels missing randomly for each variable (Flexible). We measure the effectiveness [15] of our counterfactual generations by abducting and intervening on test images with random interventions before using classifiers or regressors, $q(\cdot|x)$, trained independently on uncorrelated data, to measure how well the desired change has been captured.

Table 1: Colour Morpho-MNIST: Log likelihoods ($\uparrow$) of the child variables. Colour log likelihoods $\in (-\infty, 2.239]$, intensity log likelihoods $\in (-\infty, -1.336]$.

| Model | Labelled | MAE ($\downarrow$) | $p_\theta(f)$ | $p_\theta(b)$ | $p_\theta(i,t)$ | $q(f|\tilde{x})$ | $q(b|\tilde{x})$ | $q(i,t|\tilde{x})$ |
|---|---|---|---|---|---|---|---|---|
| Supervised | 1000 | 3.91 | -1.46 | -1.49 | -38.21 | -1.01 | -1.17 | -32.98 |
|  | 5000 | 3.84 | -1.31 | -1.38 | -21.38 | -0.63 | -0.93 | -28.44 |
|  | 60,000 | 3.75 | 1.20 | 1.24 | -5.55 | 1.17 | 1.18 | -14.42 |
| SSL | 1000 | 3.85 | 1.05 | 1.10 | -14.26 | 0.81 | 1.08 | -26.02 |
|  | 5000 | 3.83 | 1.10 | 1.12 | -7.40 | 1.01 | 1.19 | -22.39 |
| Flexible | 1000 | 3.86 | 1.11 | 1.13 | -17.93 | 0.77 | 1.15 | -28.20 |
|  | 5000 | 3.84 | 1.14 | 1.16 | -13.98 | 1.07 | 1.10 | -19.56 |

Fig. 3a highlights the improvement by our method over the purely supervised approach. Even when only 1000 samples ($\sim 1.67\%$) are labelled, we achieve near perfect effectiveness for changed digit counterfactuals. This holds both when the data is split into distinct labelled-unlabelled sets and when these labels are

missing randomly. Moreover, in Fig. 3b, counterfactual regularisation improves performance for very low labelled sample sizes by an average of $\sim 2.2\%$. Table 1 demonstrates how the causal relationships are learned significantly better using our method, with regards to both the distributions inferred from the DAG, $p_\theta(\cdot)$, and the manifestations of these changes in the counterfactual image, $q(\cdot|\tilde{x})$.

**Independence of Cause and Mechanism**  Inspired by insights on the ICM [19], we analyse how our method performs in the specific cases of the cause variable missing and the effect present, and vice-versa, by varying the number of thickness and intensity labels while keeping the others. Table 2 (left) shows that the settings with greater proportions of intensity (effect) labels tend to produce better joint distributions, supporting the ICM hypothesis [19]. This is significant for domains with limited labelled data such as healthcare, as it suggests that, given an identified cause-effect relationship and limited capability to obtain labels, focusing on labelling the effect should provide improved performance.

Table 2: (left) Colour Morpho-MNIST: Cause and mechanism experiment. (right) MIMIC-CXR: For each intervention $do(\cdot)$, the 3 rows correspond to training with 10%, 20%, 30% of variables labelled, all using CF regularisation. Semi-supervision in both settings (SSL, Flexible) outperforms pure supervision (Sup.).

| $i$ labels | $t$ labels | $do(t)$ $p_\theta(i,t)$ (↑) | $q(i,t|\tilde{x})$ (↑) | $do(i)$ $||t-\tilde{t}||$ (↓) | $||i-\tilde{i}||$ (↓) |
|---|---|---|---|---|---|
| 300 | 2700 | -21.84 | -29.47 | **0.089** | 0.293 |
| 600 | 2400 | -18.89 | -26.17 | 0.102 | 0.203 |
| 1200 | 1800 | -18.01 | -24.78 | 0.099 | 0.221 |
| 1500 | 1500 | -19.40 | -25.43 | 0.132 | 0.172 |
| 1800 | 1200 | -16.27 | -22.13 | 0.120 | 0.161 |
| 2400 | 600 | -14.62 | -19.32 | 0.146 | 0.093 |
| 2700 | 300 | **-14.15** | **-17.56** | 0.152 | **0.088** |

| | | Disease (↑) | | | Age (↓) | | | Sex (↑) | | | Race (↑) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Sup. | SSL | Flex. | Sup. | SSL | Flex. | Sup. | SSL | Flex. | Sup. | SSL | Flex. |
| $do(d)$ | 10% | 0.55 | 0.70 | **0.71** | 15.62 | 9.12 | 9.19 | 0.95 | 0.95 | 0.94 | 0.71 | 0.71 | 0.68 |
| | 20% | 0.57 | **0.78** | 0.77 | 14.21 | 8.65 | 8.52 | 0.94 | 0.99 | 0.99 | 0.74 | 0.77 | 0.77 |
| | 30% | 0.68 | 0.83 | **0.84** | 12.95 | 8.53 | 7.82 | 0.98 | 1.00 | 0.95 | 0.77 | 0.82 | 0.81 |
| $do(a)$ | 10% | 0.87 | 0.91 | 0.93 | 15.01 | **13.38** | 13.75 | 0.90 | 0.96 | 0.90 | 0.72 | 0.81 | 0.77 |
| | 20% | 0.87 | 0.96 | 0.96 | 15.40 | **12.57** | 13.21 | 0.94 | 1.00 | 0.99 | 0.71 | 0.83 | 0.80 |
| | 30% | 0.90 | 0.96 | 0.95 | 14.26 | **12.07** | 12.15 | 0.98 | 1.00 | 1.00 | 0.78 | 0.85 | 0.84 |
| $do(s)$ | 10% | 0.84 | 0.91 | 0.91 | 14.15 | 9.31 | 9.31 | 0.69 | **0.97** | 0.90 | 0.69 | 0.74 | 0.77 |
| | 20% | 0.86 | 0.97 | 0.96 | 13.57 | 8.25 | 7.87 | 0.73 | **1.00** | 0.99 | 0.69 | 0.80 | 0.83 |
| | 30% | 0.89 | 0.97 | 0.98 | 12.92 | 7.99 | 7.95 | 0.78 | **0.99** | 0.99 | 0.77 | 0.80 | 0.82 |
| $do(r)$ | 10% | 0.84 | 0.93 | 0.95 | 15.08 | 9.76 | 9.73 | 0.96 | 0.98 | 0.95 | 0.46 | **0.53** | 0.52 |
| | 20% | 0.88 | 0.95 | 0.95 | 14.27 | 7.87 | 7.79 | 0.95 | 1.00 | 0.99 | 0.50 | **0.57** | 0.56 |
| | 30% | 0.93 | 0.96 | 0.97 | 14.11 | 7.37 | 7.61 | 0.98 | 1.00 | 1.00 | 0.55 | 0.62 | **0.63** |

### 4.2  Counterfactuals for Medical Imaging Data

To evaluate our method on medical imaging data, we apply it to the MIMIC-CXR dataset (50k training, 20k test samples) [9, 12]. We assume the casual structure used in [7] with variables disease ($d$), age ($a$), sex ($s$), race ($r$), with the only non-independence being that $a$ causes $d$ (Fig. 2b). For disease, we use the presence of pleural effusion as a binary variable and we train models using 10%, 20%, 30%, 40% and 50% of the total labels for each of the three models (Supervised, SSL, Flexible). As the causal structure is simpler, we measure performance, over 3 seeds, by intervening on each variable separately before measuring the ROCAUC for the discrete variables and the MAE for age.

From the cells on the diagonal of Table 2 (right), we see that our method tends to improve upon the supervised approach with regards to implementing interven-
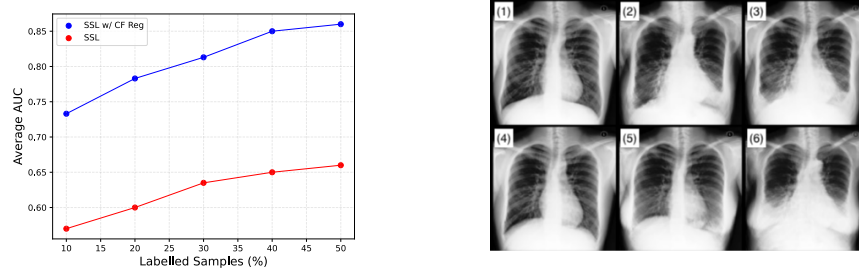
Fig. 4: (a) CF Regularisation on MIMIC-CXR. (b) MIMIC-CXR CFs from model trained on 40% labels. From top-left: (1) original: white, healthy, 20-year-old male, (2) do(age=90), (3) do(diseased), (4) do(asian), (5) do(female), (6) do(all).

tions. The other cells are essentially a measure of reconstruction quality, since they involve evaluating variables that have not been intervened on. As such, the closeness of these values for the various models suggests that the achieved counterfactual generation gains are primarily due to differences in the causal inference component. This indicates that it would be fruitful to focus future efforts on improving this section of the model. This holds for both SSL and Flexible, demonstrating that practitioners implementing our approach need not prioritise achieving full labelling for any given sample over collecting as many labels as possible, bolstering the usability of the model. Fig. 4a demonstrates the increased interventional accuracy provided by CF regularisation. Moreover, as shown in Figure 4b, our model is able to exhibit clear visual changes for the various CFs, indicating the numerical results are not due to minute changes undetectable to the human eye [2]. To build upon this, an avenue of future research would be to use this approach to generate additional training data for underrepresented populations in medical datasets and evaluate how this aids downstream tasks.

## 5    Conclusion

This study introduces a semi supervised deep causal generative model to enable training on causal data with missing labels in medical imaging. Experiments on a coloured Morpho-MNIST dataset, where the whole generative process is known, along with experiments on real clinical data from MIMIC-CXR, demonstrate that our approach uses unlabelled and partially labelled data effectively and improves over the state of the art fully supervised causal generative models. The key practical contribution of this work is that it enables training causal models on clinical databases where patient data may have missing labels, which previous models could not use, relaxing one of the main requirements for training a causal model. A limitation of our work is that we assume the DAG structure is known a priori. Hence, if this is misspecified, there are no guarantees on the correctness of the generated counterfactuals. A possible next step could thus be to explore cases with limited information on the DAG structure of the causal variables.

**Disclosure of Interests.** The authors have no competing interests to declare that are relevant to the content of this article.

# References

1. Bachman, P., Alsharif, O., Precup, D.: Learning with pseudo-ensembles. ArXiv **abs/1412.4864** (2014) 2
2. Banerjee, I., Bhimireddy, A.R., Burns, J.L., Celi, L.A., Chen, L.C., Correa, R., Dullerud, N., Ghassemi, M., Huang, S.C., Kuo, P.C., Lungren, M.P., Palmer, L.J., Price, B., Purkayastha, S., Pyrros, A., Oakden-Rayner, L., Okechukwu, C., Seyyed-Kalantari, L., Trivedi, H., Wang, R., Zaiman, Z., Zhang, H., Gichoya, J.W.: Ai recognition of patient race in medical imaging: a modelling study. The Lancet. Digital health **4**, e406 – e414 (2021) 8
3. Bareinboim, E., Correa, J.D., Ibeling, D., Icard, T.F.: On pearl's hierarchy and the foundations of causal inference. Probabilistic and Causal Inference (2022) 2, 3
4. Bengio, Y., Courville, A.C., Vincent, P.: Representation learning: A review and new perspectives. IEEE Transactions on Pattern Analysis and Machine Intelligence **35**, 1798–1828 (2012) 1
5. Coelho de Castro, D., Walker, I., Glocker, B.: Causality matters in medical imaging. Nature Communications **11** (12 2020) 1
6. Castro, D.C., Tan, J., Kainz, B., Konukoglu, E., Glocker, B.: Morpho-MNIST: Quantitative assessment and diagnostics for representation learning. Journal of Machine Learning Research **20**(178) (2019) 2, 6
7. De Sousa Ribeiro, F., Xia, T., Monteiro, M., Pawlowski, N., Glocker, B.: High fidelity image counterfactuals with probabilistic causal models. In: Proceedings of the 40th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 202, pp. 7390–7425 (07 2023) 1, 3, 6, 7
8. Gagrani, M., Rainone, C., Yang, Y., Teague, H., Jeon, W., Hoof, H.V., Zeng, W.W., Zappi, P., Lott, C., Bondesan, R.: Neural topological ordering for computation graphs (2022) 4
9. Goldberger, A.L., Amaral, L.A.N., Glass, L., Hausdorff, J.M., Ivanov, P.C., Mark, R.G., Mietus, J.E., Moody, G.B., Peng, C.K., Stanley, H.E.: Physionet: Components of a new research resource for complex physiologic signals". circu-lation vol (2000) 2, 7
10. Hess, K., Melnychuk, V., Frauen, D., Feuerriegel, S.: Bayesian neural controlled differential equations for treatment effect estimation. In: The Twelfth International Conference on Learning Representations (2024) 1
11. Hu, W., Miyato, T., Tokui, S., Matsumoto, E., Sugiyama, M.: Learning discrete representations via information maximizing self-augmented training. ArXiv (2017) 2
12. Johnson, A.E.W., Pollard, T.J., Berkowitz, S.J., Greenbaum, N.R., Lungren, M.P., ying Deng, C., Mark, R.G., Horng, S.: Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. Scientific Data **6** (2019) 2, 7

13. Jones, C., Castro, D.C., Ribeiro, F.D.S., Oktay, O., McCradden, M., Glocker, B.: No fair lunch: A causal perspective on dataset bias in machine learning for medical imaging (2023) 1
14. Kingma, D.P., Rezende, D.J., Mohamed, S., Welling, M.: Semi-supervised learning with deep generative models 4
15. Monteiro, M., Ribeiro, F.D.S., Pawlowski, N., Castro, D.C., Glocker, B.: Measuring axiomatic soundness of counterfactual image models (2023) 6
16. Paleyes, A., Urma, R.G., Lawrence, N.D.: Challenges in deploying machine learning: A survey of case studies. ACM Computing Surveys **55**, 1 – 29 (2020) 1
17. Pawlowski, N., Castro, D.C., Glocker, B.: Deep structural causal models for tractable counterfactual inference. In: Advances in Neural Information Processing Systems (2020) 1, 5
18. Peters, J., Janzing, D., Schölkopf, B.: Elements of Causal Inference: Foundations and Learning Algorithms. Adaptive Computation and Machine Learning, MIT Press (2017) 1, 2, 3
19. Schölkopf, B., Janzing, D., Peters, J., Sgouritsa, E., Zhang, K., Mooij, J.: On causal and anticausal learning. Proceedings of the 29th International Conference on Machine Learning, ICML 2012 **2** (06 2012) 2, 7
20. Sønderby, C.K., Raiko, T., Maaløe, L., Sønderby, S.K., Winther, O.: Ladder variational autoencoders 3
21. Xia, K., Lee, K.Z., Bengio, Y., Bareinboim, E.: The causal-neural connection: Expressiveness, learnability, and inference. In: Neural Information Processing Systems (2021) 1, 2
22. Xie, Q., Dai, Z., Hovy, E.H., Luong, M.T., Le, Q.V.: Unsupervised data augmentation for consistency training. arXiv: Learning (2019) 2, 5