# Causal and Active Learning-Based Counterfactual Chest X-ray Generation for Supporting Clinical Decision-Making in Lung Disease

**Yifei Zhu** Y.ZHU4@EXETER.AC.UK
**Greta Mohr** G.MOHR@UCL.AC.UK
**Lei Zhang** L.ZHANG6@EXETER.AC.UK
**Chris Sainsbury** CHRIS@SAINSBURY.IM
**Feng Dong** FENG.DONG@STRATH.AC.UK
**John MacLay** JOHN.MACLAY@GLASGOW.AC.UK
**David J. Lowe** DAVID.LOWE@GLASGOW.AC.UK
**David Lagnado** D.LAGNADO@UCL.AC.UK
**Xujiong Ye** X.YE2@EXETER.AC.U

**Editors:** Bijan Mazaheri and Niels Richard Hansen

## Abstract

Lung diseases such as lung cancer are major contributors to global morbidity, requiring accurate diagnostic decisions for optimal patient outcomes. While deep learning has advanced medical imaging, the lack of causal inference limits its clinical utility. This study proposes a causal generative framework for counterfactual analysis of Chest X-rays, guided by expert model supervision to ensure clinical plausibility. To solve data imbalance and enhance robustness, we introduce a recurrent active learning strategy that utilises "forgetting rates" to select informative samples. Experimental results demonstrate effectiveness improvements of 9.25% on the MIMIC dataset and 13.40% on ChestXray8. Furthermore, two-stage human expert evaluations confirm that the model generates highly realistic synthetic data that maintains a clinical heavy-tailed distribution. These high-quality counterfactuals not only improve diagnostic accuracy but also facilitate confidence calibration for clinicians through interpretable evidence. Our findings demonstrate that integrating causal modeling with expert supervision and active learning provides a robust, clinically meaningful tool for pulmonary diagnostic decision-making.

**Keywords:** Counterfactual, structure causal model, medical imaging, active learning, expert model, hierarchical variational autoencoder

## 1. Introduction

Breathlessness is common in hospitalised patients and is often a harbinger of serious and potentially fatal pathology Rao and Gray (2003). Patients with undifferentiated respiratory distress require urgent attention in the emergency department DeVos and Jacobson (2016). With lung diseases affecting millions globally and lung cancer maintaining high incidence-mortality ratios Choi and Jung (2023), accurate diagnosis and effective treatment planning are crucial for managing these diseases. Recent advancements in artificial intelligence (AI), particularly in deep learning, have demonstrated considerable promise in supporting the diagnosis of pulmonary diseases through medical imaging modalities Cohen et al. (2022); Jasmine Pemeena Priyadarsini et al. (2023); Al-Sheikh et al. (2023); Hage Chehade et al. (2025). A growing body of research has validated the efficacy of deep learning

© 2026 Y. Zhu, G. Mohr, L. Zhang, C. Sainsbury, F. Dong, J. MacLay, D.J. Lowe, D. Lagnado & X. Ye.

algorithms in the automated detection and classification of various lung conditions, such as pneumonia, tuberculosis, and lung cancer, by leveraging large-scale annotated imaging datasets Johnson et al. (2019); rsn (2018); Wang et al. (2017b). Despite these technological advances, the integration of deep learning algorithms into safety-critical domains like healthcare remains limited, primarily due to persistent real-world challenges that limit their reliability and generalizability in clinical settings D'Amour et al. (2022).

A major limitation of current deep learning models lies in their inability to perform causal inference Bengio et al. (2013); Kusner et al. (2017); Peters et al. (2017); Ribeiro et al. (2023). Conventional architectures, such as generative adversarial networks (GANs) and variational autoencoders (VAEs), primarily capture statistical associations between variables and outcomes, thereby lacking the capacity to answer counterfactual or "what-if" questions that are central to clinical decision-making Schölkopf et al. (2021); Schölkopf (2022); Ribeiro et al. (2023). In contrast, integrating a causal inference framework enables reasoning about the effects of hypothetical interventions, offering the potential to uncover how specific changes in clinical variables may influence patient outcomes Pearl (2009); Imbens and Rubin (2015). For example, estimating the impact of a particular treatment on disease progression could significantly enhance the quality and precision of clinical decisions.

Currently, most causal models are mainly developed for low-dimensional data modalities, such as tabular, numerical, or text inputs Li et al. (2022); Jensen and Andreassen (2008). However, current models remain limited in their applicability to high-dimensional data, particularly medical images. Extending these models to handle such high-dimensional data is essential, as the ability to generate counterfactual images, those showing alternative disease appearances under hypothetical conditions could significantly enhance clinical understanding. The capability would allow clinicians to observe how imaging features or pathologies might vary under different conditions, offering deeper insights into disease mechanisms and potential treatment effects Pawlowski et al. (2020); Castro et al. (2020); Glocker et al. (2023). Ribeiro et al.'s work Ribeiro et al. (2023) focuses on image-level. However, it does not provide systematic assessment of the clinical plausibility or relevance of the generated images, which raises concerns regarding their interpretability and utility in medical decision-making contexts. Additionally, within the realm of generative models, particularly those used for medical image synthesis, there remains a critical limitation in ensuring that the generated images are clinically meaningful Sun et al. (2024). Although metrics like the Fréchet Inception Distance (FID) Heusel et al. (2017) are often used to assess image quality, they fall short in validating the clinical relevance of the images. This shortfall arises not only because most current models lacked prior knowledge (such as medical conditions) as a condition for generation, but also they lack a mechanism to integrate posterior knowledge, including expert feedback, into the learning loop. Incorporating such prior knowledge and feedback could help ensure that generated images not only appear realistic but also align with clinical expectations and contribute to meaningful diagnostic or prognostic insights. We enhance the counterfactual generation model by introducing a mechanism to assess the clinical relevance of the generated data, an essential aspect for real-world medical applications Sun et al. (2024). Additionally, working with real clinical data for counterfactual generation remains challenging due to the complexity and variability inherent in medical imaging. To address these limitations, we incorporate an expert model supervision, which leverages domain expertise to ensure that the generated counterfactuals are clinically meaningful. Furthermore, we employ active learning to iteratively refine the data distribution by incorporating

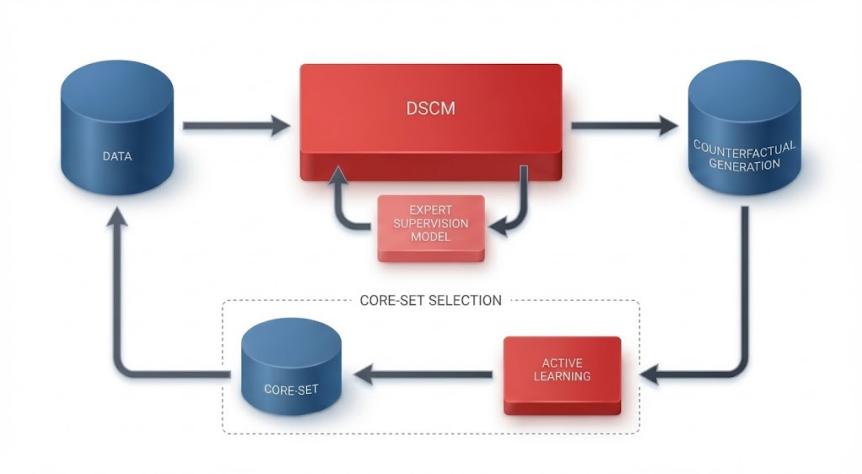



Figure 1: The proposed framework. The proposed model architecture with active learning and human in the loop. The framework begins with data input into the SCM for training, which generates counterfactuals. Expert knowledge is integrated into the SCM training pipeline to ensure the clinical validity of the generated counterfactuals. An active learning component is employed for Core-set Selection, where samples with the highest uncertainty, as generated by the SCM, are selected as the core set. This module identifies the most informative counterfactual samples based on uncertainty, which are then fed back into the SCM to enhance the data distribution as close as the real distribution.

high-uncertainty samples, thereby improving its alignment with the real data distribution and ensuring that the generated counterfactuals are clinically relevant.

This work focuses on causal generative modelling applied to high-dimensional data (e.g. image) for lung disease counterfactual analysis, with the goal of enhancing the understanding of lung diseases using real-world clinical datasets. In particular, we aim to enhance the structured causal model (SCM) by addressing key challenges related to data quality during training. Clinical datasets often contain substantial noise, inconsistencies, and skewed data, leading to deviations between the training set distribution and the true data distribution. To mitigate this issue, we employed active learning to enhance the quality of the training data, incorporating uncertainty measurement of samples to better capture and reflect the clinical disease representations. Our main contributions can be summarised as follows:

- We proposed a counterfactual generation model with expert model supervision, which ensures the generation is clinically meaningful.
- We introduced a recurrent active learning framework using the forgetting rate to balance the data distribution, which improves the robustness of the model.
- We evaluated counterfactual images using a conventional counterfactual evaluation metric, including effectiveness, composition and reversibility.

- We conducted a two-stage human expert evaluation: (1) single-image diagnosis comparing with the real X-ray images, counterfactual X-ray generated by baseline model and the proposed model; and (2) sequential diagnosis where experts first assess a real X-ray and then reevaluate after viewing a counterfactual generation.

## 2. Methods

### 2.1. Preliminary Probabilistic Causal Model Framework

Our framework leverages Structural Causal Models (SCMs) instantiated via deep generative hierarchies to enable robust counterfactual reasoning. Formally, an SCM $\mathcal{M}$ is defined by endogenous variables $\mathbf{X}$, exogenous noise $\mathbf{U}$, and deterministic mechanisms $\mathbf{F}$, satisfying the causal Markov condition Pearl (2009); Ribeiro et al. (2023). Counterfactual inference within this framework follows a three-step procedure, abduction, action, and prediction, to estimate outcomes under hypothetical interventions $do(x_k := c)$. The deep structural causal model is based on a conditional Hierarchical Variational Autoencoder (HVAE) Kingma et al. (2016); Sønderby et al. (2016), extended from standard VAEs Rezende et al. (2014). Unlike traditional architectures, our approach treats the latent variables $\mathbf{z}_{1:L}$ as the exogenous noise $\mathbf{U}$. Crucially, the causal parents $\mathbf{pa}_\mathbf{x}$ are decoupled from the prior $p_{\boldsymbol{\theta}}(\mathbf{z}_{1:L})$ to maintain independence, while injecting them into the generative path to parametrize the likelihood $p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z}_{1:L}, \mathbf{pa}_\mathbf{x})$ Ribeiro et al. (2023). To ensure generated counterfactuals $\tilde{\mathbf{x}}$ exhibit semantically meaningful changes, a variational lower bound on the mutual information between the intervention and the prediction is maximised Barber and Agakov (2004). The training is formulated as a constrained optimization problem using the Lagrangian method:

$$\mathcal{L}_{\text{LG}}(\boldsymbol{\theta}, \boldsymbol{\phi}, \lambda) = \mathcal{L}_{\text{CT}}(\mathcal{M}; \mathbf{x}, \mathbf{pa}_x) - \lambda \left(c - \mathcal{F}_{\text{FE}}(\boldsymbol{\theta}, \boldsymbol{\phi}; \mathbf{x}, \mathbf{pa}_x)\right), \tag{1}$$

where $\mathcal{L}_{\text{CT}}$ represents the counterfactual loss derived from parent predictors (detailed definition is presented in Appendix A), and $\mathcal{F}_{\text{FE}}$ is the observational free energy (negative ELBO), ensuring the model preserves generative quality on observed data $c$ while optimizing for causal consistency. The parameters of the model are denoted as $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$. The details of the baseline model can be found in Appendix A.

### 2.2. Active Learning for Data Distribution Refinement

Overlapping pathological features in chest X-rays introduce dataset noise, complicating clinical differentiation. Moreover, while the majority of samples exhibit similar visual characteristics, only a minority display distinctive features. This imbalance results in a homogenisation effect in synthetic image generation, hindering the model's ability to capture rare yet clinically significant variations in disease presentation. These limitations arise from discrepancies between training and real-world distributions. To address this, we incorporate active learning via core-set selection to enhance data representativeness. By prioritizing informative samples, this approach aligns the model with the true clinical distribution. As illustrated in Figure 1, the module identifies challenging samples and reintroduces them into the DSCM training loop to iteratively refine the synthetic distribution. Formally, Formally, given a labeled dataset $\mathcal{D} = \{(\mathbf{x}_i, y_i) \mid i \in [n]\}$, core-set selection seeks a subset of $m \leq n$ points that approximates the distribution of the complete dataset.

Selection via proxy (SVP) Coleman et al. (2020) is applied to core-set selection. A small deep learning model can serve as an inexpensive proxy for data selection in deep learning Coleman

et al. (2020). This proxy model, $A_k^P$, is created by scaling down deep learning models through layer reduction and smaller architectures. Although $A_k^P$ achieves significantly lower accuracy than the target model, $A_k^T$, it still provides useful representations for point selection. We applied the forgetting events Toneva et al. (2019) for core-set selection. An sample $x_i$ undergoes a forgetting event when its accuracy drops between two consecutive updates, $accuracy_i^{t-1} > accuracy_i^t$, where $accuracy_i^{t-1}$ and $accuracy_i^t$ are a binary variable indicating whether $x_i$ is correctly classified at step $t-1$ and $t$ respectively. In simpler terms, if $x_i$ is correctly classified at step $t-1$ but incorrectly classified at step $t$, it is considered as a forgetting event. Conversely, a learning event is identified when the classification accuracy of $x_i$ improves between two consecutive updates. Formally, this is represented as $accuracy_i^{t-1} < accuracy_i^t$.This signifies that $x_i$ transitions from being misclassified at time step $t-1$ to correctly classified at time step $t$, indicating a successful learning occurrence. Therefore, each input sample has a forgetting rate, and we select $N$ samples with highest forgetting rate to form the core set. The samples with high forgetting rates are located on the classification boundaries, which indicate that the uncertainty is high. In the case of the counterfactual, we have clear labels for the core set, since we perform interventions to obtain the counterfactual. Adding the core set to the training set helps the DSCM to learn the features of lung diseases and reduce uncertainty within the model.

### 2.3. Expert Model Integration

In contrast to conventional generative modelling applications that emphasise visual realism or artistic expression, medical image generation must be grounded in clinical and biological validity Sun et al. (2024). The primary purpose of synthetic medical images is to support downstream tasks such as disease classification, segmentation, or diagnostic decision-making. Therefore, these images must accurately reflect diverse clinical attributes, including disease types and patient demographics, especially in the context of pulmonary imaging. As described in Section Preliminaries, the baseline DSCM model may have limited capacity to generate clinically meaningful features, potentially compromising the plausibility of synthesised chest radiographs without further adaption. To address this limitation, we develop an extended model that incorporates domain knowledge from medical experts, following the approach introduced by Ouyang et al. Ouyang et al. (2022).

An expert model, TorchXrayVision Cohen et al. (2022), is incorporated into the DSCM. As shown in Figure 1 (a), the expert model is integrated into the SCM training pipeline to ensure the clinical validity of the generated counterfactuals. This expert model Ribeiro et al. (2023) is pre-trained on a large chest X-ray dataset annotated by medical professionals and outputs disease probabilities based on the input images. Therefore, we can obtain the observational, $p(\mathbf{x})$, and counterfactual disease probabilities, $p(\tilde{\mathbf{x}})$, and the goal is to maximizing the difference of probabilities between them. To align the generated counterfactuals with expert knowledge, we introduce a constraint of the form $\log(1+\beta|p(\mathbf{x})-p(\tilde{\mathbf{x}})|) \geq d$, enforcing that the predicted disease probabilities under counterfactuals differ from those of the factual inputs by at least a threshold. Scaling factors are denoted as $\alpha$ and $\beta$, and $d$ is a constant. The new Lagrangian optimisation problem of the DSCM, therefore, becomes:

$$\begin{aligned}\mathcal{L}_{\text{LG}}(\boldsymbol{\theta}, \boldsymbol{\phi}, \lambda; \mathbf{x}, \mathbf{pa}_x) = \mathcal{L}_{\text{CT}}(\mathcal{M}; \mathbf{x}, \mathbf{pa}_x) - \lambda\left(c - \mathcal{F}_{\text{FE}}(\boldsymbol{\theta}, \boldsymbol{\phi}; \mathbf{x}, \mathbf{pa}_x)\right) \\ + \alpha \log(1+\beta|p(\mathbf{x}) - p(\tilde{\mathbf{x}})|),\end{aligned} \tag{2}$$

In this work, we selected the logarithm. The logarithmic structure provides a smooth and gradual penalty for deviations, especially when differences are small, which helps mitigate the impact of outliers. The parameter $\beta$ allows for tuning the sensitivity to these differences, offering flexibility in adjusting the model's response to variations between counterfactual generations and observations.

### 2.4. Evaluation Matrix

The necessary constraints of counterfactual inference model are defined as: composition, reversibility and effectiveness Monteiro et al. (2023). **Effectiveness** is defined as that intervening on a set of variables to a specific value directly causes it to take on that valueMonteiro et al. (2023). Therefore, the equality holds as: $Pa(DSCM(\mathbf{x}, \mathbf{pa_x}, \tilde{\mathbf{pa}}_\mathbf{x})) = \tilde{\mathbf{pa}}_\mathbf{x}$, where $Pa(\cdot)$ is an oracle function that outputs the parent variables. **Composition** is defined as that intervening on a variable but keeping its value the same as it will not change other variables in the system. Therefore, the equality holds as: $DSCM(\mathbf{x}, \mathbf{pa_x}, \mathbf{pa_x}) = \mathbf{x}$ since if $\tilde{\mathbf{pa}}_\mathbf{x} = \mathbf{pa_x}$, then $\mathbf{x}$ is not affected. **Reversibility** If the causal mechanism is invertible, the mapping between the observation and counterfactual is deterministic, which follows that if $\tilde{\mathbf{x}} := DSCM(\mathbf{x}, \mathbf{pa_x}, \tilde{\mathbf{pa}}_\mathbf{x})$, and then $\mathbf{x} := DSCM(\tilde{\mathbf{x}}, \tilde{\mathbf{pa}}_\mathbf{x}, \mathbf{pa_x})$. The Composition and reversibility are evaluated by using FID $= \|\mu_1 - \mu_2\|^2 + \text{Tr}(\Sigma_1 + \Sigma_2 - 2(\Sigma_1\Sigma_2)^{1/2})$, where Tr$(\cdot)$ is the trace of a matrix. The mean and covariance, capturing the location and shape of feature distributions for real and generated images, are denoted as $\mu_i$ and $\Sigma_i$, respectively. The effectiveness of the model is quantified using anti-causal parent predictors Ribeiro et al. (2023), while composition and reversibility is quantified using FID and Structural Similarity Index (SSIM). FID evaluates global feature-level consistency and SSIM measures pixel-level structural preservation.

### 2.5. Clinician Evaluation of Generated Images

To assess the usability of the counterfactual images, we carried out a thorough evaluation using participants sourced from medical students and doctors. We particularly wanted to establish the effectiveness of counterfactual images for improving diagnostic accuracy and confidence calibration. The primary objectives of this evaluation were to examine whether the generated counterfactual images exhibit clinically meaningful features, and to determine the extent to which these images can contribute to improved clinical decision-making.

The study sample consisted of 42 participants, aged between 20 and 67 years ($\mu = 26.1$, $\sigma = 7.37$, 24 Female, 18 Male). Participants were recruited via the online platform Prolific and were directed to the online experiment platform Gorilla (https://app.gorilla.sc/). All participants were either medical students (26) or medical doctors (16), with a combined average of 6.20 years medical experience (including study time).

The task consisted of two blocks. **Block 1**: was designed to assess diagnostic capability and directly compare performance between the baseline DSCM proposed in Ribeiro et al. (2023) and the proposed counterfactual model. Participants were shown a series of 36 chest x-ray images, in a randomised order. These images were derived from 12 patients, with each patient image appearing three times; once as the original observation, once as a synthetic image generated using the baseline model and once as a synthetic image generated using our proposed counterfactual model. Importantly, the synthetic images showed the same condition (healthy or suffering from pneumonia or a pleural effusion) to allow for direct comparison. After selecting a diagnosis from the 3 options, participants were asked to rate the realness of the image on a scale from 1 – (definitely AI generated) to 100 (definitely real). They then had the option to select as many reasons they think

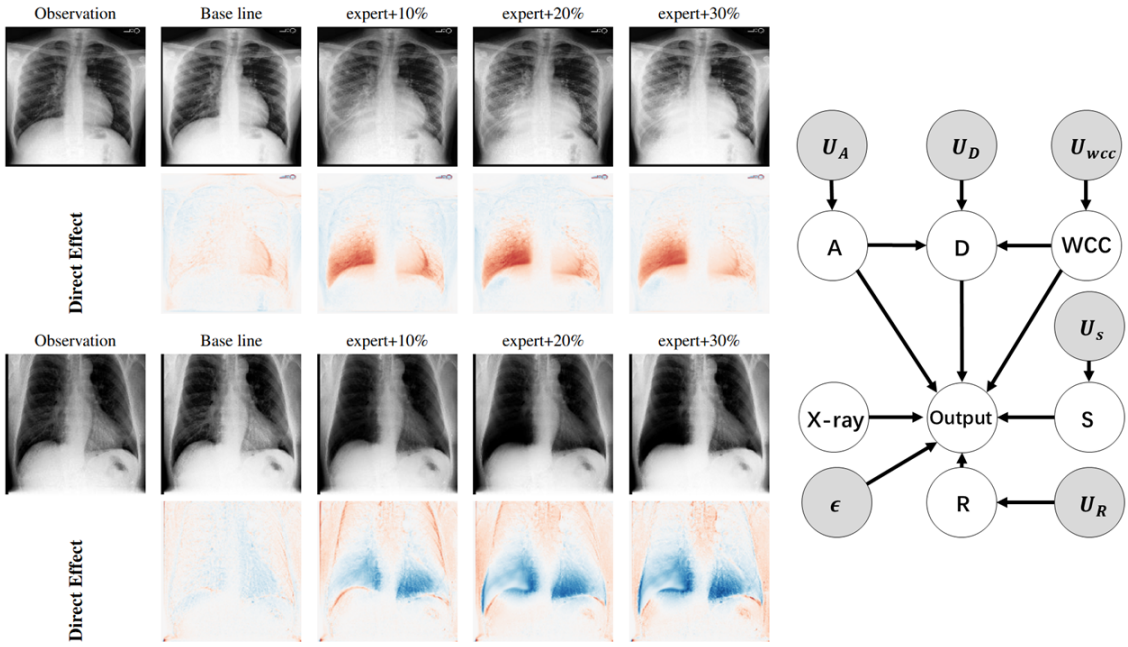



Figure 2: Comparison of Baseline with our proposed counterfactual generated models by adding both expert model, and different percentage of synthetic data via active learning. The red color in the direct effect represents the pixel value increases, and blue color represents the pixel value decreases. The DAG of the structure causal model for MIMIC dataset: age (A), disease (D), race (R), sex (S) and white cell count (WCC).

the image may be synthetic from this list: 1. The lungs or other organs appear too symmetrical; 2. The anatomical structures appear slightly distorted or unnatural; 3. The image feels synthetic or computer-generated; 4. The pathology does not match typical clinical presentation; 5. There are subtle inconsistencies in shading or lighting; 6. The image looks too smooth or lacks natural texture; 7. The image looks overly clean or noise-free.

**Block 2**: Block 2 consisted of 12 cases, in which participants were shown a real chest-ray image depicting either pneumonia or pleural effusion. Participants were asked to make a diagnostic decision and rate their confidence on a scale from 1 (not confident at all) to 6 (very confident). On the next screen, participants were then shown the same image again alongside a synthetic image showing what the chest x-ray would look like if the patient were healthy (generated using either the base line or our model). They were then asked to make a final diagnostic decision on the real X-ray and rate their confidence.

## 3. Results

### 3.1. Dataset and Preprocessing

We evaluated the proposed approach with the MIMIC-CXR dataset Johnson et al. (2019) and ChestX-ray8 Wang et al. (2017a). **MIMIC**: The generative model assumed the following observed

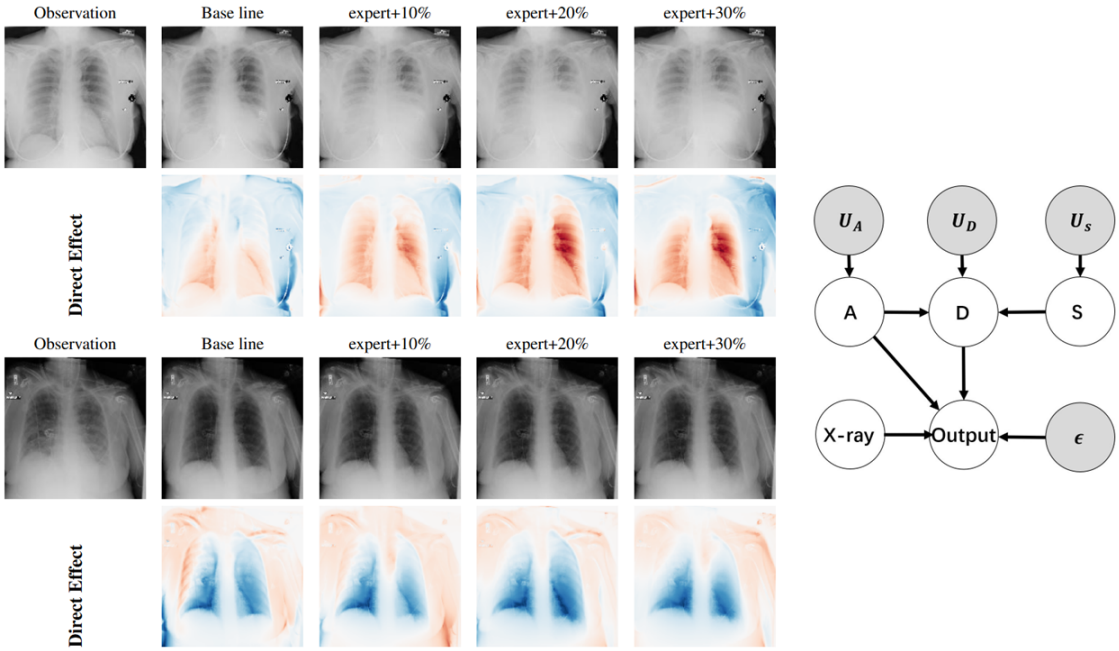



Figure 3: Comparison of Baseline with our proposed counterfactual generated models by adding both expert model, and different percentage of synthetic data via active learning. The red color in the direct effect represents the pixel value increases, and blue color represents the pixel value decreases. The DAG of the structure causal model for ChestXray8 dataset: age (A), disease (D), and sex (S).

variables: age (A), sex (S), race (R), disease (D), white cell count (WCC) and X-ray images. The corresponding directed acyclic graph (DAG) is shown in Figure (1) b. The MIMIC-CXR dataset comprises 227,835 imaging studies from 65,379 patients treated at the Beth Israel Deaconess Medical Center Emergency Department between 2011 and 2016 Johnson et al. (2019). Each imaging study is accompanied by a semi-structured, free-text radiology report describing the findings. The disease labels were provided by the CheXpert system Irvin et al. (2019), which automatically interprets the presence of 14 disease observations from the reports. **ChestX-ray8**: The generative model assumed the following observed variables: age (A), sex (S), disease (D) and X-ray images. The ChestX-ray8 dataset comprises 112,120 X-ray images with disease labels form 30,805 unique patients. The disease labels were classified by using natural language processing to text-mine from the associated radiological reports. We resized the chest X-ray images to $192 \times 192$ resolution. For this study, we focused on two diseases: pneumonia and pleural effusion. Accordingly, the disease (D) node takes on three discrete values: 0 (no finding), 1 (pleural effusion), and 2 (pneumonia). The training dataset consisted of 60,000 samples, with 20,000 for each scenario. Experiment setup and implementation is listed in Appendix B, and the participants information is introduced in Appendix C.

| Dataset | Effectiveness | Composition | | Reversibility | |
|---|---|---|---|---|---|
| MIMIC | DISEASE (d) | SSIM | FID | SSIM | FID |
| Base-line | 0.8487 | 0.9496 | 0.5439 | 0.9450 | 0.6425 |
| Ours+10% | 0.9056 | 0.9715 | 0.1451 | 0.9670 | 0.2361 |
| Ours+20% | 0.9193 | 0.9783 | 0.1387 | 0.9768 | 0.2357 |
| Ours+30% | 0.9272 | 0.9879 | 0.0608 | 0.9761 | 0.2942 |
| CXR-8 | | | | | |
| Base-line | 0.8155 | 0.9999 | 4.5e-10 | 0.9862 | 0.1242 |
| Ours+10% | 0.8687 | 0.9999 | 4.4e-10 | 0.9901 | 0.1410 |
| Ours+20% | 0.9088 | 0.9999 | 4.4e-10 | 0.9855 | 0.2784 |
| Ours+30% | 0.9248 | 0.9999 | 4.3e-10 | 0.9813 | 0.3511 |

Table 1: **The counterfactual evaluation of the proposed model.**

### 3.2. Counterfactual Image Generation with Active Learning

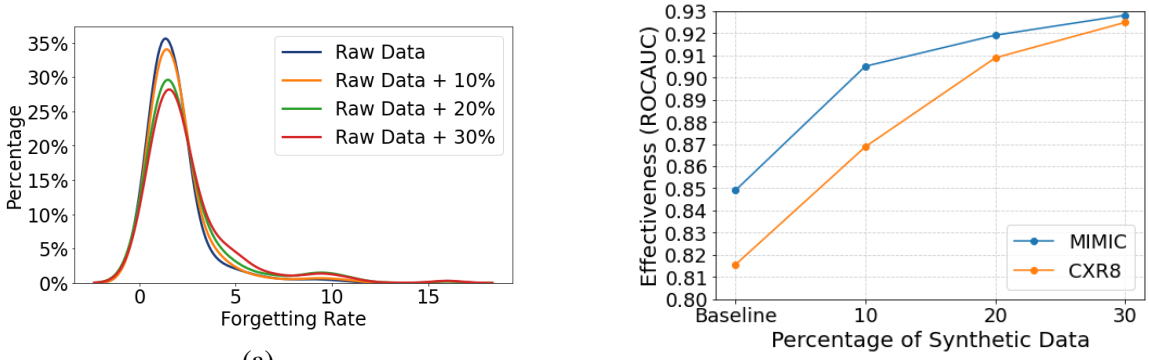
In our experiments, we first trained a baseline model proposed in Ribeiro et al. (2023) with the DAG shown in Figure (2) and (3) using only data from the MIMIC dataset or ChestXray8 dataset. The counterfactuals generated by this baseline model showed limited variation and minimal degrees of change, as shown in Figure (2) and (3). We used the forgetting rate as a score function to quantify the uncertainty of the dataset. We incrementally introduced synthetic data generated from the SCM with the highest forgetting rates, adding an amount equivalent to 10% of the training set size in each iteration. Figure (2) and (3) present the impact of incrementally incorporating synthetic data into the training process, where each column represents an increasing proportion of synthetic data (10%-30%). As the proportion of synthetic data increased, the causal model learnt more effectively, leading to counterfactual generations that better capture clinically relevant patterns. Figure (3) and (3) demonstrate that with more synthetic data, the model produced clearer and more meaningful counterfactual effects, enhancing its ability to reflect underlying causal mechanisms in medical imaging. To evaluate our model, we evaluated effectiveness, composition and reversibility, which are fundamental properties of counterfactuals that are universally valid across all causal models Monteiro et al. (2023). Table (1) illustrates the evaluation matrix and the evaluation curve shown in Figure (4) (b). The source code and pre-trained models are publicly available at [1].

### 3.3. Clinician-assisted Evaluation

This section focusses on the evaluation of our results conducted by clinicians. The clinicians evaluate our results by completing the tasks described in Section Expert Evaluation. The goal of these experiments is to evaluate whether the generated counterfactual X-rays are clinically meaningful and practically useful in supporting clinical decision-making.

The results of **Block 1** shows that participants achieved higher accuracy when assessing images generated by our proposed model (36.9%) compared with the base line model (34.6%), reflecting a numeric improvement of $\Delta = 2.3\%$. Specifically, the within-patient comparisons demonstrated that our proposed model facilitated better diagnostic judgments for pleural effusion ($\Delta = 2.94\%$) and healthy cases ($\Delta = 4.41\%$) compare to the baseline. Regarding visual fidelity, a paired-samples

1. https://github.com/YeefeiZZZ/causal-gen-lung




(a) (b)

Figure 4: a) Uncertainty Distribution of the datasets, where the x-axis is forgetting rate representing a quantification of uncertainty and the y-axis is the percentage of the data. b) X-axis represents the percentage of synthetic data added into the training set, and the y-axis represents the Receiver Operating Characteristic Area Under the Curve (ROCAUC).

t-test demonstrated that images grenerated by our proposed model were rated as significantly more realistic than the baseline model ($t(33) = -3.29, p = .002, \Delta = 7.02, \Delta = 7.02$). Notably, realism rating across all conditions hovered near the scale's midpoint ( $50\%$), suggesting that participants generally found it difficult to definitively distinguish between real and synthesised images. Examining the number of reasons provided by participants for these judgments, participants selected an average of 15.65 reasons for images generated by our proposed model, compared to 14.29 for real images. The difference between the counterfactuals generated by our model and real images was not significant, further supporting the indistinguishability of the generated images.

The results of **Block 2** shows that participants' first decision accuracy was again relatively low, with a mean of $51\%$, indicating a high level of difficulty or uncertainty during initial judgments. To evaluate how effectively each counterfactual model supported diagnostic revision, we analysed only those cases where the participant's initial decision was incorrect. Figure (5) (a) in Appendix D visualises these differences and highlights that regardless of model, having a counterfactual image was always helpful as average performance always improved.This focused the analysis on instances where correction was possible. Our model showed the greatest improvement, with participants correcting their decisions on $27.91\%$ of these cases, compared to $26.32\%$ for the base line model. Importantly, when excluding the highest improvement patients for each counterfactual model, the difference in accuracy change increases from $1.59\%$ to $3.87\%$, suggesting that the baseline model's performance was disproportionately elevated by outliers, while the our model showed more consistent, robust improvements across participants.

Beyond accuracy, we examined whether counterfactual images helped participants align their confidence with their actual diagnostic performance. Both models showed negligible correlations between confidence and accuracy ($r \approx 0$) at initial diagnose. Results are visualised in Figure (5) (b) in Appendix D. However, a critical divergence emerged upon the presentation of the counterfactual images. For our proposed model, the correlation between confidence and accuracy improved substantially during the second decision($r(32) = 0.11, p = 0.106$), indicating that our counterfactual images helped participants better assess the correctness of their judgments. Conversely, the baseline

model's calibration declined ($r(32) = -0.5, p = 0.46$), implying that the counterfactuals generated by baseline model were not beneficial and may have introduced confusion. The improvement was particularly notable for specific pathologies. While the baseline model showed consistently poor calibration across conditions, our proposed model drove significant improvements in pneumonia cases. Calibration for pneumonia improved from a non-significant baseline to a statistically significant positive correlation at the second decision ($r(32) = 0.20, p = 0.046$). This indicates that our counterfactuals successfully bridge the gap between clinical judgment and diagnostic reality, allowing clinicians to form more justified and accurate confidence assessments.

## 4. Discussion

Model accuracy alone does not ensure that the clinical observations represented in the counterfactuals are reliable or clinically meaningful. To address this, we incorporated an expert model (eg. TorchXrayVison Cohen et al. (2022)) into the training loop, selecting TorchXrayVision due to its training on several large chest X-ray datasets and its high accuracy in disease classification tasks. The results indicate that the inclusion of expert input significantly enhance the clinical relevance of the generated images. For MIMIC dataset, when expert knowledge supervision was integrated, the probability values for pneumonia (P(Pneumonia)) showed a marked improvement in alignment with clinical expectations, increasing from 0.6640 to 0.8858 when applying $do$ operation on pneumonia and adjusting more realistically from 0.6315 to 0.4248 when applying $do$ operation on no disease. For ChestXray8 dataset, when expert knowledge supervision was integrated, the probability values for effusion (P(Effusion)) showed a marked improvement in alignment with clinical expectations, increasing from 0.5865 to 0.7183 when applying $do$ operation on effusion and adjusting more realistically from 0.2153 to 0.0889 when applying $do$ operation on no disease. This suggests that expert model supervised counterfactuals provided a more accurate and clinically meaningful representation, which is crucial for reliable diagnostic decision-making. The visualisation of incorporating expert knowledge into the generation can be found in Figures 6 and 7 in the appendix.

As shown in table 1, with the incorporation of synthetic data selected through active learning, the model's performance improves as additional synthetic data were introduced. Figure (4 a) presents the forgetting rate distributions across various data configurations, including raw data alone and with incremental additions of synthetic data (10%-30%). All distributions exhibited a pronounced peak in the low forgetting rate range, indicating that most samples were relatively less challenging. As the proportion of synthetic data increased, the peak decreased while the distribution broadened, suggesting enhanced sample diversity and the introduction of more challenging instances. The resulting heavy-tailed distribution, closely resembling the real data distribution, contributed to the improved performance of the SCM. The results in Table (1) demonstrate that the expert model not only enhanced the effectiveness but also composition and reversibility. **MIMIC dataset**: For composition, SSIM increases from $0.9715$ to $0.9879$ while FID decreased from $0.1451$ to $0.0608$ as the proportion of synthetic data increased, indicating enhanced structural consistency. Similarly, for reversibility, SSIM remained high, $0.9761$, and FID decreased from $0.6425$ to $0.2942$, suggesting better reconstruction fidelity. **ChestXray8 dataset**: For composition, SSIM remians from $0.9999$ while FID decreased from $4.5e-10$ to $4.3e-10$ as the proportion of synthetic data increased, indicating enhanced structural consistency. Similarly, for reversibility, SSIM remained high, $0.9813$, and FID increased from $0.1242$ to $0.3511$, suggesting that reconstruction fidelity, while structurally robust, experiences a slight distributional divergence. The observed increase in FID indicates that the in-

clusion of synthetic data introduces subtle statistical discrepancies. While the pixel-wise structural fidelity remains intact, as evidenced by the sustained SSIM, the feature distribution of the reconstructed images diverges slightly from the original real-world manifold, thereby adversely affecting perceptual quality metrics.

Expert evaluation aims to evaluate the clinical utility of counterfactual X-rays generated by our proposed model, focussing on their impact on diagnostic accuracy among medical professionals. The results indicate that our model offers distinct advantages in medical decision-making processes compared to baseline approaches. In Block 1, while overall diagnostic performance reflected the difficulty of the task, participants achieved higher accuracy using the proposed model compared to the baseline, particularly for pleural effusion and healthy cases. The consistent numerical trend suggests the proposed model better support clinical judgments. Additionally, X-rays generated by the proposed model were rated as significantly more realistic than those generated by the baseline model, indicating higher visual plausibility. Notably, participants found it difficult to distinguish our synthetic images from real X-rays, as evidenced by the realism rating. This confirms that the proposed model produces synthetic data with a high degree of fidelity, effectively mimicking authentic medical imaging. In Block 2, the provision of counterfactual X-rays successfully facilitated diagnostic revision following initial errors. Our proposed model demonstrated robust and consistent improvements across participants, outperforming the baseline model, particularly when controlling for outlier cases. Crucially, our proposed model imporved confidence-accuracy calibration, whereas the baseline model did not, especially in pneumonia cases, where the correlation between confidence and correctness reached statistical significance. Collectively, these findings demonstrate that high-quality counterfactuals generated by our model effectively support clinical reasoning, aiding both diagnostic correction and the alignment of clinician confidence with accuracy.

## 5. Conclusion

In this paper, we proposed a clinically meaningful counterfactual generation model capable of effectively intervening across various lung diseases. By introducing active learning with uncertainty measurement, we refined the distribution of training samples, demonstrating that a heavy-tailed data distribution, maintained through the selection of core sample sets, can significantly enhance the structural causal model. Additionally we incorporated expert model into the framework to ensure that generated counterfactual images retain clinical relevance. Our evaluation method not only assesses the effectiveness, composition and reversibility, but also leverages the expert model to calculate disease probabilities for the generated counterfactual images, providing a more comprehensive assessment by comparing these probabilities to those derived from the original data. Our model yielded effectiveness improvement of $9.25\%$ for MIMIC Dataset and $13.40\%$ for ChestXray8 dataset, respectively. Finally, the results were evaluated by human experts. We demonstrated that the proposed counterfactual model generates realistic and clinically meaningful synthetic chest X-ray images. Crucially, the use of these high-quality counterfactuals successfully aids in diagnostic error correction and enhances confidence-accuracy calibration, thereby offering a reliable mechanism to support and refine clinical judgments.

## Acknowledgments

This work was supported by the Engineering and Physical Sciences Research Council (EPSRC) Grant (EP/X029778/1).

## References

Rsna pneumonia detection challenge. Radiological Society of North America (RSNA), AI Image Challenge, 2018. URL https://www.kaggle.com/c/rsna-pneumonia-detection-challenge. Accessed: 2024-10-28.

Mona Hmoud Al-Sheikh, Omran Al Dandan, Ahmad Sami Al-Shamayleh, Hamid A Jalab, and Rabha W Ibrahim. Multi-class deep learning architecture for classifying lung diseases from chest x-ray and ct images. *Scientific Reports*, 13(1):19373, 2023.

Jason Ansel, Edward Yang, Horace He, Natalia Gimelshein, Animesh Jain, Michael Voznesensky, Bin Bao, Peter Bell, David Berard, Evgeni Burovski, Geeta Chauhan, Anjali Chourdia, Will Constable, Alban Desmaison, Zachary DeVito, Elias Ellison, Will Feng, Jiong Gong, Michael Gschwind, Brian Hirsh, Sherlock Huang, Kshiteej Kalambarkar, Laurent Kirsch, Michael Lazos, Mario Lezcano, Yanbo Liang, Jason Liang, Yinghai Lu, CK Luk, Bert Maher, Yunjie Pan, Christian Puhrsch, Matthias Reso, Mark Saroufim, Marcos Yukio Siraichi, Helen Suk, Michael Suo, Phil Tillet, Eikan Wang, Xiaodong Wang, William Wen, Shunting Zhang, Xu Zhao, Keren Zhou, Richard Zou, Ajit Mathews, Gregory Chanan, Peng Wu, and Soumith Chintala. PyTorch 2: Faster Machine Learning Through Dynamic Python Bytecode Transformation and Graph Compilation. In *Proceedings of ACM International Conference on Architectural Support for Programming Languages and Oauthor = MetZen, Jan and Lemaitre, Guillaume, title = Gaussian process regression (GPR) on Mauna Loa CO2 data, note = https://scikit-learn.org/stable/auto_examples/gaussian_process/plot_gpr_co2.html,*. ACM, April 2024. doi: 10.1145/3620665.3640366. URL https://pytorch.org/assets/pytorch2-2.pdf.

David Barber and Felix Agakov. The im algorithm: a variational approach to information maximization. *Advances in neural information processing systems*, 16(320):201, 2004.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

Eli Bingham, Jonathan P. Chen, Martin Jankowiak, Fritz Obermeyer, Neeraj Pradhan, Theofanis Karaletsos, Rohit Singh, Paul A. Szerlip, Paul Horsfall, and Noah D. Goodman. Pyro: Deep universal probabilistic programming. *J. Mach. Learn. Res.*, 20:28:1–28:6, 2019. URL http://jmlr.org/papers/v20/18-403.html.

Daniel C Castro, Ian Walker, and Ben Glocker. Causality matters in medical imaging. *Nature Communications*, 11(1):3673, 2020.

Hyundong Choi and Heechul Jung. Deep generative models for health. In *Proceedings of Annual Conference on Neural Information Processing Systems*, 2023.

Joseph Paul Cohen, Joseph D. Viviano, Paul Bertin, Paul Morrison, Parsa Torabian, Matteo Guarrera, Matthew P Lungren, Akshay Chaudhari, Rupert Brooks, Mohammad Hashir, and Hadrien Bertrand. Torchxrayvision: A library of chest x-ray datasets and models. In Ender Konukoglu, Bjoern Menze, Archana Venkataraman, Christian Baumgartner, Qi Dou, and Shadi Albarqouni, editors, *Proceedings of The 5th International Conference on Medical Imaging with Deep Learning*, volume 172 of *Proceedings of Machine Learning Research*, pages 231–249. PMLR, 06–08 Jul 2022. URL https://proceedings.mlr.press/v172/cohen22a.html.

Cody Coleman, Christopher Yeh, Stephen Mussmann, Baharan Mirzasoleiman, Peter Bailis, Percy Liang, Jure Leskovec, and Matei Zaharia. Selection via proxy: Efficient data selection for deep learning. In *Proceedings of International Conference on Learning Representations*, 2020.

Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D Hoffman, et al. Underspecification presents challenges for credibility in modern machine learning. *Journal of Machine Learning Research*, 23(226):1–61, 2022.

Elizabeth DeVos and Lisa Jacobson. Approach to adult patients with acute dyspnea. *Emergency Medicine Clinics*, 34(1):129–149, 2016.

Ben Glocker, Charles Jones, Mélanie Bernhardt, and Stefan Winzeck. Algorithmic encoding of protected characteristics in chest x-ray disease detection models. *EBioMedicine*, 89, 2023.

Aya Hage Chehade, Nassib Abdallah, Jean-Marie Marion, Mathieu Hatt, Mohamad Oueidat, and Pierre Chauvet. Advancing chest x-ray diagnostics: A novel cyclegan-based preprocessing approach for enhanced lung disease classification in chestx-ray14. *Computer Methods and Programs in Biomedicine*, 259:108518, 2025. ISSN 0169-2607. doi: https://doi.org/10.1016/j.cmpb.2024.108518. URL https://www.sciencedirect.com/science/article/pii/S016926072400511X.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Guido W. Imbens and Donald B. Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, 2015.

Jeremy A. Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn L. Ball, Katie S. Shpanskaya, Jayne Seekins, David Andrew Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, C. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and A. Ng. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *AAAI Conference on Artificial Intelligence*, 2019. URL https://api.semanticscholar.org/CorpusID:58981871.

M Jasmine Pemeena Priyadarsini, Ketan Kotecha, GK Rajini, K Hariharan, K Utkarsh Raj, K Bhargav Ram, V Indragandhi, V Subramaniyaswamy, and Sharnil Pandya. Lung diseases detection using various deep learning algorithms. *Journal of healthcare engineering*, 2023(1):3563696, 2023.

Karsten Jensen and Steen Andreassen. Generic causal probabilistic networks: A solution to a problem of transferability in medical decision support. *Computer Methods and Programs in Biomedicine*, 89(2):189–201, 2008. ISSN 0169-2607. doi: https://doi.org/10.1016/j.cmpb.2007.10.015. URL https://www.sciencedirect.com/science/article/pii/S0169260707002635. The 6th IFAC Symposium on Modelling and Control in Biomedical Systems.

Alistair Johnson, Tom Pollard, Seth Berkowitz, Nathaniel Greenbaum, Matthew Lungren, Chih-ying Deng, Roger Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6:317, 12 2019. doi: 10.1038/s41597-019-0322-0.

Durk P Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling. Improved variational inference with inverse autoregressive flow. *Advances in neural information processing systems*, 29, 2016.

Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. *Advances in neural information processing systems*, 30, 2017.

Zongyu Li, Zheng Hua Zhu, Xiaoning Guo, Shuai Zheng, Zhenyu Guo, Siwei Qiang, and Yao Zhao. A survey of deep causal models and their industrial applications. *Artif. Intell. Rev.*, 57:298, 2022. URL https://api.semanticscholar.org/CorpusID:253523500.

Miguel Monteiro, Fabio De Sousa Ribeiro, Nick Pawlowski, Daniel C. Castro, and Ben Glocker. Measuring axiomatic soundness of counterfactual image models. In *the Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=lZOUQQvwI3q.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744, 2022.

Nick Pawlowski, Daniel Coelho de Castro, and Ben Glocker. Deep structural causal models for tractable counterfactual inference. *Advances in neural information processing systems*, 33:857–869, 2020.

Judea Pearl. *Causality*. Cambridge university press, 2009.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

AB Rao and D Gray. Breathlessness in hospitalised adult patients. *Postgraduate medical journal*, 79(938):681–685, 2003.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *Proceedings of International conference on machine learning*, pages 1278–1286. PMLR, 2014.

Fabio De Sousa Ribeiro, Tian Xia, Miguel Monteiro, Nick Pawlowski, and Ben Glocker. High fidelity image counterfactuals with probabilistic causal models. In *Proceedings of International Conference on Machine Learning*, 2023.

Bernhard Schölkopf. *Causality for Machine Learning*, page 765–804. Association for Computing Machinery, New York, NY, USA, 1 edition, 2022. ISBN 9781450395861. URL https://doi.org/10.1145/3501714.3501755.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

Casper Kaae Sønderby, Tapani Raiko, Lars Maaløe, Søren Kaae Sønderby, and Ole Winther. Ladder variational autoencoders. *Advances in neural information processing systems*, 29, 2016.

Shenghuan Sun, Greg Goldgof, Atul Butte, and Ahmed M Alaa. Aligning synthetic medical images with clinical knowledge using human feedback. *Advances in Neural Information Processing Systems*, 36, 2024.

Mariya Toneva, Alessandro Sordoni, Remi Tachet des Combes, Adam Trischler, Yoshua Bengio, and Geoffrey J. Gordon. An empirical study of example forgetting during deep neural network learning. In *Proceedings of International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=BJlxm30cKm.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2097–2106, 2017a.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers. ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases. In *Proceedings of Conference on Computer Vision and Pattern Recognition*, 2017b.

## Appendix A. Preliminary Probabilistic Causal Model Framework

A SCM, denoted as $\mathcal{M}$, consists of two variable sets, the endogenous variables $\mathbf{X} = x_1, x_2, \cdots, x_N$ and the exogenous variables $\mathbf{U} = u_1, u_2, \cdots, u_N$, along with a set of causal mechanisms $\mathbf{F} = f_1, f_2, \cdots, f_N$ Pearl (2009). In an SCM, each endogenous variable $x_k \in \mathbf{X}$ is defined by a function $f_k$, depending on its direct causes $\mathbf{pa}_k$ and an exogenous noise term $u_k \in \mathbf{U}$, expressed as $x_k := f_k(\mathbf{pa}_k, u_k)$. Graphically represented as a directed acyclic graph (DAG), an SCM satisfies the causal Markov condition, meaning each variable is conditionally independent of its non-descendants given its direct causes Ribeiro et al. (2023). SCMs also enable counterfactual reasoning, allowing questions about hypothetical outcomes under alternative conditions. Counterfactuals are calculated through a three-step process Pearl (2009): 1) Abduction—updating the exogenous variable distribution $P(\mathbf{U})$ based on observed evidence to infer $P(\mathbf{U}|\mathbf{X})$; 2) Action—applying an

intervention $do(x_k := c)$ to generate the submodel $\mathcal{M}_c$ and 3) Prediction—estimating the counterfactual distribution using $< \mathcal{M}_c, P(\mathbf{U}|\mathbf{X}) >$. This structured approach enables SCMs to model and answer complex causal and counterfactual questions effectively.

The deep structural causal model is based on a hierarchical variational autoencoder (HVAE) Kingma et al. (2016); Sønderby et al. (2016), and the HVAE is extended from a standard Variational Autoencoder (VAE) Rezende et al. (2014) by applying hierarchical latent variable model (HLVM). A HVAE model for data $x$ is defined by a hierarchical latent mediator model using a prior over $L$ layers of hierarchical latent variables $\mathbf{z}_i$, where $i = 1, \cdots, L$, and it can be factorizing as:

$$p(\mathbf{x}, \mathbf{z}_{1:L}) = p(\mathbf{x}|\mathbf{z}_{1:L})p(\mathbf{z}_L) \prod_{i=1}^{L-1} p(\mathbf{z}_i|\mathbf{z}_{>i}). \tag{3}$$

HVAEs train the hierarchical generative model $p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z}_{1:L})$, where $\boldsymbol{\theta}$ is the parameters of the model, by demonstrating variational inference model $q_{\boldsymbol{\phi}}(\mathbf{z}_{1:L}|\mathbf{x})$ and maximizing the evidence lower bound (ELBO) on the marginal log-likelihood:

$$\log p_{\boldsymbol{\theta}}(\mathbf{x}) \geq \mathbb{E}_{q_{\boldsymbol{\theta}}(\mathbf{z}_{1:L}|\mathbf{x})}[\log p_{\boldsymbol{\phi}}(\mathbf{x}|\mathbf{z}_{1:L})] \tag{4}$$

$$- D_{KL}(q_{\boldsymbol{\phi}}(\mathbf{z}_{1:L}|\mathbf{x}) \parallel p_{\boldsymbol{\theta}}(\mathbf{z}_{1:L})), \tag{5}$$

where $D_{KL}$ is Kullback-Leibler Divergence. By optimizing the ELBO via trainable parameters $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$, the marginal $p_{\boldsymbol{\theta}}(\mathbf{x})$ is close to a given data distribution $p_{data}(\mathbf{x})$. In Ribeiro et al. (2023), the deep structure causal models (DSCMs) treat $\mathbf{z}$ as part of $\mathbf{x}$'s exogenous noise, requiring an unconditional prior $p(\mathbf{z})$. Using a conditional HVAE framework decouples the conditioning on $\mathbf{pa_x}$ in the prior while enabling conditional sampling, the generative model is:

$$p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z}_{1:L}|\mathbf{pa_x}) = p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z}_{1:L}, \mathbf{pa_x})p_{\boldsymbol{\theta}}(\mathbf{z}_L) \prod_{i=1}^{L-1} p_{\boldsymbol{\theta}}(\mathbf{z}_i|\mathbf{z}_{>1}), \tag{6}$$

where $\mathbf{z}_i$ and $\mathbf{pa_x}$ are introduced into each layer of the top-down hierarchy as:$h_i = h_{i+1} + f_i^w(\mathbf{z}_i, \mathbf{pa_x}), \mathbf{z}_i \sim p_{\boldsymbol{\theta}}(\mathbf{z}_i|\mathbf{z}_{>i})$, where $h_i$ is the hierarchy latent parameter. With this conditioning architecture, the prior $p_{\boldsymbol{\theta}}(\mathbf{z}_{1:L})$ becomes independent of $\mathbf{pa_x}$, but the likelihood is: $p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z}_{1:L}, \mathbf{pa_x}) = \mathcal{N}(\mathbf{x}|\mu_{\boldsymbol{\theta}}(\mathbf{h}_1), \sigma_{\boldsymbol{\theta}}(\mathbf{h}_1))$, where $h_1$ is the initial hierarchy mechanism, and $u_{\boldsymbol{\theta}}$ and $\sigma_{\boldsymbol{\theta}}$ are mean and variance respectively. Here, $\mathcal{N}(\mathbf{x}|\cdot, \cdot)$denotes a Gaussian distribution over $\mathbf{x}$.

The counterfactual $\tilde{\mathbf{x}}$ should adhere to counterfactual conditioning on $\tilde{\mathbf{pa}}_k$ by exhibiting semantically meaningful changes from $\mathbf{x}$. As the mutual information (MI) term is intractable, a variational technique Barber and Agakov (2004) is applied to define the lower bound of the MI,

$$I(\tilde{pa}_k; \tilde{x}) = \mathbb{E}_{p(\tilde{\mathbf{pa}}_k, \tilde{\mathbf{x}})} \left[ \log \left( \frac{p(\tilde{\mathbf{pa}}_k|\tilde{\mathbf{x}})}{p(\mathbf{pa}_k)} \cdot \frac{q_\psi(\tilde{\mathbf{pa}}_k|\tilde{\mathbf{x}})}{q_\psi(\tilde{\mathbf{pa}}_k|\tilde{\mathbf{x}})} \right) \right] \geq \mathbb{E}_{p(\tilde{\mathbf{x}})} \left[ \log q_\psi(\tilde{\mathbf{pa}}_k|\tilde{\mathbf{x}}) \right] + H(\tilde{\mathbf{pa}}_k) \tag{7}$$

where $q_\psi(\tilde{\mathbf{pa}}_k|\tilde{\mathbf{x}})$ is the approximated variational distribution for $p(\tilde{\mathbf{pa}}_k|\tilde{\mathbf{x}})$. In practice the random interventions on $\mathbf{pa_x}$ are performed by independently sampling each parent from its marginal distribution, and maximise the log-likelihood of the predictors with a given $\tilde{\mathbf{x}}$ sampled from the counterfactual distribution Ribeiro et al. (2023), $\max_{P_\mathcal{M}, q_\psi} \mathbb{E}_{p_{\text{data}}(\mathbf{x}, \mathbf{pa}_x)} \left[ -\mathcal{L}_{\text{CT}}(\mathcal{M}; \mathbf{x}, \mathbf{pa}_x) \right]$, where the counterfactual loss is, $\mathcal{L}_{\text{CT}}(\mathcal{M}; \mathbf{x}, \mathbf{pa}_x) =$

$$- \sum_{k=1}^{K} \mathop{\mathbb{E}}_{\substack{\tilde{\mathbf{pa}}_k \sim p(\mathbf{pa}_k), \\ \tilde{\mathbf{x}} \sim P_\mathcal{M}(\tilde{\mathbf{x}}|do(\tilde{\mathbf{pa}}_k), \mathbf{x})}} \left[ \log q_{\psi_k}(\tilde{\mathbf{pa}}_k|\tilde{\mathbf{x}}) \right], \tag{8}$$

**Algorithm 1:** Forgetting Events (Toneva et al., 2019; Coleman et al., 2020)

**Initialize:** $\text{acc}_i^{(t-1)} \leftarrow 0, \quad \forall i \in [n]$ **Initialize:** $\text{forget}_i \leftarrow 0, \quad \forall i \in [n]$

**while** *training is not done* **do**

  Sample mini-batch $B$ from dataset $\mathcal{D}$ **for** *each example* $i \in B$ **do**

    Compute current accuracy $\text{acc}_i^{(t)}$ **if** $acc_i^{(t-1)} > acc_i^{(t)}$ **then**

      $\text{forget}_i \leftarrow \text{forget}_i + 1$

    **end**

    $\text{acc}_i^{(t-1)} \leftarrow \text{acc}_i^{(t)}$

  **end**

  Perform gradient update on classifier using $B$

**end**

**return** $\{\text{forget}_i\}_{i=1}^n$

and $q_\psi$ are parent predictors. The constraint of counterfactual training, denoted as $c$, negative ELBO (free energy $\mathcal{F}_{\text{FE}}$) of the pre-trained HVAE over the observational data, ensuring it does not increase during counterfactual training. In Lagrangain form, the optimisation problem is reformulated as minimising $\mathcal{L}_{\text{LG}}(\boldsymbol{\theta}, \boldsymbol{\phi}, \lambda; \mathbf{x}, \mathbf{pa}_x)$

$$= \mathcal{L}_{\text{CT}}(\mathcal{M}; \mathbf{x}, \mathbf{pa}_x) - \lambda \left(c - \mathcal{F}_{\text{FE}}(\boldsymbol{\theta}, \boldsymbol{\phi}; \mathbf{x}, \mathbf{pa}_x)\right), \tag{9}$$

where $\boldsymbol{\phi}$ is the trainable parameter.

## Appendix B. Setup and Implementation Details

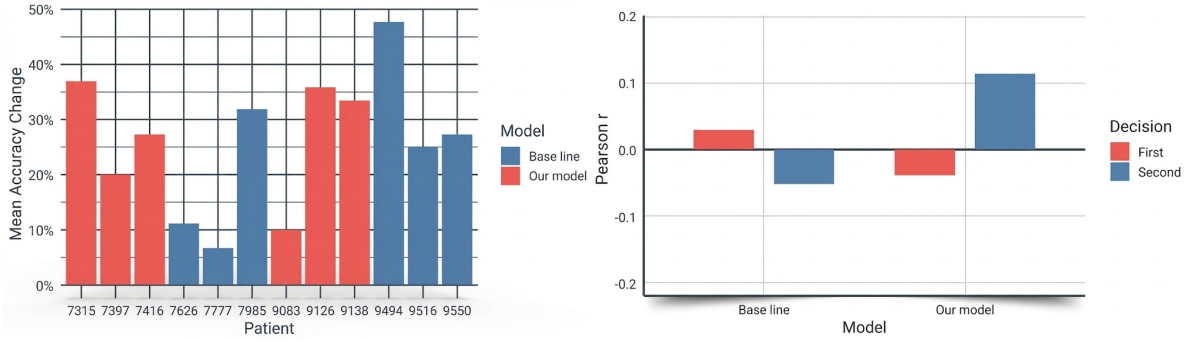
The stochastic latent variables in our HVAE spanned 5 resolution scales, reaching up to half the input resolution: $1^2, 6^2, 12^2, 24^2, 48^2, 96^2$. Each resolution scale incorporated the following number of residual blocks: $2, 4, 8, 12, 8, 4$. Each latent variable had 16 channels, and the feature map widths at each resolution scale were $512, 192, 160, 128, 96, 64, 32$, where 32 represented the width of the final (deterministic) $192 \times 192$ upsampling residual block. We trained our HVAEs 200 epoches with batch size of 32 and the AdamW optimizer. The initial learning rate was $5e^{-5}$ with 100 iterations of linear warm-up, $[\beta_1, \beta_2] = [0.9, 0.9]$ in Equation 2 and $0.1$ weight decay. The gradient clipping was set to 350 and gradient update skipping threshold was 500 based on the L2 norm of the gradients. The final model was an exponential moving average of the parameters with a rate of $0.999$. In the aspect of data augmentation, zero-padding of 9 to all borders was applied and the random horizontal flips with probability $0.5$ was performed. The pixel intensities were normalised to a range of $[-1, 1]$. Since node D (disease) was not a continuous variable, normalizing flows can not be directly invertible. To address this problem, the Gumbel-max parametrisation was applied Pawlowski et al. (2020). The Gumbel-max trick was a method to draw a sample for discrete distribution, given its probabilities over categories Ribeiro et al. (2023).

All experiments were performed on a Linux machine with i9-14900K, NVIDIA RTX 4090 and 128 GB memory. We implemented HVAEs based on Pytorch Ansel et al. (2024) and Pyro Bingham et al. (2019), a universal probabilistic programming language supported by Pytorch.

## Appendix C. Participants

The study sample consisted of 42 participants, aged between 20 and 67 years ($\mu = 26.1$, $\sigma = 7.37$, 24 Female, 18 Male). Participants were recruited via the online platform Prolific and were directed to the online experiment platform Gorilla (https://app.gorilla.sc/). The recruitment period was the 2nd - 5th June 2025 (02/06/2025 - 05/06/2025). All participants were either medical students (26) or medical doctors (16), with a combined average of 6.20 years medical experience (including study time). The study was approved by the University College London Institutional Review Board (IRB): 0487. They were provided with an information sheet outlining the study, including the nature of the task, consent procedures, and data usage. Formal informed consent was obtained before participation. All participants had the opportunity to ask questions or seek clarification via the email address provided on the participant information sheet or via the researcher-participant chat on Prolific (the online platform used to recruit). All consent records were automatically logged and timestamped by the Gorilla platform, ensuring secure documentation. The study did not involve minors; all participants confirmed they were 18 years of age or older before proceeding. Therefore, parental or guardian consent was not required.

## Appendix D. Extended Results




(a) (b)

Figure 5: a) Mean accuracy change between first and second decision by Patient and Model. b) Pearson correlation of accuracy and confidence for first and second decision for each counterfactual model.

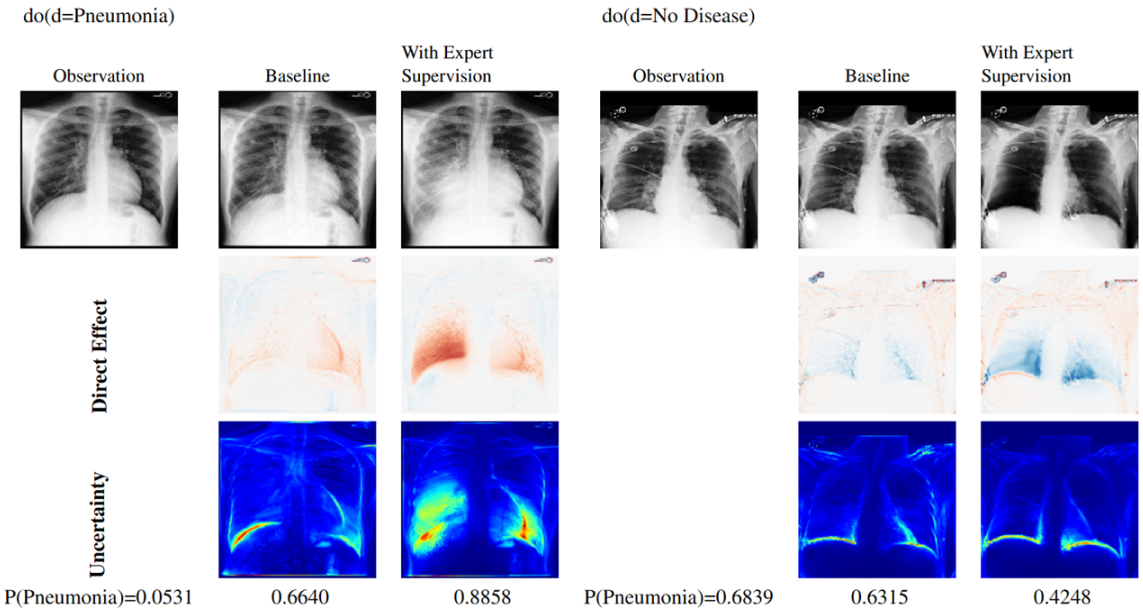



Figure 6: Examples of counterfactual for MIMIC dataset. The probabilities of disease generated by both baseline model, and our model are presented. The uncertainty is standard deviation at each pixel.

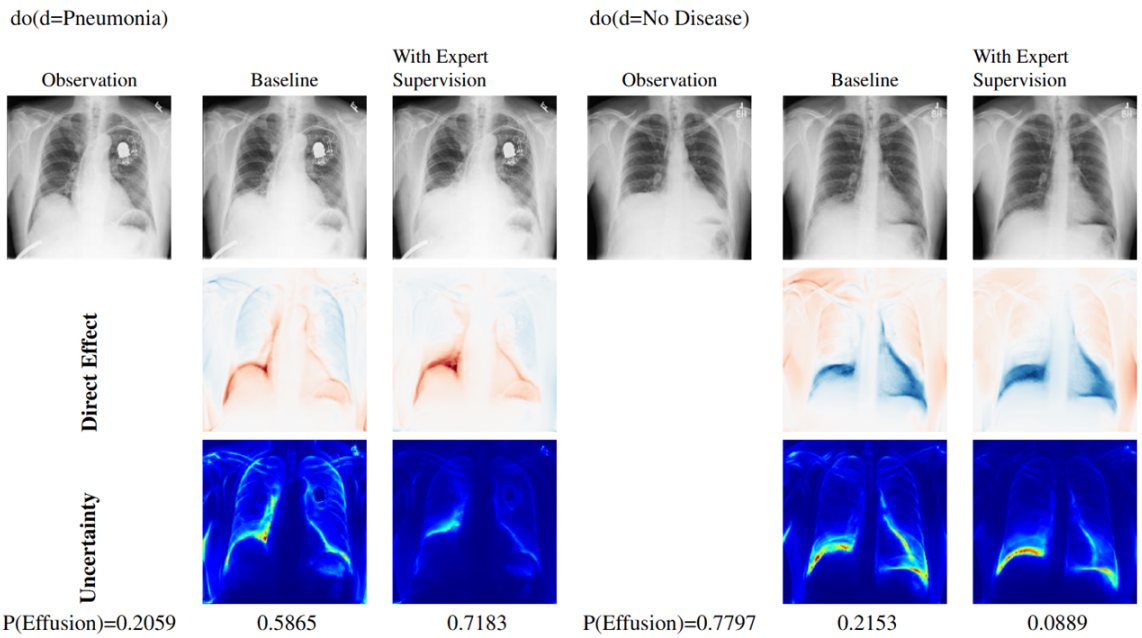



Figure 7: Examples of counterfactual for ChestXray8 dataset. The probabilities of disease generated by both baseline model, and our model are presented. The uncertainty is standard deviation at each pixel.