Computers in Biology and Medicine 213 (2026) 111807

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Causal counterfactual simulation for treatment decisions in multimodal lung disease data

Yifei Zhu[a,*], Lei Zhang[a], Chris Sainsbury[b,e], Feng Dong[c], John MacLay[e,f], David J. Lowe[d,g], Xujiong Ye[a]

[a] Department of Computer Science, University of Exeter, Exeter, UK
[b] Institute of Applied Health Research, University of Birmingham, Birmingham, UK
[c] Department of Computer and Information Sciences, University of Strathclyde, Glasgow, UK
[d] Digital Health Validation Lab, University of Glasgow, Glasgow, UK
[e] NHS Greater Glasgow and Clyde, Glasgow, UK
[f] Respiratory Department, Glasgow Royal Infirmary, Glasgow, UK
[g] Emergency Department, Queen Elizabeth University Hospital, Glasgow, UK

HIGHLIGHTS

- Clinician-guided longitudinal DAG models treatment-specific outcomes.
- Anatomy-aware longitudinal model uses images and lung-mask attention to track disease evolution.
- Semi-synthetic MIMIC data removes confounding, resolves the treatment paradox, and recovers true effects.

ARTICLE INFO

Keywords:
Counterfactual
Structure causal model
Medical imaging
Generative model

ABSTRACT

While deep learning has advanced the automated diagnosis of pulmonary diseases from chest X-rays, transitioning these models to treatment optimisation is fundamentally hindered by the reliance on biased observational data. In datasets, such as MIMIC, confounding factors create a "Treatment Paradox", where active interventions are statistically correlated with higher mortality simply because they are preferentially assigned to the most severely ill patients. Consequently, standard models learn clinical assignment patterns rather than true physiological efficacy. To address this, we propose a causal synthesis strategy that generates de-confounded, high-dimensional treatment counterfactuals. Guided by our experts' clinical priors, we construct a clinical transition matrix to simulate genuine physiological treatment effects. Using patient white blood cells as measure of severity, we applied a disease counterfactual model to generate the corresponding post-treatment images based on this matrix. Our longitudinal generative architecture uses both pre- and post-treatment chest X-rays to capture disease evolution over time. Furthermore, to resolve spatial misalignments caused by varying patient postures across temporal scans, we introduce an explicit lung mask constraint. This anatomy-aware mechanism confines counterfactual modifications to pulmonary regions, preventing clinically implausible alterations to background anatomy. By decoupling true treatment effects from baseline severity and posture variations, our framework successfully resolves the treatment paradox. Ultimately, this approach generates realistic treatment counterfactuals, providing clinicians with a trustworthy visual tool to evaluate patient-specific outcomes under alternative treatment scenarios and support medical decision-making.

* Corresponding author.
Email address: y.zhu4@exeter.ac.uk (Y. Zhu).

https://doi.org/10.1016/j.compbiomed.2026.111807
Received 18 March 2026; Received in revised form 21 May 2026; Accepted 8 June 2026
Available online 11 June 2026
0010-4825/© 2026 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

1. Introduction

Breathlessness is common in hospitalised patients and is often a harbinger of severe and potentially fatal pathologies [34]. Patients with undifferentiated respiratory distress require urgent attention in the emergency department [12]. With lung diseases affecting millions globally and lung cancer maintaining high incidence-to-mortality ratios [9], 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 [2,11,17,22]. A growing body of research has validated the efficacy of deep learning 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 [1,23,41].

While these diagnostic advancements are significant, the transition from disease classification to treatment optimisation remains fraught. AI in healthcare relies heavily on "Big Data", such as Medical Information Mart for Intensive Care (MIMIC) dataset [1,23,41]. However, these datasets are primarily observational, reflecting clinical workflows and human decision-making rather than controlled experimental conditions. Consequently, data such as MIMIC are sourced exclusively from intensive care units (ICUs) and are therefore subject to a specific and severe form of selection bias. In the ICU environment, pulmonary pathologies do not exist in isolation, but they are confounded by multi-organ failure, systemic inflammation, and the physiological noise of critical care interventions [3,4,13]. Consequently, the treatment outcomes recorded in such datasets may differ substantially from those typically observed in broader clinical practice or general practice (GP) settings. This divergence means that the data represent only the most extreme cases, where the patient's baseline health is already severely compromised. If models are trained naively on these distributions, they risk learning spurious correlations that reflect terminal illness or institutional practice rather than physiological truth, leading to a catastrophic lack of generalisability in broader clinical deployment [15,29]. In this work, we propose a semi-synthetic data generation strategy to overcome this issue.

The most critical presentation of selection bias is the treatment paradox, which is a form of Simpson's paradox [38] driven by confounding [32,39]. With these biases, empirical evidence often reveals a counter-intuitive statistical relationship, such as the death rate with treatment being higher than the death rate without treatment. This suggests that patients receiving active intervention have a higher probability of mortality. However, clinicians prioritise aggressive treatments for most critically ill patients who possess an inherently higher baseline risk of death [7], creating a statistical artefact driven by unobserved clinical severity. As the influence of ICU-related factors is often embedded in high-dimensional data, standard models fail to account for them [14], incorrectly attributing the poor outcome to the treatment itself. This phenomenon, where models learn treatment assignment patterns rather than true physiological responses, has been identified as a primary barrier to the clinical deployment of risk models [14]. Specifically, instead of capturing the physiological efficacy of an intervention, the model merely learns the decision-making logic used by doctors to assign the treatment. Consequently, the model could recommend against clinically beneficial treatments, falsely predicting that the intervention itself drives clinical deterioration.

To address these challenges, we apply a shift from traditional data cleaning to a causal synthesis framework guided by clinical priors. We use a previously developed counterfactual model for lung disease [45], which is further augmented by expert clinical knowledge to reconstruct the underlying data generation process. Specifically, we formalise expert clinical knowledge into a clinical transition matrix to simulate plausible patient trajectories. We then use a pre-trained generative backbone to generate synthetic X-rays based on the real-world dataset, ensuring that the outcomes are explicitly defined and controllable [46]. A core innovation of this work is enabling generative modelling of high-dimensional treatment counterfactuals. Unlike conventional methods restricted to low-dimensional tabular data, we project causal dynamics directly into the visual domain. By synthesising counterfactual Chest X-ray (CXR) images, we visualise the hypothetical progression of lung pathology under differing treatment scenarios (e.g., treated vs. untreated). This process creates "probabilistic twins" for each patient, effectively decoupling the causal impact of the treatment from the confounding effects of baseline severity, thereby empowering clinicians with the causal insights necessary for robust, evidence-based decision-making.

This work presents a novel methodology for causal correction in large-scale multimodal high dimensional clinical data. Our primary contributions are:

- Clinician-Guided Longitudinal Causal Graph: Guided by clinical experts, we construct a longitudinal Directed Acyclic Graph (DAG) to model outcomes under different treatment options.
- Anatomy-Aware Longitudinal Architecture: We propose a novel treatment effect model that leverages pre-treatment images to capture disease evolution and utilises Lung Masks as a spatial attention mechanism to focus learning on relevant pulmonary regions.
- We built a large semi-synthetic dataset to address the bias in the raw MIMIC data. In the original data, treatment often appears harmful simply because the most severely ill patients receive it. Our new dataset removes this confounding pattern by explicitly simulating the true medical efficacy of treatment. By training on this revised data with known outcomes, our method is able to resolve the Treatment Paradox and accurately recover the genuine effects of medical intervention.

2. Related work

While deep learning has achieved expert-level performance in medical diagnostics, standard models rely heavily on statistical correlations, making them vulnerable to spurious correlations and biases that often exist in real-world medical data [31,37]. To address this, structural causal models (SCMs) [33] have been introduced to formalise the data generation process and enable counterfactual reasoning. Recently, deep structural causal models (DSCMs) have emerged to bridge the gap between high-dimensional imaging data and causal inference. For instance, Pawlowski et al. [30] proposed a tractable framework for counterfactual inference using normalising flows, while Castro et al. [8] highlighted the necessity of causal graphs in medical imaging to avoid interpretation errors. Despite these advances, most existing DSCMs focus on static disease detection or simplified causal graphs. They often lack the capability to model complex, longitudinal treatment effects where unobserved confounders create paradoxical associations in real-world datasets like MIMIC-CXR.

Parallel to the developments in causal modelling, deep generative models, such as generative adversarial networks (GANs) [16] and denoising diffusion probabilistic models [20], have been widely applied for medical data augmentation [43]. Techniques such as CycleGAN are frequently used for domain adaptation or cross-modality synthesis. However, these correlation-based generative models lack explicit causal control mechanisms. They typically perform image-to-image translation based on population-level distributions rather than individual-level counterfactuals. As a result, attempts to modify pathological features may introduce anatomically implausible artifacts or unintended identity shifts [10]. To illustrate current limitations, Fig. 1 presents example images generated by a diffusion model for chest X-ray generation [27]. As shown, even diffusion models frequently fail to produce clinically meaningful images and often contain unrealistic artifacts. Crucially, they remain unable to generate personalised counterfactuals. The absence of explicit structural constraints and causal identifiability limits the utility of these models for treatment effect estimation, where preserving the integrity of patient characteristics prior to treatment is critical.

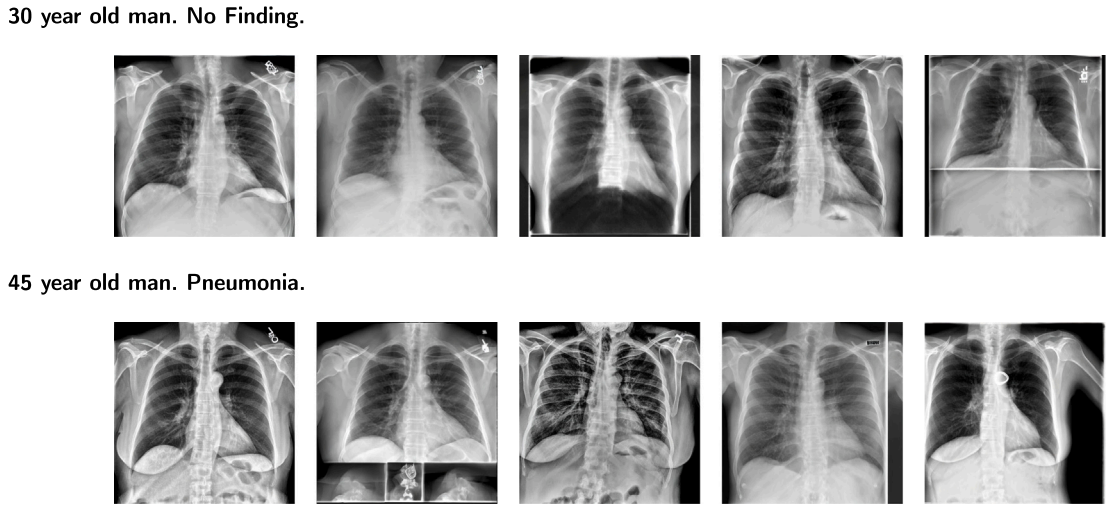


Fig. 1. Examples generated by diffusion model [27]. Diffusion model frequently fails to produce clinically meaningful chest X-ray images.

Furthermore, while counterfactual regression (CFR) models and recent causal representation learning (CRL) methods have significantly advanced the field, their direct application to our task presents substantial challenges. Traditional CFR models are primarily architected to estimate scalar individual treatment effects (ITE) from low-dimensional tabular data, lacking the generative capacity to synthesise high-dimensional post-treatment images. Similarly, while recent CRL methods provide robust theoretical frameworks for latent disentanglement, extending them from standard benchmarks to disentangle complex and longitudinal confounders within high-resolution medical imaging remains a significant computational challenge. Consequently, standard conditional generative models (such as diffusion models) often fail to preserve patient identity, while traditional causal models lack high-dimensional generative capabilities. The proposed HVAE architecture bridges this gap by explicitly embedding longitudinal causal graphs into the generative process.

To address the dual challenges of causal identifiability and high-fidelity generation, our work builds directly upon the probabilistic causal framework recently proposed by Zhu et al. [45] and the architectural advancements in hierarchical variational autoencoders (HVAEs) by Ribeiro et al. [36], which is shown in Fig. 2. In [45], the authors developed a DSCM framework with active learning and an expert model for lung disease. Crucially, the validity of this framework was established through rigorous clinical user studies, where senior respiratory physicians confirmed that the generated counterfactuals exhibited semantically correct pathological changes while maintaining high visual realism. However, the original framework was developed for static diagnostic tasks and did not model the temporal dynamics of treatment interventions or the complex confounding structures, such as Simpson's paradox, that are common in ICU settings. In this work, we first evaluate whether the clinically validated generative backbone can produce images that are consistent with clinical semantics. We then extend this model by incorporating longitudinal constraints and expert-specified confounders to enable the generation of treatment counterfactuals, augmenting the data synthesis for training counterfactual treatment effect models.

3. Methods

3.1. Preliminary lung disease counterfactual generation model

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 [32,36]. 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) [24,40], extended from standard VAEs [35]. 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 $p_\theta(\mathbf{z}_{1:L})$ to maintain independence, while injecting them into the generative path to parametrise the likelihood $p_\theta(\mathbf{x}|\mathbf{z}_{1:L}, \mathbf{pa}_{\mathbf{x}})$ [36]. 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 [5]. The training is formulated as a constrained optimisation problem using the Lagrangian method:

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

where $\mathcal{L}_{\mathrm{CT}}$ represents the counterfactual loss derived from parent predictors, and $\mathcal{F}_{\mathrm{FE}}$ is the observational free energy (negative ELBO), ensuring the model preserves generative quality on observed data c while optimising 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 [36,45].

3.2. Clinician-guided longitudinal deep causal model

To formulate the causal problem, we first define the standard longitudinal clinical workflow for pneumonia management, which serves as the basis for our data modelling. The clinical timeline commences at the baseline admission ($t = 0$), where a patient presenting with respiratory symptoms undergoes an initial diagnostic workup, comprising a baseline chest X-ray (X_{pre}) and a blood test (I_1). Based on these indicators, an initial diagnosis is established, and a therapeutic intervention (T) is administered. The treatment episode spans the period between admission and the scheduled follow-up visit, which typically occurs four to twelve weeks post-intervention ($t = 1$). Crucially, the follow-up protocol follows a sequential assessment: the secondary blood count (I_2) is obtained prior ($t = 2$) to the acquisition of the outcome X-ray (X_{post}).

3.2.1. Longitudinal DAG

Estimating the causal effect of treatment is inherently a longitudinal problem [25]. Existing DSCMs in medical imaging predominantly rely on static snapshots [8,30], a limitation that fails to disentangle pre-existing anatomical conditions from the dynamic effects of treatment interventions. Unlike lung disease counterfactual models that rely on

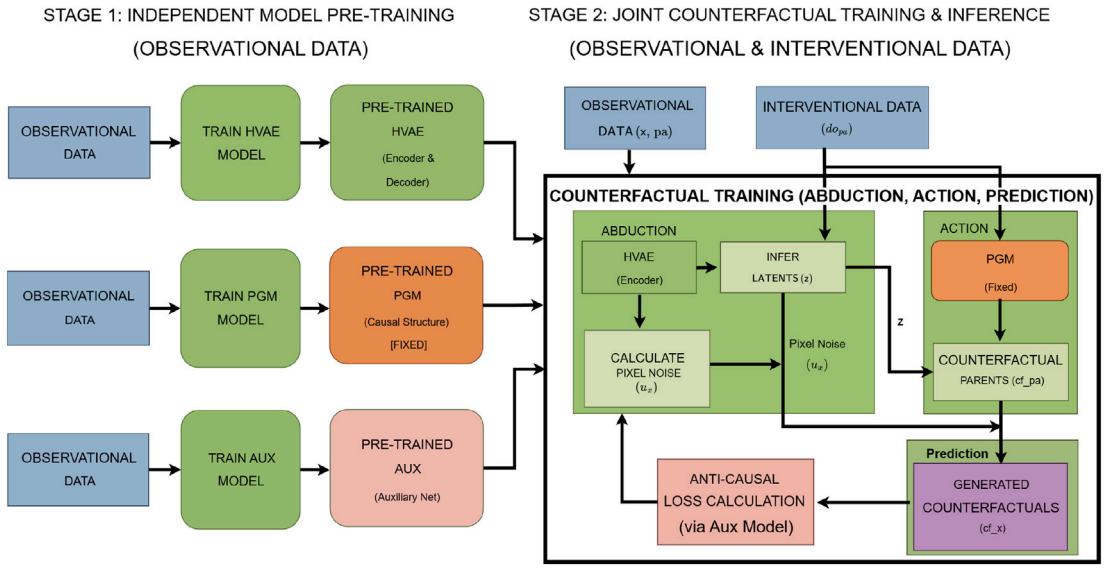


Fig. 2. Overview of the two-stage counterfactual framework. Stage 1 illustrates the independent pre-training of the Hierarchical Variational Autoencoder (HVAE), the probabilistic Graphical model (PGM), and the auxiliary anti-causal predictor using purely observational data. Stage 2 describes the joint counterfactual training and inference process, structured upon the standard three-step counterfactual procedure: (1) Abduction: The HVAE encoder infers the exogenous latent variables (z) and pixel-level noise (u_x) from the factual observational data to capture the patient's invariant anatomical identity. (2) Action: The fixed PGM applies the specific clinical intervention (do_{pa}) to generate the counterfactual parents (cf_pa). (3) Prediction: The generative framework synthesizes the final counterfactual image (cf_x) by integrating the preserved identity features (z, u_x) with the intervened causal parents (cf_pa). Crucially, during the joint training in Stage 2, an anti-causal loss is calculated via the pre-trained auxiliary model on the generated counterfactuals (cf_x). This loss provides essential gradient feedback to the HVAE to guarantee the semantic correctness of the synthesized pathological shifts.

static and single-time-point chest X-rays, assessing treatment efficacy requires modelling the temporal evolution of the patient's state before and after treatment. When evaluating only the post-treatment image, standard generative models cannot determine whether the absence of disease is caused by the treatment or if the patient was always healthy. To address this, we propose a novel longitudinal deep structural causal model (LDSCM) that models the causal mechanisms governing the patient's temporal evolution from a pre-treatment status to a post-treatment outcome. By structuring the causal graph to reflect this time-dependent progression, we enable the model to distinguish between the patient's invariant biological identity and the specific pathological deviations induced by the intervention, thereby providing the necessary temporal context for accurate counterfactual inference. This sequence presents the complete "Treatment Case" that our model is designed to analyse.

3.2.2. Clinical confounders

In clinical practices, treatment allocation is not random but is strongly influenced by the patient's underlying infection severity and physiological condition. If this underlying severity is omitted from the SCM, it acts as an unobserved confounder that simultaneously affects both treatment assignment and the patient outcome, creating a bias that obscures the true effect of the treatment. To address this, we incorporate the white blood cell count (WCC) as an observed severity-related variable within the causal graph [32]. Our modelling choice is supported by established clinical guidelines and severity assessment protocols for respiratory infection, chronic lung disease assessment, and Systemic Inflammatory Response Syndrome (SIRS) [6]. Specifically, the British Thoracic Society (BTS) guideline for bronchiectasis mandates full blood count assessment as a standard investigation to identify aetiology and immune status [19]. Similarly, inflammatory markers are central to the severity assessment in lower respiratory tract infections as outlined in National Institute for Health and Care Excellence (NICE) guidelines [28], where they complement clinical scoring systems to evaluate pneumonia severity and guide intensive care interventions [42].

To formally justify the identifiability of the treatment effect within our SCM, we explicitly analyse the assumed causal graph using Backdoor Criterion. In our clinical setting, the pre-treatment total WCC (I_1) is incorporated as an observed pre-treatment severity indicator that captures important aspects of the patient's underlying biological condition prior to treatment assignment. Within the DAG, I_1 functions as a common cause of both the treatment decision T and the post-treatment outcome $\mathbf{X}_{post}$, thereby inducing non-causal backdoor paths between treatment and outcome. In particular, the following backdoor pathways are present:

$$T \rightarrow I_1 \rightarrow \mathbf{X}_{post} \tag{2}$$

and

$$T \rightarrow I_1 \rightarrow I_2 \rightarrow \mathbf{X}_{post}. \tag{3}$$

According to the Backdoor Criterion, a valid adjustment set must block all non-causal paths entering the treatment node through incoming arrows. Conditioning on the observed pre-treatment variable I_1 blocks both of the above backdoor paths, thereby d-separating T and $\mathbf{X}_{post}$ along these non-causal routes and enabling identification of the causal effect $do(T) \rightarrow \mathbf{X}_{post}$ under the assumed SCM. Importantly, we intentionally condition only on the pre-treatment WCC (I_1) and not on the post-treatment WCC (I_2). Structurally, I_2 is descendant of the treatment variable and lies on the causal mediation pathway,

$$T \rightarrow I_2 \rightarrow \mathbf{X}_{post}, \tag{4}$$

which represents part of the genuine biological mechanism through which treatment affects the outcome. Conditioning on I_2 would therefore block part of the total causal effect that we aim to estimate. Furthermore, I_2 also forms a collider in the structure,

$$I_1 \rightarrow I2 \leftarrow T. \tag{5}$$

Under the rules of d-separation, this collider naturally blocks association between I_1 and T unless conditioning is performed on I_2.

Conditioning on I_2 would open the previously blocked non-causal pathway,

$$T \rightarrow I_2 \leftarrow I_1 \rightarrow \mathbf{X}_{post}, \tag{6}$$

thereby introducing collider bias and inducing a spurious association between treatment and outcome. Consequently, restricting adjustment to the pre-treatment variable I_1 is theoretically justified both by the Backdoor Criterion and by the structural properties of the assumed SCM.

3.2.3. Capturing pathological shift with temporal anatomical anchors

To quantify the visual changes attributable to treatment, we move beyond the single-image reconstruction approaches used in prior disease counterfactual models [36,45] by implementing a dual-image input strategy that explicitly encourages the model to capture pathology evolution over time. First, within the generative HVAE, we use the pre-treatment chest X-ray, $\mathbf{X}_{pre}$, as an exogenous temporal anatomical anchor. We extract high-dimensional features from $\mathbf{X}_{pre}$ and treat them as one of the parent nodes. This constraint preserves invariant anatomical identity, compelling the latent space to focus exclusively on modelling the pathological variations driven by the intervention. Second, within the anti-causal parent predictors, we introduce a feature-difference integration to robustly infer latent confounders. Using a pre-trained expert model as the feature extractor, we extract deep features from both chest X-ray before treatment, $\mathbf{X}_{pre}$, and outcome chest X-ray, $\mathbf{X}_{post}$. A representation vector is constructed by concatenating the post-treatment features, $\mathbf{F}_{post}$, and the element-wise difference of pre- and post-treatment features, $(\mathbf{F}_{post} - \mathbf{F}_{pre})$. By explicitly conditioning the predictor on this difference term, the model is guided to attend specifically to the pathological shift rather than static background noise.

3.2.4. Mitigating inter-scan variability inconsistency via anatomy-aware mask guidance

However, the utilisation of longitudinal image pairs introduces the challenge of inter-scan variability. Since X_{pre} and X_{post} are acquired at different time points, patients inevitably exhibit variations in positioning, respiratory volume, or projection angle. A generative model trained on such pairs without constraints risks entangling these non-pathological geometric shifts, potentially leading to "hallucinated" artifacts where the ribcage or spine is physically distorted to match the target pose [10]. To address this, we incorporate an anatomy-aware lung mask constraint. We use the PSPNet architecture [44] to extract fine-grained anatomical priors. Crucially, rather than applying a threshold binary mask, we utilise the raw pixel-wise probability maps ($\mathbf{P}_{lung} \in [0,1]$) representing the likelihood of lung tissue presence at every spatial location. These probabilistic maps serve as a fundamental component of the input data. The mask is channel-wise concatenated to the corresponding raw chest X-ray, resulting in a dual-channel tensor $\mathbf{X} \in \mathbb{R}^{H \times W \times 2}$. By feeding this anatomy-enhanced input into the HVAE, we provide the encoder with an explicit spatial attention mechanism from the very first layer. This structurally forces the model to distinguish between the pulmonary parenchyma ($P_{lung} \approx 1$) and surrounding thoracic structures ($P_{lung} \approx 0$), ensuring that the learned latent features are strictly confined to the relevant pathological regions while preserving the patient's rigid anatomy. The workflow of the proposed framework is presented in Fig. 3.

3.3. Evaluation matrix

A reliable counterfactual inference model must satisfy three fundamental constraints: effectiveness, composition, and reversibility [26]. First, effectiveness ensures that intervening on a variable guarantees it takes the specified value, denoted as $Pa(DSCM(\mathbf{x}, \mathbf{pa}_{\mathbf{x}}, \tilde{\mathbf{pa}}_{\mathbf{x}})) = \tilde{\mathbf{pa}}_{\mathbf{x}}$, where $Pa(\cdot)$ is a prediction function outputting parent variables. The effectiveness of the model is quantified using anti-causal parent predictors [36]. Second, composition requires that intervening on a variable without changing its value does not affect the rest of the variables, which satisfies the equation $DSCM(\mathbf{x}, \mathbf{pa}_{\mathbf{x}}, \mathbf{pa}_{\mathbf{x}}) = \mathbf{x}$. Third, reversibility requires that the causal mechanism is invertible, ensuring a deterministic mapping between observations and counterfactuals. This means that if a counterfactual is generated as $\tilde{\mathbf{x}} := DSCM(\mathbf{x}, \mathbf{pa}_{\mathbf{x}}, \tilde{\mathbf{pa}}_{\mathbf{x}})$, the original state is uniquely recoverable via $\mathbf{x} := DSCM(\tilde{\mathbf{x}}, \tilde{\mathbf{pa}}_{\mathbf{x}}, \mathbf{pa}_{\mathbf{x}})$. Both composition and reversibility are evaluated using the Structural Similarity Index (SSIM), Fréchet Inception Distance (FID) [18] and Cosine Similarity (Cos Sim) for global feature-level consistency. SSIM, FID and Cosine Similarity are used to assess the causal properties of reversibility and composition. Their primary purpose is to quantitatively ensure that counterfactual interventions preserve the patient's invariant anatomical structure without introducing structural distortions. Meanwhile, the clinical and semantic correctness of the synthesised counterfactual is quantified via effectiveness. By employing an anti-causal predictor, this metric objectively verifies whether the intended pathological shift has been accurately simulated. Furthermore, the foundational clinical correctness and visual realism of this generative architecture have been comprehensively validated through expert blinded evaluations by clinicians in our prior work [46].

3.4. Probabilistic synthetic strategy

3.4.1. Simulating clinical reality

A fundamental challenge in causal inference in real-world settings is the prevalence of the treatment paradox, which is a manifestation

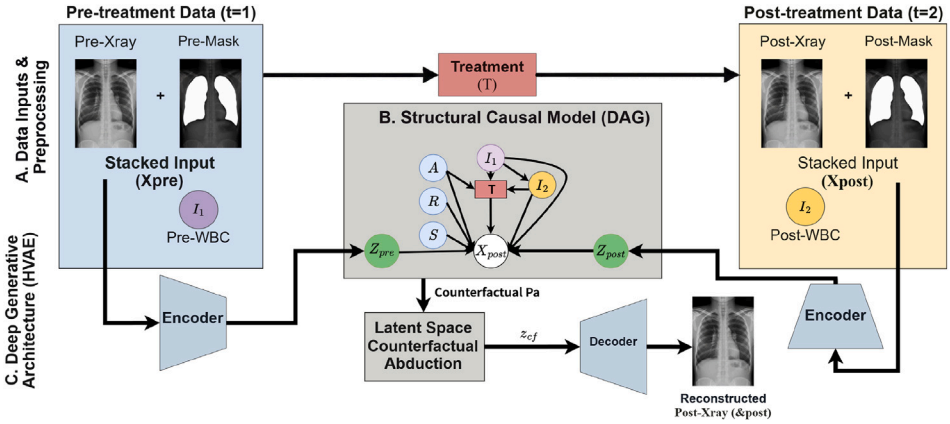


Fig. 3. Workflow for longitudinal counterfactual generation. The framework consists of three main components. (A) Data inputs & Preprocessing: pre-treatment (t = 1) and post-treatment (t = 2) chest X-rays are stacked with segmentation masks to form the image inputs, $\mathbf{X}_{pre}$ and $\mathbf{X}_{post}$, accompanied by clinical variables such as white blood cell counts, I_1 and I_2. (B) Structural causal model (DAG): A causal graph defines the relationships between variables. (C) Deep generative architecture: an HVAE encodes the image inputs, performs counterfactual abduction in the latent space based on the DAG, and uses a decoder to synthesise the reconstruction.

of Simpson's paradox [38] driven by confounding clinical severity. In observational datasets like MIMIC-CXR collected from ICU settings, aggressive interventions are preferentially administered to the most critically ill patients. This creates a counter-intuitive statistical association where treatment appears to correlate with higher mortality [7]. Consequently, standard models trained on such data risk learning treatment assignment patterns rather than the actual efficacy of the treatment [14]. To address this issue and generate valid counterfactual supervision, we construct a semi-synthetic pipeline that enforces biologically plausible causal dynamics, effectively overcoming bias inherent in the observational data.

We formulate the data-generation process using a Clinical Transition Matrix, which acts as a corrective probabilistic prior. Unlike the observational distribution where treatment is statistically linked to remaining stable or deteriorating, this matrix encodes the true physiological efficacy of the intervention. It assumes that, while the assignment of treatment depends on severity (proxied by the WCC vector), the treatment promotes clinical recovery. For instance, the matrix assigns a high probability of "recovery" to treated patients, thereby correcting the paradoxical artifact found in empirical data.

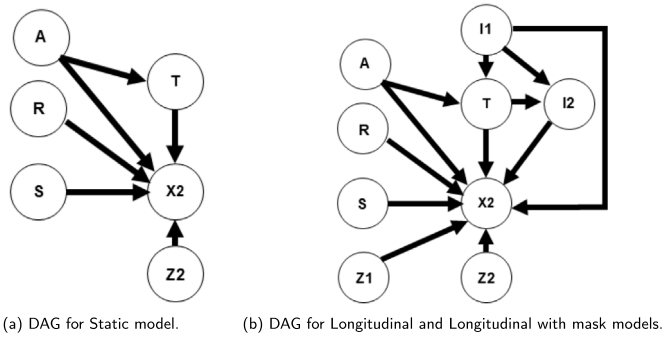
For each patient, we perform Monte Carlo sampling over this transition matrix to determine a potential clinical outcome. To convert this assigned outcome into a realistic synthetic X-ray image, we employ our previously developed disease counterfactual model [45]. This generative framework links the chosen clinical state directly to corresponding visual changes. Specifically, it modifies the patient's baseline image to match the target state, by either removing pathological features to simulate recovery or enhancing them to simulate deterioration. This procedure yields a longitudinal synthetic dataset, D_{syn}. Unlike the raw observational data, this dataset reflects valid causal mechanisms rather than spurious statistical correlations, providing a trustworthy ground truth for model training.

4. Treatment results

4.1. Dataset and preprocessing

We evaluated the proposed approach with the MIMIC-CXR dataset [23]. The disease counterfactual generative model assumed the following observed variables: age (A), sex (S), race (R), treatment (T), pre-treatment white blood cell count (I1), post-treatment white blood cell count (I2) and X-ray images. The corresponding DAG is shown in Fig. 4(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 [23]. 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 [21], which automatically interprets the presence of 14 disease observations from the reports. We resized the chest X-ray images to 192×192 resolution.

To enable longitudinal causal analysis, we filtered the dataset to identify distinct treatment episodes. We selected a single patient record that must perfectly align: (1) a pre-treatment chest X-ray, (2) an explicit clinical intervention occurring within a four-to-twelve-week interval, (3) a complete, corresponding set of systemic inflammatory biomarkers (such as WCC), and (4) a subsequent post-treatment chest X-ray. This selection process yielded a cohort of 1015 paired samples, and they are partitioned into training set (70%), validation set (10%) and test set (10%). Given the selection bias inherent in the raw MIMIC-CXR outcomes, we constructed a semi-synthetic dataset to serve as a reliable, proxy ground truth. We formulated a clinical transition matrix to guide the generation of synthetic outcomes based on infection severity and treatment status. The synthetic data generation follows three clinical rules based on patient severity and treatment. First, for effective intervention ($T = 1$), if a patient has abnormal WCC (indicating high severity) and receives treatment, we assign a high probability (e.g., 80%) to recovery. The model then generates a healthy image by removing pathological features. Second, for natural progression ($T = 0$), if a high-risk patient does not receive treatment, we assign a high probability to deterioration. Consequently, the model generates an image with enhanced disease patterns to simulate the worsening condition. Finally, for low-risk stability, patients with normal WCC are assigned outcomes with higher randomness, reflecting the uncertain and limited impact of intervention on mild cases.



(a) DAG for Static model. (b) DAG for Longitudinal and Longitudinal with mask models.

Fig. 4. Directed Acyclic Graphs (DAGs) for the proposed models: age (A), disease (D), race(R), sex (S) and pre-treatment white blood cell count (I1), post-treatment white blood cell count (I2) and treatment (T). Latent features of the pre- and post-treatment X-rays are denoted as Z1 and Z2 respectively.

4.2. Setup and implementation details

Our HVAE model used stochastic latent variables across 6 resolution scales, reaching up to half the input resolution: $1^2, 6^2, 12^2, 24^2, 48^2, 96^2$. These scales contained $2, 4, 8, 12, 8 and 4$ residual blocks, respectively. Each latent variable had 16 channels. The feature map widths at each scale were $512, 192, 160, 128, 96, 64, 32$, with 32 being the width of the final 192×192 deterministic upsampling block.

We trained the HVAE model for 200 epochs using a batch size of 16 and the AdamW optimiser. The initial learning rate was 5×10^{-5} with a 100-step warm-up. We set $[\beta_1, \beta_2] = [0.9, 0.9]$ and the weight decay to 0.05. Based on the L2 norm, we clipped gradients at 350 and skipped parameter updates if they exceeded 500. The final model parameters were calculated using an exponential moving average with rate of 0.999.

For data augmentation, we added a 9-pixel zero-padding to all borders and applied random horizontal flips with a 50% probability. Pixel intensities were scaled to a range of $[-1, 1]$. Because the pre-treatment and post-treatment white blood cell counts are discrete variables, standard normalising flows are not invertible. We address this by applying the Gumbel-max trick, which allows for sampling from discrete distributions [30,36].

4.3. Semi-synthetic data generation

To evaluate our models against the treatment paradox found in the MIMIC dataset, we established a semi-synthetic generation pipeline. In raw observational data, aggressive interventions are confounded by patient severity, creating a spurious statistical correlation where treatment appears linked to clinical deterioration or higher mortality. To rectify this and provide valid counterfactual supervision, we constructed the synthetic dataset, D_{syn}, by enforcing biologically plausible causal dynamics.

The generation process is guided by a clinician-informed clinical transition matrix, which acts as a corrective probabilistic prior based on the patient's pre-treatment white blood cell count severity ranges and treatment status. We define three specific severities:

- High Severity (white cell count higher than upper range): If treated (T = 1), patient has an 80% probability of transitioning to no finding and a 20% probability of remaining with pneumonia. If

Table 1
The counterfactual evaluation of the proposed model. The static model is based on a single post-treatment X-ray (DAG is presented in Fig. 4(a)). The Longitudinal model uses both pre- and post-treatment X-rays (DAG is shown in Fig. 4(b)). HVAE is the Longitudinal model excluding the causal component.

Model	Effectiveness	Reversibility			Composition		
	DRUG (d)	SSIM	FID	Cos Sim	SSIM	FID	Cos Sim
HVAE	0.1302	0.9998	0.0004	0.9999	0.9996	0.0006	0.9999
Ours (Static)	0.3208	0.9626	0.1973	0.9520	0.9848	0.0735	0.9759
Ours (Longitudinal)	0.9992	0.9787	0.1581	0.9652	0.9936	0.0471	0.9874
Ours (Longitudinal with Lung Mask)	0.8312	0.9706	0.1527	0.9593	0.9895	0.0549	0.9847

untreated (T = 0), the patient has a 30% probability of remaining with pneumonia and a 70% probability of deteriorating.
- Moderate Severity (white cell count lower than lower range): Treated (T = 1) patient has an 80% probability of 'no finding' and a 20% probability of remaining with pneumonia. An untreated (T = 0) patient shows a 60% probability of 'no finding', a 20% probability of remaining with pneumonia, and a 20% probability of deteriorating.
- Low Severity (white cell count within normal range): Treated (T = 1) patient 100% recovers to 'no finding'. An untreated (T = 0) patient has a 70% probability of recovering to 'no finding', a 10% probability of remaining with pneumonia, and a 20% probability of deteriorating.

For each patient, we perform Monte Carlo sampling over these defined, proxy ground truth distributions to determine their potential clinical outcome. Once the target semantic label is selected, we use a pre-trained disease counterfactual generative model to intervene on corresponding outcome, either by removing pathological features to simulate recovery or introducing them to simulate deterioration.

4.4. Treatment counterfactual generation

To accurately evaluate the models on the synthetic dataset, the quantitative metrics, shown in Table 1 must be analysed jointly. A high effectiveness (DRUG(d)) indicates that the generated images correctly reflect the intended clinical intervention. However, this must be paired with high reversibility and composition to prove that the SCM has learned consistent and accurate causal mechanisms. The baseline HVAE model achieved near-perfect reversibility and composition, such as SSIM for reversibility of 0.9998. However, it recorded the lowest effectiveness of 0.1302. This discrepancy shows that the HVAE simply copies the input image without applying meaningful treatment effects. Because it lacks a causal structure, it achieves high image similarity but fails to perform true counterfactual generation.

Incorporating an SCM improved the causal generation. The Static model, which relies on a single post-treatment X-ray (DAG shown in Fig. 4(a)), achieved an effectiveness of 0.3208. The Longitudinal model, which uses both pre- and post-treatment X-rays (DAG shown in Fig. 4(b)), significantly outperformed the Static model. It reached a remarkably high effectiveness of 0.9992 while simultaneously improving reversibility FID to 0.1581 and composition FID to 0.0471. This joint success proves that providing temporal context allows the SCM to capture the evolution of disease, while maintaining causal consistency. The Longitudinal with lung mask model, which adds a lung mask to focus on pulmonary features, also demonstrated strong joint performance. It achieved a solid effectiveness of 0.8312 alongside high reversibility FID of 0.9706 and composition FID of 0.9895. This indicates that the Longitudinal with mask model successfully learns consistent SCM dynamics even when restricted by anatomical constraints.

Although the unmasked Longitudinal model achieved the highest joint quantitative performance, qualitative visual inspection, shown in Fig. 5, revealed an important clinical limitation. The visual results demonstrate that the Longitudinal model tends to modify the patient anatomy, such as the neck and arm. It incorrectly changes non-pulmonary regions, altering background tissues, bones, and anatomical structures that should remain completely unaffected by the treatment intervention. In contrast, the Longitudinal with mask model successfully resolves this issue. While its quantitative performances are slightly lower than those of Longitudinal model, the explicit lung mask forces the model to restrict its modifications strictly to the lung areas. This results in more accurate and visually realistic counterfactuals. The Longitudinal with mask model changes only the relevant disease areas while preserving the rest of the patient's invariant anatomy, making it far more appropriate for clinical applications.

4.5. Clinical semantic evaluation

To validate the clinical semantic correctness of our generated counterfactuals, we employed a pre-trained, independent diagnostic classifier (TorchXrayVision) to estimate the average treatment effect (ATE) on the generated post-treatment images. As presented in Table 2, the quantitative results align closely with expected physiological trajectories. Under the simulated active treatment intervention ($T = 0 \rightarrow 1$), the generated counterfactuals exhibited a reduced the probability of pathological presence, yielding an ATE of −0.2069 for pneumonia and −0.1034 for effusion. Concurrently, this treatment intervention yielded a positive ATE of 0.3103 for "no finding", quantitatively demonstrating that structural causal model successfully generates semantic visual features indicative of clinical recovery.

Conversely, when simulating the withdrawal of treatment of previously treated patients ($T = 1 \rightarrow 0$), the model accurately reflects clinical deterioration. This is evidenced by an expected mean increase in pneumonia (ATE = 0.2184) and a corresponding reduction in "no finding" (ATE = −0.2414). Furthermore, the standard deviation of the individual treatment effect (Std ITE), which ranges from 0.2835 to 0.5140 across conditions, reveals significant heterogeneity in individual patient responses. This high variance empirically validates the fundamental necessity of our proposed high-dimensional, patient-specific generative framework, demonstrating that proportional-level statistical ATE alone is insufficient to capture the individualised pathological shift required for precision clinical decision-making.

5. Discussion

Our work evaluates the proposed model for treatment counterfactual generation using synthetic data based on the real-world MIMIC dataset. The original MIMIC data was collected from the ICU. This provenance creates a major challenge for causal learning. Within this dataset, there is a strong confounding bias. Patients who received treatment often show worse clinical outcomes, while patients who did not receive treatment show better outcomes. This happens because only the most severely ill patients are given the drug. This data distribution contradicts general clinical knowledge. Therefore, we synthesise a dataset following the clinicians' guidance.

Our results show that providing an SCM with temporal context successfully generates clinically meaningful treatment counterfactuals. As shown in Table 1, the Longitudinal model achieved the best overall

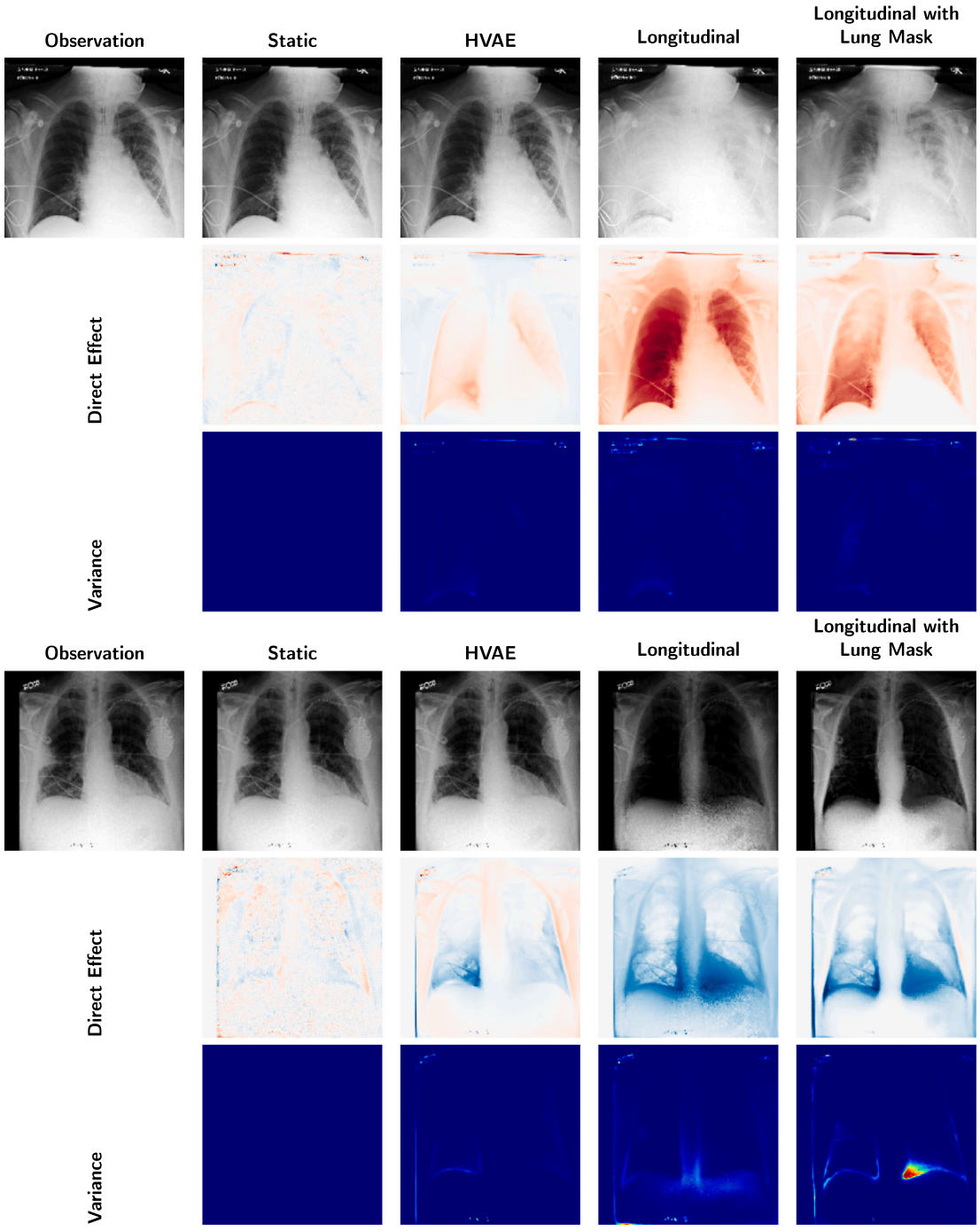


Fig. 5. Visualisation of treatment counterfactuals. Comparison of the baseline with our proposed counterfactual generation models, which incorporate a longitudinal DAG and an anatomy-aware mask. Red regions in the direct effect maps indicate an increase in pixel values, while blue regions indicate a decrease. The static model is based on a single post-treatment X-ray (DAG is presented in Fig. 4(a)). The Longitudinal model uses both pre- and post-treatment X-rays (DAG is shown in Fig. 4(b)). HVAE is the Longitudinal model excluding the causal component.

quantitative performance for effectiveness, reversibility, and composition. Despite these strong evaluation metrics, Fig. 5 column 4 illustrates a critical failure in causal learning. The visual results show that the Longitudinal model excessively modifies the counterfactual images. When applying a counterfactual intervention, the model incorrectly changes non-pulmonary regions like bones and background tissues. This over-modification occurs because the Longitudinal architecture relies on both pre-treatment and post-treatment X-rays. In a real clinical setting, these two scans are taken at different times. This leads to differences in patient pose and spatial alignment. The deep learning model fails to disentangle the true treatment effect from the spatial changes caused by these pose inconsistencies. Consequently, it incorrectly alters background variables and anatomical structures that should remain causally invariant to the intervention. To avoid overfitting, we implemented strict data partitioning and training protocols. The dataset was strictly partitioned into separate training, validation, and testing splits. Throughout the training phase, the model's generalisation was continuously monitored via the hold-out validation set, and no empirical evidence of overfitting was observed. Importantly, all quantitative results were computed on the independent test set.

Table 2
Comparative analysis of causal inference metrics across different pathologies. This table presents the factual probabilities and treatment effects under two scenarios: transitioning from control to treated (Drug 0 → 1) and vice versa (Drug 1 → 0). Prob(0/1) denotes the factual probability under the respective drug condition; ATE (Average Treatment Effect) represents the expected mean change in probability; and Std ITE (Standard Deviation of Individual Treatment Effect) quantifies the heterogeneity of the effect.

Pathology	Control to Treated (T = 0 → 1)			Treated to Control (T = 1 → 0)		
	Prob(T = 0)	ATE	Std ITE	Prob(T = 1)	ATE	Std ITE
Pneumonia	0.2759	−0.2069	0.4913	0.1897	0.2184	0.5140
Effusion	0.3103	−0.1034	0.4093	0.0575	0.0230	0.2835
No Finding	0.4138	0.3103	0.4708	0.7529	−0.2414	0.4919

The Longitudinal with lung mask model successfully solves this problem stemming from pose inconsistencies. Fig. 5 last column demonstrates that introducing an explicit lung mask acts as a strong causal constraint. This spatial prior forces the deep learning model to restrict all counterfactual changes strictly to the pulmonary regions. Therefore, the model correctly disentangles disease-specific causal features from background anatomical structures. This strict constraint leads to a reduction in global performance metrics compared with the Longitudinal model. However, this trade-off is necessary to ensure the generation of clinically valid counterfactuals.

Finally, we consider the ethical implications of using synthetic counterfactuals. While observational databases like MIMIC are already rigorously de-identified to protect patient privacy, generative models introduce an additional layer of structural anonymity, as the synthesized outputs represent mathematically simulated physiological states rather than real individuals. However, the primary ethical concern lies in algorithmic fairness and clinical safety. Generative models may inherit and potentially amplify underlying demographic or clinical biases present in the retrospective training data. Consequently, from both an ethical and patient-safety perspective, we strongly advocate for a strict 'human-in-the-loop' paradigm. These causal generative frameworks should be deployed exclusively as auxiliary decision-support tools, with certified clinicians retaining the ultimate diagnostic and therapeutic authority to ensure synthetic outputs do not autonomously determine treatment decisions.

6. Conclusion

Observational clinical data often contain bias that makes active treatments appear harmful. To solve this problem, we developed a longitudinal deep structural causal model that generates unbiased synthetic chest X-ray images. Initially, we introduce pre- and post-treatment X-ray images to capture the evolution of the disease over time. However, variations arising from inter-scan variability caused the model to incorrectly alter background anatomical regions, such as bones, in the counterfactual generations. Therefore, we introduced lung masks to force the model to focus on the pulmonary regions. Although this mask slightly reduces performance, it produces much more realistic and clinically valid X-ray images. By separating true treatment effects from patient severity and positional differences, our model provides doctors with a robust decision-support tool to visualise hypothetical patient outcomes and facilitate informed clinical decision-making.

CRediT authorship contribution statement

Yifei Zhu: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Formal analysis, Conceptualization. **Lei Zhang:** Writing – review & editing, Supervision, Methodology, Conceptualization. **Chris Sainsbury:** Writing – review & editing, Conceptualization. **Feng Dong:** Writing – review & editing, Supervision, Conceptualization. **John MacLay:** Writing – review & editing, Conceptualization. **David J. Lowe:** Writing – review & editing, Conceptualization. **Xujiong Ye:** Writing – review & editing, Supervision, Methodology, Conceptualization.

Ethic statement

In this study,data were sourced from publicly available repositories, and no personally identifiable information was used. Efforts were made to mitigate biases in the training dataset and to ensure fair outcomes. This study uses de-identified patient data from the MIMIC-III/IV database (Medical Information Mart for Intensive Care), a publicly available database developed by the Massachusetts Institute of Technology (MIT) and Beth Israel Deaconess Medical Center (BIDMC). Ethical approval for the use of this database was provided by the Institutional Review Boards of both MIT and BIDMC, and the data are de-identified in accordance with the Health Insurance Portability and Accountability Act (HIPAA) standards. Therefore, additional ethical approval was not required for this study. No patient was directly involved in the research process, and no personally identifiable information was used in this study. This study has no foreseeable harmful effects on humans or the environment, and it complies with GDPR regulations.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Yifei Zhu reports that financial support was provided by Engineering and Physical Sciences Research Council. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

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

Data availability

This study utilised data from two publicly available resources hosted on PhysioNet: the MIMIC-IV database (v[2.2]) and the MIMIC-CXR-JPG database (v[2.0.0]). Both datasets contain de-identified patient data. Access is granted to credentialed researchers who complete the required human subjects research training and sign a Data Use Agreement. Further information about accessing these datasets is available on the PhysioNet website (https://physionet.org/).

References

[1] RSNA Pneumonia Detection Challenge, Radiological society of North America (RSNA), AI image challenge, 2018. https://www.kaggle.com/c/rsna-pneumonia-detection-challenge (Accessed: 28 October 2024).

[2] M.H. Al-Sheikh, O. Al Dandan, A.S. Al-Shamayleh, H.A. Jalab, R.W. Ibrahim, Multi-class deep learning architecture for classifying lung diseases from chest X-Ray and CT images, Sci. Rep. 13 (2023) 19373.

[3] M. Asim, F. Amin, A. El-Menyar, Multiple organ dysfunction syndrome: contemporary insights on the clinicopathological spectrum, Qatar medical journal 2020 (2020) 22.

[4] M.C. Bachmann, M. Benites, V. Oviedo, N.H. Vadeghani, D. Soto, R. Basoalto, P. Cruces, Y. Jalil, L.F. Damiani, G. Bugedo, et al., Association between controlled mechanical ventilation and systemic inflammation in acute hypoxemic respiratory failure: an observational cohort study, Crit. Care 29 (2025) 482.

[5] D. Barber, F. Agakov, The im algorithm: a variational approach to information maximization, Adv. Neural Inf. Process. Syst. 16 (2004) 201.

[6] R.C. Bone, R.A. Balk, F.B. Cerra, R.P. Dellinger, A.M. Fein, W.A. Knaus, R.M. Schein, W.J. Sibbald, Definitions for sepsis and organ failure and guidelines for the use of innovative therapies in sepsis, Chest 101 (1992) 1644–1655.

[7] R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, N. Elhadad, Intelligible models for healthcare: predicting pneumonia risk and hospital 30-day readmission, in: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2015, pp. 1721–1730.

[8] D.C. Castro, I. Walker, B. Glocker, Causality matters in medical imaging, Nat. Commun. 11 (2020) 3673.

[9] H. Choi, H. Jung, Deep generative models for health, in: Proceedings of Annual Conference on Neural Information Processing Systems, 2023.

[10] J.P. Cohen, M. Luck, S. Honari, Distribution matching losses can hallucinate features in medical image translation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2018, pp. 529–536.

[11] J.P. Cohen, J.D. Viviano, P. Bertin, P. Morrison, P. Torabian, M. Guarrera, M.P. Lungren, A. Chaudhari, R. Brooks, M. Hashir, H. Bertrand, TorchXRayVision: a library of chest X-ray datasets and models, in: E. Konukoglu, B. Menze, A. Venkataraman, C. Baumgartner, Q. Dou, S. Albarqouni (Eds.), Proceedings of the 5th International Conference on Medical Imaging with Deep Learning, PMLR, 2022, pp. 231–249, https://proceedings.mlr.press/v172/cohen22a.html.

[12] E. DeVos, L. Jacobson, Approach to adult patients with acute dyspnea, Emergency Medicine Clinics 34 (2016) 129–149.

[13] S. Dhanireddy, W.A. Altemeier, G. Matute-Bello, D.S. O'Mahony, R.W. Glenny, T.R. Martin, W.C. Liles, Mechanical ventilation induces inflammation, lung injury, and extra-pulmonary organ dysfunction in experimental pneumonia, Lab. Investig. 86 (2006) 790–799.

[14] J. Futoma, M. Simons, T. Panch, F. Doshi-Velez, L.A. Celi, The myth of generalisability in clinical research and machine learning in health care, Lancet Digit. Health 2 (2020) e489–e492.

[15] R. Geirhos, J.H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, F.A. Wichmann, Shortcut learning in deep neural networks, Nat. Mach. Intell. 2 (2020) 665–673.

[16] I.J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, Adv. Neural Inf. Process. Syst. 27 (2014).

[17] A. Hage Chehade, N. Abdallah, J.M. Marion, M. Hatt, M. Oueidat, P. Chauvet, Advancing chest X-ray diagnostics: a novel CycleGAN-based preprocessing approach for enhanced lung disease classification in ChestX-Ray14, Comput. Methods Programs Biomed. 259 (2025) 108518, https://doi.org/10.1016/j.cmpb.2024.108518, https://www.sciencedirect.com/science/article/pii/S016926072400511X.

[18] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, S. Hochreiter, GANs trained by a two time-scale update rule converge to a local Nash equilibrium, Adv. Neural Inf. Process. Syst. 30 (2017).

[19] A.T. Hill, A.L. Sullivan, J.D. Chalmers, A. De Soyza, J.S. Elborn, R.A. Floto, L. Grillo, K. Gruffydd-Jones, A. Harvey, C.S. Haworth, E. Hiscocks, J.R. Hurst, C. Johnson, W.P. Kelleher, P. Bedi, K. Payne, H. Saleh, N.J. Screaton, M. Smith, M. Tunney, D. Whitters, R. Wilson, M.R. Loebinger, British thoracic society guideline for bronchiectasis in adults, Thorax 74 (2019) 1–69, https://doi.org/10.1136/thoraxjnl-2018-212463

[20] J. Ho, A. Jain, P. Abbeel, Denoising diffusion probabilistic models, Adv. Neural Inf. Process. Syst. 33 (2020) 6840–6851.

[21] J.A. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R.L. Ball, K.S. Shpanskaya, J. Seekins, D.A. Mong, S.S. Halabi, J.K. Sandberg, R. Jones, D.B. Larson, C. Langlotz, B.N. Patel, M.P. Lungren, A. Ng, CheXpert: a large chest radiograph dataset with uncertainty labels and expert comparison, in: AAAI Conference on Artificial Intelligence, 2019, https://api.semanticscholar.org/CorpusID:58981871.

[22] M. Jasmine Pemeena Priyadarsini, K. Kotecha, G. Rajini, K. Hariharan, K. Utkarsh Raj, K. Bhargav Ram, V. Indragandhi, V. Subramaniyaswamy, S. Pandya, Lung diseases detection using various deep learning algorithms, J. Healthc. Eng. 2023 (2023) 3563696.

[23] A. Johnson, T. Pollard, S. Berkowitz, N. Greenbaum, M. Lungren, C.Y. Deng, R. Mark, S. Horng, MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports, Sci. Data 6 (2019) 317, https://doi.org/10.1038/s41597-019-0322-0

[24] D.P. Kingma, T. Salimans, R. Jozefowicz, X. Chen, I. Sutskever, M. Welling, Improved variational inference with inverse autoregressive flow, Adv. Neural Inf. Process. Syst. 29 (2016).

[25] W.W. Loh, D. Ren, Estimating time-varying treatment effects in longitudinal studies, Psychol. Methods 30 (2025) 240.

[26] M. Monteiro, F.D.S. Ribeiro, N. Pawlowski, D.C. Castro, B. Glocker, Measuring axiomatic soundness of counterfactual image models, in: The Eleventh International Conference on Learning Representations, 2023, https://openreview.net/forum?id=lZOUQQvwI3q.

[27] S.L. Moroianu, C. Bluethgen, P. Chambon, M. Cherti, J.B. Delbrouck, M. Paschali, B. Price, J. Gichoya, J. Jitsev, C.P. Langlotz, et al., Improving performance, robustness, and fairness of radiographic AI models with finely-controllable synthetic data, arXiv preprint arXiv:2508.16783, 2025.

[28] National Institute for Health and Care Excellence, Pneumonia: diagnosis and management, NICE Guideline NG250, National Institute for Health and Care Excellence, London, 2025. https://www.nice.org.uk/guidance/ng250.

[29] Z. Obermeyer, B. Powers, C. Vogeli, S. Mullainathan, Dissecting racial bias in an algorithm used to Manage the health of populations, Science 366 (2019) 447–453.

[30] N. Pawlowski, D. Coelho de Castro, B. Glocker, Deep structural causal models for tractable counterfactual inference, Adv. Neural Inf. Process. Syst. 33 (2020) 857–869.

[31] J. Pearl, Causality: Models, Reasoning and Inference, Cambridge University Press, New York, 2000.

[32] J. Pearl, Causality, Cambridge University Press, 2009.

[33] J. Peters, D. Janzing, B. Schölkopf, Elements of Causal Inference: Foundations and Learning Algorithms, Adaptive Computation and Machine Learning, MIT Press, Cambridge, MA, 2017, https://mitpress.mit.edu/books/elements-causal-inference.

[34] A. Rao, D. Gray, Breathlessness in hospitalised adult patients, Postgraduate medical journal 79 (2003) 681–685.

[35] D.J. Rezende, S. Mohamed, D. Wierstra, Stochastic backpropagation and approximate inference in deep generative models, in: Proceedings of International Conference on Machine Learning, PMLR, 2014, pp. 1278–1286.

[36] F.D.S. Ribeiro, T. Xia, M. Monteiro, N. Pawlowski, B. Glocker, High fidelity image counterfactuals with probabilistic causal models, in: Proceedings of International Conference on Machine Learning, 2023.

[37] B. Scholkopf, F. Locatello, S. Bauer, N.R. Ke, N. Kalchbrenner, A. Goyal, Y. Bengio, Towards causal representation learning, ArXiv arXiv:2102.11107, 2021, https://api.semanticscholar.org/CorpusID:231986372.

[38] E.H. Simpson, The interpretation of interaction in contingency tables, Journal of the Royal Statistical Society: Series B (Methodological) 13 (1951) 238–241.

[39] M.W. Sjoding, K. Luo, M.A. Miller, T.J. Iwashyna, When do confounding by indication and inadequate risk adjustment bias critical care studies? A simulation study, Crit. Care 19 (2015) 195.

[40] C.K. Sønderby, T. Raiko, L. Maaløe, S.K. Sønderby, O. Winther, Ladder variational autoencoders, Adv. Neural Inf. Process. Syst. 29 (2016).

[41] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, R.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, 2017.

[42] G.W. Waterer, L.A. Kessler, R.G. Wunderink, Infection-directed clinical prediction rules for community-acquired pneumonia, J. Clin. Epidemiol. 54 (2001) 1253–1259.

[43] X. Yi, E. Walia, P. Babyn, Generative adversarial network in medical imaging: a review, Med. Image Anal. 58 (2019) 101552, https://doi.org/10.1016/j.media.2019.101552, https://www.sciencedirect.com/science/article/pii/S1361841518308430.

[44] H. Zhao, J. Shi, X. Qi, X. Wang, J. Jia, Pyramid scene parsing network, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 6230–6239, https://api.semanticscholar.org/CorpusID:5299559.

[45] Y. Zhu, L. Zhang, C. Sainsbury, F. Dong, J. MacLay, D.J. Lowe, X. Ye, Counterfactual medical images generation for lung disease diagnosis using probabilistic causal models and active learning, IEEE Access 13 (2025) 170817–170826, https://doi.org/10.1109/ACCESS.2025.3615683.

[46] Y. Zhu, L. Zhang, C. Sainsbury, F. Dong, J. MacLay, D.J. Lowe, X. Ye, Causal and active learning-based counterfactual chest X-ray generation for supporting clinical decision-making in lung disease, in: Proceedings of the Fifth Conference on Causal Learning and Reasoning, 2026. To appear.