RadEdit: stress-testing biomedical vision models via diffusion image editing

Fernando Pérez-García[⋆,1], Sam Bond-Taylor[⋆,1], Pedro P. Sanchez[+,2], Boris van Breugel[+,3], Daniel C. Castro[1], Harshita Sharma[1], Valentina Salvatelli[1], Maria T.A. Wetscherek[1], Hannah Richardson[1], Matthew P. Lungren[1,4,5], Aditya Nori[1], Javier Alvarez-Valle[1], Ozan Oktay [†,1], and Maximilian Ilse [†,1]

[1] Microsoft Health Futures
[2] University of Edinburgh
[3] University of Cambridge
[4] University of California
[5] Stanford University
[⋆] Shared first author
[+] Work done at Microsoft Health Futures
[†] Shared last author

Abstract. Biomedical imaging datasets are often small and biased, meaning that real-world performance of predictive models can be substantially lower than expected from internal testing. This work proposes using generative image editing to simulate dataset shifts and diagnose failure modes of biomedical vision models; this can be used in advance of deployment to assess readiness, potentially reducing cost and patient harm. Existing editing methods can produce undesirable changes, with spurious correlations learned due to the co-occurrence of disease and treatment interventions, limiting practical applicability. To address this, we train a text-to-image diffusion model on multiple chest X-ray datasets and introduce a new editing method, RadEdit, that uses multiple image masks, if present, to constrain changes and ensure consistency in the edited images, minimising bias. We consider three types of dataset shifts: acquisition shift, manifestation shift, and population shift, and demonstrate that our approach can diagnose failures and quantify model robustness without additional data collection, complementing more qualitative tools for explainable AI.

Keywords: Image editing · diffusion models · biomedical imaging

1 Introduction

Developing accurate and robust models for biomedical image analysis requires large and diverse datasets that are often difficult to obtain due to ethical, legal, geographical, and financial constraints [41]. This leads to biased training datasets that affect the performance of trained models and generalisation to real-world scenarios [59, 40]. Such data mismatch may arise from genuine differences in upstream data acquisition as well as from the selection criteria for dataset creation, which materialise as various forms of dataset shifts (population, acquisition, annotation, prevalence, manifestation) [7].

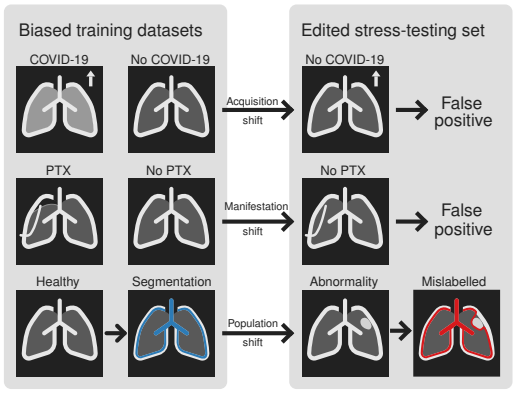


Fig. 1: Stress-testing models by simulating dataset shifts via image editing. *Top*: editing out COVID-19 features results in false positives since the classifier relies on acquisition differences, e.g., radiographic markers (white arrow). *Middle*: editing out a pneumothorax (PTX) results in false positives since the classifier instead detects chest drains. *Bottom*: editing abnormalities into lungs causes a lung segmentation model to mislabel (blue: ground-truth segmentation; red: model prediction).

Biomedical vision models, when put into real-world use, can be unhelpful or potentially even harmful to patients if they are affected by dataset shifts, leading to missed diagnoses [22, 74, 76, 56]. For example, the COVID-19 pandemic led to hundreds of detection tools being developed, with some put into use in hospitals; yet Roberts et al. [56] found that *"none of the models identified are of potential clinical use due to methodological flaws and/or underlying biases."* It is therefore crucial to properly assess models for bias, prior to real-world use.

Recent deep generative models have made remarkable improvements in terms of sample quality, diversity, and steerability [57, 48, 35, 28]. These models have been shown to generalise to out-of-distribution domains [42, 6, 32, 19], opening up avenues for new applications. One such application is generating synthetic data for stress-testing models [51, 42, 72]. This involves creating data that is realistic, yet can represent settings, domains, or populations that do not appear (enough) in the real training/test data.

In this work, we investigate how deep generative models can be used for stress-testing biomedical imaging models. We consider three dataset shift scenarios:

1. **Acquisition shift:** classifying COVID-19 cases when the positive and negative cases were acquired at different hospitals (Sec. 5.2).
2. **Manifestation shift:** detecting if pneumothorax[1] was resolved when chest drains (inserted to treat pneumothorax) are present (Sec. 5.3).
3. **Population shift:** segmenting lungs in the presence of abnormalities rarely or never seen in the training dataset (Sec. 5.4).

For each of these scenarios, we simulate dataset shifts, producing stress-test sets which can occur in the real world but do not appear or are underrepresented in the original training/test sets. Following prior work, these test sets are synthesised using generative image editing, which unlike generating images from scratch, only minimally modifies the images, hence, better retains fidelity and diversity [51, 42]. For the above scenarios, we use generative editing to 1. remove only COVID-19 features while keeping visual indicators of the different hospitals; 2. remove only pneumothorax while keeping the chest drain; and 3. add abnormalities that occlude lung structures in the image.

[1] We provide descriptions of the medical terms used throughout the paper in Appendix A

We train a diffusion model [27, 57] on a large collection of chest X-rays from a variety of biomedical imaging datasets (Sec. 5.1), enabling us to add and remove a wide variety of pathologies and support devices when editing. Despite the diversity within these datasets, substantial biases are still present, some of which are learned by the generative model; as a result, correlated features may also be modified. For example, in Scenario 2, removing the pneumothorax might also remove the chest drains as both features typically co-occur in datasets [58], since chest drains are used to treat pneumothorax. Furthermore, when editing only within editing masks, artefacts often appear at the border of the masks. Lastly, artefacts occur when editing images outside of the training dataset domain of the diffusion model used for editing. To overcome these challenges, we propose using multiple masks to break existing correlations. This involves defining which regions must change, and explicitly forcing correlated regions to remain unchanged. In addition, we allow the area outside of the masks to be modified by the diffusion model to ensure image consistency. Since our proposed editing method, which we call RadEdit, leads to only minimal overall changes of chest X-rays, we are able to generate synthetic datasets that can be used to stress-test segmentation models (Scenario 3), which, to the best of our knowledge, we are the first to demonstrate.

In summary, our contributions are as follows:

- We introduce a novel editing approach that reduces the presence of artefacts in edited images and simplifies prompt construction compared to prior work [10, 51].
- Our editing approach allows us to construct synthetic datasets with specific data shifts by performing zero-shot edits on datasets/abnormalities not seen in training.
- We conduct a broad set of experiments using these synthetic datasets to stress-test and expose biases in biomedical classification and, for the first time, segmentation models, introducing a new use case of synthetic data into the medical setting.

2 Preliminaries

In this section, we introduce background context for stress-testing biomedical imaging models: failure modes of biomedical imaging models caused by different dataset shifts; diffusion models as versatile generative models; and diffusion-based image editing.

2.1 Dataset shifts

Dataset shift refers to a discrepancy between the training and test data distributions due to external factors [7, 34]. Such shifts are regularly observed in machine learning for biomedical imaging, often due to data scarcity. For example, collected training datasets might consist primarily of healthy patients. However, when the model is used in practice after training, there could be a shift towards unhealthy patients. A taxonomy of different types of dataset shifts in the context of biomedical imaging was developed by Castro et al. [7]. In this paper, we consider three dataset shifts of particular interest.

Acquisition shift results from the use of different scanners (manufacturer, hardware, and software) or imaging protocols as often encountered when using data from multiple cohorts. These changes affect factors such as image resolution, contrast, patient positioning, and image markings.

Manifestation shift results from the way the prediction targets physically manifest in anatomy changes between domains. For example, training datasets could consist of more severe pathological cases than observed in practice, or a pathology may co-occur with different visual features, e.g., support devices.

Population shift results from differences in intrinsic characteristics of the populations under study, changing the anatomical appearance distribution. This definition encompasses examples such as age, sex, ethnicity, and comorbidities, but also abnormalities such as pleural effusion and support devices. In contrast to manifestation shift, the shift in anatomical appearance is not affected by prediction targets.

2.2 Diffusion models

Denoising diffusion probabilistic models (DDPMs) [27, 67] are a versatile and effective class of generative models that enable sampling from the data distribution by learning to denoise samples corrupted with Gaussian noise. DDPMs are formed by defining a forward time process that gradually adds noise to data points x_0 through the recursion

$$x_t = \sqrt{1-\beta_t}x_{t-1} + \sqrt{\beta_t}\epsilon_t, \quad t = 1, \dots, T, \quad \text{s.t. } x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\bar{\epsilon}_t\,, \tag{1}$$

where $\epsilon_{1:T}, \bar{\epsilon}_{1:T} \sim \mathcal{N}(0, I)$, $\beta_{1:T}$ is a predefined noise schedule that determines how quickly to corrupt the data and ensures that x_T contains little to no information about x_0, and $\bar{\alpha}_t = \prod_{s=1}^{t}(1-\beta_s)$. To form a generative model, the process is reversed in time, gradually transforming Gaussian noise into samples from the learned distribution. While the exact reversal is intractable, a variational approximation is defined by [68]:

$$x_{t-1} = \hat{\mu}_t(x_t, f_\theta(x_t, t, c)) + \sigma_t z_t, \tag{2}$$

$$\hat{\mu}_t(x_t, \epsilon_t) = \sqrt{\bar{\alpha}_{t-1}}\frac{x_t - \sqrt{1-\bar{\alpha}_t}\epsilon_t}{\sqrt{\bar{\alpha}_t}} + \sqrt{1-\bar{\alpha}_{t-1}-\sigma_t^2}\epsilon_t, \tag{3}$$

where c is a conditioning signal such as a text description, $f_\theta(x_t, t, c)$ is a learned approximation of the noise $\bar{\epsilon}_t$ that corrupted the image x_0 to obtain x_t, $z_{1:T} \sim \mathcal{N}(0, I)$, and $\sigma_{1:T}$ controls how much noise is introduced. The process is Markovian and known as a DDPM [27] when $\sigma_t = \sqrt{(1-\bar{\alpha}_{t-1})/(1-\bar{\alpha}_t)}\sqrt{1-\bar{\alpha}_t/\bar{\alpha}_{t-1}}$, while for $\sigma_t = 0$ the process is deterministic and is called a denoising diffusion implicit model (DDIM) [68].

2.3 Image editing

Since DDIMs have a one-to-one correspondence with latent vectors x_T, we can deterministically map data points to latents by running the DDIM generative process in reverse [68], called DDIM inversion. Several approaches [10, 46] have shown that images can be edited by running the reverse diffusion process augmented by the latent vectors and a modified prompt c. However, this method can lead to undesired artefacts in the edited images. For example, structures unrelated to the desired edit may also change shape, size, or location. To address this, Huberman-Spiegelglas et al. [29] propose DDPM inversion. Here, the original forward process defined in Eq. (1) is adapted, replacing the correlated vectors $\bar{\epsilon}_{1:T}$ with statistically independent vectors $\tilde{\epsilon}_{1:T}$ (more details in Appendix B). These noise vectors are then used in the generative process, retaining the structure of the original image better than DDIM inversion.

3 Related work

3.1 Generative image editing

With advances in deep generative modelling, several approaches to image editing have emerged. Many of these early approaches use compressed latent manipulation [12, 52, 64, 71] where fine-grained edits are difficult to achieve and can result in unwanted changes. More recently, the unparalleled flexibility of diffusion models, together with advances in plain text conditioning, have opened up new avenues for editing techniques.

Here, we describe some notable diffusion editing methods. SDEdit [46] shows that diffusion models trained solely on real images can be used to generate images from sketches by perturbing sketches with noise, then running the reverse diffusion process. Palette [60] is a diffusion model trained for inpainting by filling regions with noise and denoising. Blended diffusion [2, 3] uses masks with CLIP [53] conditioning to guide local edits. Multiple works show that injecting U-Net activations, obtained by encoding the original image into the generation process, better retains image structure [24, 70]. DiffEdit [10] uses text prompts to determine the appropriate region to edit. Mokady et al. [47] improve diffusion inversion quality by optimising the diffusion trajectory.

Crucially, in the works which use masks for editing, a single type of mask is always used to define the region of interest. In this work, we argue that a second type of mask is required to avoid the loss of features caused by spurious correlations. As better editing approaches are developed, this requirement should be kept in mind.

3.2 Biomedical imaging counterfactuals

Generative models have also been applied to biomedical counterfactual generation. Reinhold et al. [55] manipulate causes of multiple sclerosis in brain MRI with deep structural causal models [49]. Sanchez et al. [62] and Fontanella et al. [15] use editing to remove pathologies for abnormality detection. Ktena et al. [39] generate out-of-distribution samples to improve classifier performance. Gu et al. [21] train a diffusion model to model disease progression by conditioning on a prior X-ray and text progression description. Unlike our approach, these methods do not use masks to enforce which regions may or may not be edited, meaning that spurious correlations might affect edits. Additionally, these methods use synthetic data to augment and improve model performance whereas we focus on using synthetic medical data for stress-testing.

3.3 Stress-testing

Several approaches have used non-deep-generative-model methods to stress-test networks. Hendrycks and Dietterich [23] evaluate classification models' robustness to corruptions such as blurring, Gaussian noise, and JPEG artefacts. Sakaridis et al. [61] stress-test a segmentation model for roads by using an optical model to add synthetic fog to scenes. Koh et al. [38] collate a dataset presenting various distribution shifts.

More recent models have made use of conditional generative models to simulate shifts. Prabhu et al. [51] propose LANCE, which stress-tests classification models by using diffusion-based image editing to modify image subjects via caption editing with

a large language model (LLM); Kattakinda et al. [36] do similar, but instead modify the background. Li et al. [42] use diffusion models with a single subject mask to separately edit backgrounds and subjects. Van Breugel et al. [72] use generative adversarial networks to simulate distribution shifts on tabular data. This line of research is partially related to adversarial attacks [20], where the focus is on minimally modifying images such that they are visually indistinguishable to a human, but the attacked model fails.

3.4 Limitations of existing editing-based stress-testing methods

Recent advancements in diffusion modelling have drastically improved image editing. However, two prevalent approaches, LANCE [51] and DiffEdit [10], produce artefacts in medical images, making them unsuitable for stress-testing biomedical vision models.

LANCE only uses a global prompt (no mask) for image editing. While effective in the natural image domain, it leads to artefacts in the biomedical domain. For example in Sec. 5.4, we add pathologies and support devices to images of healthy lungs to stress-test lung segmentation models; in such cases we must ensure that the position and shape of the lung borders are not altered during editing. However, we find that LANCE changes the position and shape of the lung border thus making edited images unsuitable for stress-testing such models (Sec. 5.5). In addition, we find that LANCE potentially removes support devices when prompted to remove pathologies, a direct effect of the correlations in the diffusion model's training datasets (Sec. 5.1), making LANCE unsuited for testing the robustness of models to manifestation shift.

DiffEdit addresses these issues by editing only inside an automatically predicted mask m_{edit}. However, these predicted masks often mismatch the manually annotated ground-truth, especially for small and complex abnormalities like pneumothorax[1] (see Sec. 5.5). Moreover, spurious correlations learned by the diffusion model can lead to the inclusion of support devices in the automatically predicted masks. Furthermore, even when relying on manually annotated masks, DiffEdit can introduce sharp discrepancies at mask boundaries, leading to unrealistic artefacts, such as when adding consolidation that should partially occlude the lung border (Fig. 10b in the Appendix).

4 Method

Our objective is to create synthetic test data through image editing that simulates specific data shifts, to rigorously evaluate biomedical imaging models. This synthetic data is used to predict model robustness, eliminating need for additional real-world test data.

4.1 Improved editing with RadEdit

To address the issues outlined in Sec. 3.4, we propose RadEdit: by introducing 'keep' and 'edit' masks into the editing process, RadEdit explicitly specifies which areas must remain unchanged (keep) and which should be actively modified based on the conditioning signal (edit). Crucially, these masks need not be mutually exclusive, allowing changes in the unmasked regions to ensure global consistency. Using masks, we assume that spurious correlations are mostly non-overlapping [44].

Algorithm 1 RadEdit (ours) uses multiple masks to decouple spurious correlations

Require: original image x_0, inversion prompt c_{inv}, editing prompt c, edit mask m_{edit}, keep mask m_{keep}, CFG weight w, diffusion model f_θ

$(\hat{x}_{1:T}, z_{1:T}) \leftarrow \text{DDPMINVERSION}(x_0, c_{\text{inv}})$ ▷ *Encode image. Procedure in Appendix B*

$x_T \leftarrow \hat{x}_T$

for $t \leftarrow T$ to 1 **do**

- $\epsilon_{\text{cond},t} \leftarrow f_\theta(x_t, t, c)$ ▷ *Predict conditional noise*
- $\epsilon_{\text{uncond},t} \leftarrow f_\theta(x_t, t, c = \emptyset)$ ▷ *Predict unconditional noise*
- $\epsilon_t \leftarrow \epsilon_{\text{uncond},t} + w(\epsilon_{\text{cond},t} - \epsilon_{\text{uncond},t})$ ▷ *Combine noise predictions with CFG*
- $\epsilon_t \leftarrow m_{\text{edit}} \odot \epsilon_t + (1 - m_{\text{edit}}) \odot \epsilon_{\text{uncond},t}$ ▷ *Use CFG only within m_{edit}*
- $x_{t-1} \leftarrow \hat{\mu}_t(x_t, \epsilon_t) + \sigma_t z_t$ ▷ *Move to next time step*
- $x_{t-1} \leftarrow m_{\text{keep}} \odot \hat{x}_{t-1} + (1 - m_{\text{keep}}) \odot x_{t-1}$ ▷ *Undo edits within m_{keep}*

return edited version of x_0

RadEdit is detailed in Algorithm 1, where a number of key properties make RadEdit more suitable for biomedical image editing than prior editing methods. Firstly, since we aim to edit only within the edit mask m_{edit}, classifier-free guidance (CFG) [26] is used only within this region, with high guidance values (following [29], we use a value of 15) ensuring that pathologies are completely removed without drastically changing the rest of the image. This approach also simplifies choosing a prompt for editing since we do not have to take into account the effect of the prompt on the rest of the image. Secondly, we allow the area outside m_{edit} to be modified via unconditional generation to ensure image consistency. Lastly, the region of the keep mask m_{keep} is reverted to the encoding, ensuring that this region remains the same. Instead of initiating our generating process from pure noise we set $x_T = \hat{x}_T$, where $\hat{x}_T$ is the last output of the DDPM inversion.

In Fig. 3c, 10c, we show that RadEdit enables artefact-free editing while preserving structures of interest. Because the anatomical layout remains intact after editing, masks still correspond to the same structures, therefore the same masks can be reused to stress-test segmentation models (Sec. 5.4). In practice, we use a latent diffusion model [57], therefore all operations in Algorithm 1 are performed in the latent space of a variational autoencoder (VAE) [57]; this does not limit the generality of the approach.

4.2 Using synthetic images to uncover bias

Despite advancements in biomedical computer vision, recent studies have shown that bias in training and test data can lead to unrealistically high performance of machine learning models on the test set [59, 11]. In our experiments, we use RadEdit to create high quality synthetic test datasets that realistically capture specific dataset shifts, allowing us to quantify the robustness of models to these dataset shifts. By using masks, we can precisely edit the original training data to represent either acquisition shift, manifestation shift, or population shift [7] (Secs. 5.2 to 5.4). These synthetic test sets are used to stress-test (potentially biased) biomedical vision models by comparing performance to the real (biased) test set; a significant drop in performance indicates that the vision model is not robust to the dataset shift that can occur in clinical settings. This serves as a complementary tool to visual explainable AI tools like Grad-CAM [63] and saliency maps [66, 1], which offer qualitative insight into the robustness of models.

4.3 BioViL-T editing score

Since generative models result in samples of varying quality, poor-quality samples can be filtered out using image–text alignment scores, which quantitatively assess how closely related image–text pairs are via a pre-trained model that embeds similar images and text to nearby vectors [4, 54, 53, 14]. For image editing, we instead assess how similar the change in text and image embeddings are after editing: for a real image–text pair $(I_{\text{real}}, T_{\text{real}})$, edited image–text pair $(I_{\text{edit}}, T_{\text{edit}})$, image encoder E_I, and text encoder E_T, the editing score is defined based on directional similarity [17]:

$$S_{\text{BioViL-T}} = \frac{\Delta I \cdot \Delta T}{\|\Delta I\|\|\Delta T\|}, \quad \text{where} \quad \begin{array}{l} \Delta I = E_I(I_{\text{edit}}) - E_I(I_{\text{real}}), \text{ and} \\ \Delta T = E_T(T_{\text{edit}}) - E_T(T_{\text{real}}). \end{array} \tag{4}$$

Given the focus on biomedical data, we use the BioViL-T [5] image and text encoders: domain-specific vision–language models trained to analyse chest X-rays and radiology reports, therefore well suited to measure changes in the edited image, such as removed pathologies. Following Prabhu et al. [51], we discard images with $S_{\text{BioViL-T}} < 0.2$. This is not only effective for filtering out poor quality edits but is also able to detect whether the original image I_{real} does not match the original text description T_{real} well.

5 Experiments

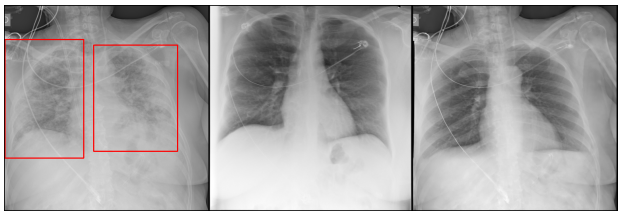
5.1 Diffusion model

Our editing method is heavily dependent on a latent diffusion model [57] that can generate realistic chest X-rays. We use the VAE [37, 25] of SDXL [50] which can adequately reconstruct chest X-rays [8]. The VAE is frozen, and the denoising U-Net is trained on three datasets downsampled and centre-cropped to 512×512 pixels: MIMIC-CXR [33], ChestX-ray8 [75], and CheXpert [30], totalling 487 680 training images. This data diversity allows us to perform *zero-shot edits* on datasets not seen during training.

For MIMIC-CXR, we only include frontal view chest X-rays, and condition the denoising U-Net on the corresponding impression section in the radiology report (a short clinically actionable outline of the main findings). We employ the tokeniser and frozen text encoder from BioViL-T [5]. For ChestX-ray8 and CheXpert, we condition on a list of all abnormalities present in an image as indicated by the labels, e.g., *'Cardiomegaly. Pneumothorax.'*. If the list of abnormalities is empty, we use the string *'No findings'*. More details on the datasets and diffusion model training can be found in Appendix E.

5.2 Acquisition shift

Background To show how RadEdit can be used to quantify the robustness of models to acquisition shift, we closely follow the experimental setup of DeGrave et al. [11], who show that deep learning systems built to detect COVID-19 from chest X-rays rely on confounding factors rather than pathology features. This problem arises when COVID-19-positive and -negative images come from disparate sources. In our setup, all COVID-19-positive cases come from the BIMCV dataset [73] (denoted BIMCV+),



(a) Original Image **(b)** LANCE [51] **(c)** RadEdit (ours)

Fig. 2: Removing COVID-19 features with LANCE[2] (b) also changes the laterality markers and reduces contrast. In contrast, RadEdit (c; ours) preserves anatomical structures and laterality markers, and retains the original contrast.

and all COVID-19-negative cases from MIMIC-CXR [33] (see Fig. 11). A classifier trained on these datasets will rely on spurious features indicative of the data's origin, e.g., laterality markers or the field of view, instead of features caused by the pathology.

Setup A synthetic test set is created by applying RadEdit to remove COVID-19 features[1] from BIMCV+ images using the prompt *'No acute cardiopulmonary process'*[3] (Fig. 2); the included bounding boxes of COVID-19 features are used as the edit mask m_{edit}. Since this is the only mask available, we set the keep mask as $m_{\text{keep}} = 1 - m_{\text{edit}}$. After filtering using the BioViL-T editing (Sec. 4.3), this results in a synthetic dataset of 2774 COVID-19-negative images containing the same spurious features as BIMCV+. Both the weak and strong predictor models are classifiers trained using a ResNet50 architecture (see Appendix F for more implementation details).

Findings Tab. 1, shows the performance of a COVID-19 classifier (weak predictor) trained on BIMCV+ and MIMIC-CXR. In accordance with DeGrave et al. [11], we find that the weak predictor performs exceptionally well on the real test set (i.e. test splits of both datasets) since the model learned to distinguish the two data sources instead of learning visual features related to COVID-19. However, in the second row of Tab. 1, we see a drop of 95% in accuracy meaning that the model fails to classify the synthetic images as COVID-19-negative. The weak predictor is not robust to a shift in acquisition.

To show that the decreased performance of the weak predictor is not caused by artefacts in the edited images, we train a more robust COVID-19 classifier (strong predictor), using the BIMCV+ and BIMCV- datasets, as in [11], where the BIMCV- dataset consists of only COVID-19-negative cases from BIMCV, and test on the same two test

Table 1: Quantifying robustness of COVID-19 detectors to acquisition shift. We train a weak predictor on the 'Biased' dataset—a combination of BIMCV+ [73] and MIMIC-CXR [33]; and a strong predictor on an unbiased dataset—a combination of BIMCV+ and BIMCV-; the 'Synthetic' test set consists of 2774 COVID-19-negative images with the same spurious features as the BIMCV+ datasets, e.g. laterality markers. We report mean accuracy and standard deviation across 5 runs.

Predictor	Test data	Accuracy
Weak	Biased	99.1 ± 0.2
Weak	Synthetic	5.5 ± 2.1
Strong	Biased	74.4 ± 3.0
Strong	Synthetic	76.0 ± 7.7

[2]For LANCE, we perform the text perturbation manually.

[3]This is a common radiological description of a 'normal' chest X-ray.

(a) Original Image **(b)** LANCE [51] **(c)** RadEdit (ours)

Fig. 3: Removing pneumothorax (red) with LANCE[2] (b) also removes the spuriously correlated chest drain (blue) and reduces contrast. In contrast, RadEdit (c; ours) preserves the chest drain and better preserves anatomical structures.

datasets. Comparing rows one and three of Tab. 1, we find that the strong predictor performs worse on the test set containing samples from BIMCV+ and MIMIC-CXR than the weak predictor (row one). This is expected as the strong predictor relies on actual pathology features. Lastly, rows three and four of Tab. 1 show that the strong predictor performs similarly on the real and synthetic test sets, attesting to the quality of our edits.

5.3 Manifestation shift

Background In this section, we show how RadEdit can be used to quantify the robustness of biomedical vision models to manifestation shift. We closely follow the experimental setup of Rueckel et al. [59], who demonstrate that pneumothorax[1] classification models are strongly biased by the presence of chest drains: while the average performance of pneumothorax classifiers is high, performance on the subset of images with a chest drain but no pneumothorax is significantly lower. This is due to chest drains being a common treatment for pneumothorax, resulting in the majority of images in datasets like CANDID-PTX [13] containing a chest drain only if there is a pneumothorax. As a result, only 1% of images in CANDID-PTX contain a chest drain but no pneumothorax.

Setup We use RadEdit to create a synthetic dataset containing images with chest drains but no pneumothorax, by editing out the pneumothorax from CANDID-PTX images using the prompt *'No acute cardiopulmonary process'*[3] (Fig. 3). The edit mask m_{edit} is set as a mask of the pneumothorax, and the keep mask m_{keep} set as the chest drain mask. This ensures that the chest drain is not removed, while preventing border artefacts. After filtering using the BioViL-T editing score (Sec. 4.3), 628 images are left; in contrast, the real test set contains only 16 of cases with drains but no pneumothorax. The weak predictor is a ResNet50 trained to classify pneumothorax. Following Rueckel et al. [59], the strong predictor is an EfficientNet U-Net [69] trained on SIIM-ACR [77] to both segment pneumothorax and classify pneumothorax (more details in Appendix G).

Findings In accordance with [59], we show in Tab. 2 that a pneumothorax classifier (weak predictor) trained on CANDID-PTX performs exceptionally well on the test split of CANDID-PTX, since very few images contain a chest drain and no pneumothorax. However, in row two of Tab. 2, we show a drastic drop in performance on the synthetic test set, i.e., the weak predictor is not robust to manifestation shift. To show that the drop in performance on the synthetic dataset does not come from editing artefacts, we evaluate the more robust segmentation model (strong predictor) on the same test datasets

Table 2: Quantifying robustness of pneumothorax detectors to manifestation shift. The weak predictor is trained on the biased CANDID-PTX [13] dataset to classify pneumothorax; the strong predictor is trained on SIIM-ACR [77] to classify and segment the pneumothorax. Real 'Biased' test data comes from CANDID-PTX which exhibits strong confounding between the pneumothorax and chest tubes; 'Synthetic' test data is 629 solely edited images containing chest drains but no pneumothorax. We report mean accuracy and standard deviation across 5 runs.

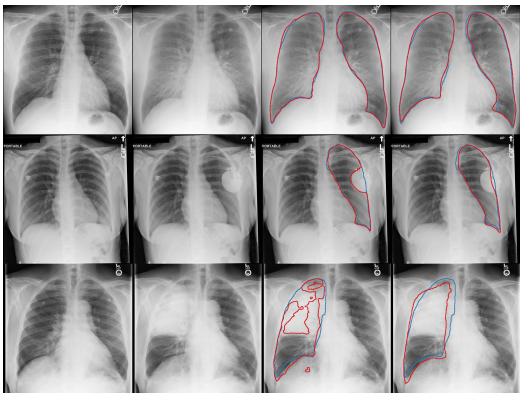
Predictor	Test data	Accuracy
Weak	Biased	93.3 ± 0.6
Weak	Synthetic	17.9 ± 3.7
Strong	Biased	93.7 ± 1.3
Strong	Synthetic	81.7 ± 7.1

(rows three and four of Tab. 2), we find that the strong predictor performs on par with the weak predictor in row one; however, the strong predictor closes the majority of the gap between the real test set and the synthetic one, attesting to the quality of our edits. In agreement with Rueckel et al. [59], there is still a performance gap, indicating that the strong predictor still suffers from mild manifestation shift.

5.4 Population shift

Background In this section, we show how RadEdit can be used to quantify the robustness of lung segmentation models to population shifts. Manually segmenting X-ray images is labour intensive and requires high expertise, leading to small datasets often limited to single pathologies or healthy patients [65, 31], e.g., MIMIC-Seg [9]. These models are thus sensitive to occlusions such as medical devices or pathologies, which typically appear as white regions on X-rays [43]. Evaluating model robustness requires further image collection for each occlusion type, which is time-consuming and costly.

Setup Here, abnormalities are added to the lung region in healthy X-rays from MIMIC-Seg (Fig. 4). The edit mask m_{edit} is set as a mask of the lung(s). When editing a single



(a) Original **(b)** Edited **(c)** Weak predictor **(d)** Strong predictor

Fig. 4: Adding pulmonary edema (top), pacemakers (middle), and consolidation (bottom) with RadEdit. The 'strong predictor' (d), a segmentation model trained on CheXmask [16] (a large dataset containing various abnormalities) is more robust to these abnormalities than the 'weak predictor' (c), a segmentation model trained on MIMIC-Seg [9] (a small set of mostly healthy patients): the weak predictor traces around the pacemaker and poorly annotates the consolidated lung. Blue: ground-truth annotation; red: predicted segmentation.

Table 3: Quantifying robustness of lung segmentation models to population shift. The 'weak predictor' is trained on MIMIC-Seg (a small set of predominantly healthy patients); the 'strong predictor' is trained on CheXmask (a large mixed set of patients with various abnormalities). Synthetic test data is created by using RadEdit to add edema, pacemakers, and consolidation. We report the change (Δ) in Dice score and AHD with respect to the segmentation models evaluated on the ground-truth test set.

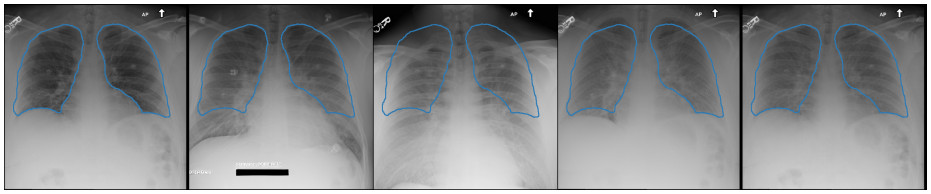
	Weak Predictor				Strong Predictor			
Test data	**Dice↑**	**Δ↓**	**AHD↓**	**Δ↓**	**Dice↑**	**Δ↓**	**AHD↓**	**Δ↓**
Real data	97.4	—	6.1	—	95.5	—	11.6	—
Healthy $\overset{\text{edit}}{\rightarrow}$ edema	93.8	3.6	21.8	15.7	93.9	**1.6**	22.8	**11.2**
Healthy $\overset{\text{edit}}{\rightarrow}$ pacemaker	85.0	12.4	49.8	43.7	87.3	**8.2**	29.5	**17.9**
Healthy $\overset{\text{edit}}{\rightarrow}$ consolidation	85.9	11.5	44.1	38.1	88.1	**7.4**	29.4	**17.8**

lung, the keep mask m_{keep} corresponds to the lung which must not change, while when editing both lungs we set $m_{\text{keep}} = 0$ to allow opacity adjustments, or for elements to be added outside of the lungs. Stress-test sets are generated for three abnormalities: pulmonary edema, pacemakers, and consolidation[1]. Prompts are phrased to match similar impressions in the training data (see Appendix H). Both predictors are EfficientNet U-Net [69] models; the weak predictor is trained on MIMIC-Seg, while the strong predictor is trained on CheXmask [16], a larger dataset with various lung abnormalities (more details in Appendix H). We evaluate segmentation quality using Dice similarity coefficient, which is the harmonic mean of the precision and recall, and 95th percentile average Hausdorff distance (AHD), a measure of the distance between two sets [45].

Findings Tab. 3 shows that the weak predictor model performs well on the real biased test data, since it is mostly composed of healthy subjects. However, testing on the synthetic lung abnormality datasets (rows two to four), causes performance to drop substantially, i.e. the weak predictor is not robust to population shift. To show that this drop in performance does not come from editing artefacts, we evaluate the strong predictor on the synthetic test sets and see considerably smaller changes in performance. This can be seen in Fig. 4: for pulmonary edema, both models can accurately segment, despite the abnormality; for pacemakers, the weak predictor incorrectly segments around the pacemakers, while the strong predictor more accurately segments the lungs; and for consolidation, both models are less able to segment the lungs accurately, however, the strong predictor gets closer to the ground-truth. See Appendix H for more examples.

5.5 Quantifying the limitations of existing editing methods

LANCE As seen in the second row of Tab. 3, adding edema leads only to a small drop in performance of the strong predictor. We hypothesise that further drops in performances stem from a mismatch of the original mask and the edited images. We therefore use this setup to quantify how well LANCE and RadEdit preserve the shape and position of the lung borders. Additionally, we study the difference between results using



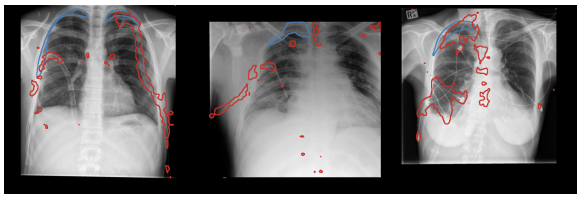
(a) (b) (c) (d) (e)

Fig. 5: Comparison of LANCE[2] and RadEdit. We measure how well the strong predictor from Tab. 3's outputs matches the ground-truth lung masks (blue) for four synthetic datasets created by adding edema using LANCE and RadEdit with DDIM or DDPM inversion. High Dice / low AHD indicates that the editing method well preserves the lung border's location and shape.

Editing Method	Dice↑	AHD↓
(a) Original data	95.5	11.6
(b) LANCE w/ DDIM	78.9	65.1
(c) LANCE w/ DDPM	80.1	69.5
(d) RadEdit w/ DDIM	86.2	39.8
(e) RadEdit w/ DDPM	**93.9**	**22.8**

DDIM or DDPM inversion. For all four methods in Fig. 5, we use the same setup as in Sec. 5.4: we first edit the original image with the prompt *'Moderate pulmonary edema. The heart size is normal'*, and then compare the outputs of the strong predictor with the original ground-truth lung masks. Here, we find that using masks and DDPM inversion is necessary for RadEdit to preserve the shape and position of the lung border.

DiffEdit We quantify how well DiffEdit's automatically predicted masks match the manual ground-truth using the same setup as in Sec. 5.3: we take an image containing pneumothorax and a chest drain, and try to remove only the pneumothorax. The editing prompt is created by keeping only the parts of the impressions related to pneumothorax and chest drains, and replacing the description of the pneumothorax with *'No pneumothorax'*. DiffEdit should therefore predict a mask containing only the pneumothorax. We perform a grid search on the MIMIC-Seg [9] validation set over DiffEdit's hyperparameters (noise strength and binarising threshold) to optimise pneumothorax segmentation metrics, then evaluate on the training set. In Fig. 6 we see that DiffEdit's predicted masks obtain poor quantitative metrics where parts of the pneumothorax are often missing, and the spuriously correlated chest drain is often included in the predicted mask. As a result, DiffEdit's predicted masks are unsuitable for stress-testing.



(a) Examples of pneumothorax masks predicted using DiffEdit [10]. Blue: ground-truth annotation; red: predicted editing mask.

Hyperparameters	Dice	AHD
Tuned per image	33.8	97.7
Tuned on validation	18.4	256.8

(b) Segmentation metrics for the pneumothorax mask predicted by DiffEdit [10], for hyperparameters tuned on the validation set (bottom) and tuned per image (top; which requires ground-truth masks).

Fig. 6: Evaluating pneumothorax masks predicted using DiffEdit [10]. (a) Predicted masks (red) are noisy, with chest drains often incorrectly segmented as well as or instead of the pneumothorax (blue); (b) this is demonstrated quantitatively with low Dice score and high AHD.

6 Limitations and future work

Despite the encouraging results presented in the paper, RadEdit is not without limitations and more work is needed to extend it to more applications. Currently, training datasets and models must be manually analysed to predict potential failure cases, simulate these failures to test the hypothesis, and finally quantitatively evaluate the model; future work could automate such failure mode discovery. Another limitation is that current editing techniques do not enable all types of stress-testing; for example, with current approaches, we are unable to test segmentation models' behaviour to cardiomegaly (enlarged heart) since this would require segmentation maps to be adjusted after editing. However, this could potentially be enabled by enlarging heart segmentations to simulate cardiomegaly and adjusting the ground-truth lung segmentation accordingly.

When using generative editing, it is not possible to guarantee that unwanted changes will not occur. With RadEdit, we minimise this by forcing spuriously correlated regions to remain the same, only using classifier-free guidance within the editing mask, and filtering via image–text alignment. Future work to better maintain structure when editing will help with this issue, but masks will still be necessary to bypass spurious correlations. Furthermore, due to potential overlap of 3D structures, editing 2D X-rays can be limited. In such cases, editing single structures using RadEdit may not be successful. To address this, 3D CT images could be edited, then projected to synthetic X-rays [18].

When producing simulated stress test sets, several factors affect edit quality including classifier-free guidance weight, number of inference steps, and time step to encode to. Additionally, the text encoder must well understand specified pathologies to provide informative features to condition the generative model on; similarly, the diffusion model must be able to capture fine details and well cover the data distribution.

Finally, more research is required to develop better approaches for quantifying edit quality for downstream tasks. In particular, observing a change in downstream performance is not necessarily indicative of real-world performance as edit quality may be poor. While the introduced BioViL-T editing score can be used to quantify edit quality, this introduces reliance on a potentially biased model. Additionally, the BioViL-T editing score is not suited to detect the artefacts introduced by LANCE and DiffEdit.

7 Conclusion

In this study, we illustrate the efficacy of generative image editing as a robust tool for stress-testing biomedical vision models. Our focus is on assessing their robustness against three types of dataset shifts commonly encountered in biomedical imaging: acquisition shift, manifestation shift, and population shift. We highlight that one of the significant challenges in biomedical image editing is the correlations learned by the generative model, which can result in artefacts during the editing process. To mitigate these artefacts, RadEdit relies on various types of masks to restrict the effects of the editing to certain areas while ensuring the consistency of the edited images. This approach enables us to generate high fidelity synthetic test sets that exhibit common dataset shifts. These synthetic test sets are used to identify and quantify the failure modes of biomedical classification and segmentation models. This provides a valuable supplement to explainable AI approaches such as Grad-CAM [63] and saliency maps [66, 1].

Bibliography

[1] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. *Advances in neural information processing systems*, 31, 2018.

[2] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18208–18218, 2022.

[3] Omri Avrahami, Ohad Fried, and Dani Lischinski. Blended latent diffusion. *ACM Transactions on Graphics (TOG)*, 42(4):1–11, 2023.

[4] Samaneh Azadi, Catherine Olsson, Trevor Darrell, Ian Goodfellow, and Augustus Odena. Discriminator rejection sampling. *arXiv preprint arXiv:1810.06758*, 2018.

[5] Shruthi Bannur, Stephanie Hyland, Qianchu Liu, Fernando Perez-Garcia, Maximilian Ilse, Daniel C. Castro, Benedikt Boecking, Harshita Sharma, Kenza Bouzid, Anja Thieme, Anton Schwaighofer, Maria Wetscherek, Matthew P. Lungren, Aditya Nori, Javier Alvarez-Valle, and Ozan Oktay. Learning to Exploit Temporal Structure for Biomedical Vision-Language Processing. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 1 2023. `https://doi.org/10.48550/arxiv.2301.04558`.

[6] Riccardo Barbano, Alexander Denker, Hyungjin Chung, Tae Hoon Roh, Simon Arrdige, Peter Maass, Bangti Jin, and Jong Chul Ye. Steerable conditional diffusion for out-of-distribution adaptation in imaging inverse problems. *arXiv preprint arXiv:2308.14409*, 2023.

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

[8] Pierre Chambon, Christian Bluethgen, Jean-Benoit Delbrouck, Rogier Van der Sluijs, Małgorzata Połacin, Juan Manuel Zambrano Chaves, Tanishq Mathew Abraham, Shivanshu Purohit, Curtis P Langlotz, and Akshay Chaudhari. Roentgen: vision-language foundation model for chest x-ray generation. *arXiv preprint arXiv:2211.12737*, 2022.

[9] Li-Ching Chen, Po-Chih Kuo, Ryan Wang, Judy Gichoya, and Leo Anthony Celi. Chest X-ray segmentation images based on MIMIC-CXR (version 1.0.0). PhysioNet, 2022.

[10] Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. DiffEdit: Diffusion-based semantic image editing with mask guidance, 2022.

[11] Alex J. DeGrave, Joseph D. Janizek, and Su-In Lee. AI for radiographic COVID-19 detection selects shortcuts over signal. *Nature Machine Intelligence*, 3(7):610–619, 2021. ISSN 2522-5839. `https://doi.org/10.1038/s42256-021-00338-7`.

[12] Alexey Dosovitskiy, Jost Tobias Springenberg, and Thomas Brox. Learning to generate chairs with convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1538–1546, 2015.

[13] Sijing Feng, Damian Azzollini, Ji Soo Kim, Cheng-Kai Jin, Simon P. Gordon, Jason Yeoh, Eve Kim, Mina Han, Andrew Lee, Aakash Patel, Joy Wu, Martin

Urschler, Amy Fong, Cameron Simmers, Gregory P. Tarr, Stuart Barnard, and Ben Wilson. Curation of the CANDID-PTX dataset with free-text reports. *Radiology: Artificial Intelligence*, 3(6):e210136, 2021. ISSN 2638-6100. https://doi.org/10.1148/ryai.2021210136.

[14] Virginia Fernandez, Pedro Sanchez, Walter Hugo Lopez Pinaya, Grzegorz Jacenków, Sotirios A Tsaftaris, and Jorge Cardoso. Privacy distillation: Reducing re-identification risk of multimodal diffusion models. *arXiv preprint arXiv:2306.01322*, 2023.

[15] Alessandro Fontanella, Grant Mair, Joanna Wardlaw, Emanuele Trucco, and Amos Storkey. Diffusion models for counterfactual generation and anomaly detection in brain images. URL http://arxiv.org/abs/2308.02062.

[16] Nicolás Gaggion, Candelaria Mosquera, Lucas Mansilla, Martina Aineseder, Diego H Milone, and Enzo Ferrante. CheXmask: a large-scale dataset of anatomical segmentation masks for multi-center chest X-ray images. *arXiv preprint arXiv:2307.03293*, 2023.

[17] Rinon Gal, Or Patashnik, Haggai Maron, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. StyleGAN-NADA: CLIP-guided domain adaptation of image generators. *ACM Transactions on Graphics*, 41(4), 2022. https://doi.org/10.1145/3528223.3530164.

[18] Cong Gao, Benjamin D Killeen, Yicheng Hu, Robert B Grupp, Russell H Taylor, Mehran Armand, and Mathias Unberath. Synthetic data accelerates the development of generalizable learning-based algorithms for x-ray image analysis. *Nature Machine Intelligence*, 5(3):294–308, 2023.

[19] Timur Garipov, Sebastiaan De Peuter, Ge Yang, Vikas Garg, Samuel Kaski, and Tommi Jaakkola. Compositional sculpting of iterative generative processes. *arXiv preprint arXiv:2309.16115*, 2023.

[20] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples, 2015.

[21] Yu Gu, Jianwei Yang, Naoto Usuyama, Chunyuan Li, Sheng Zhang, Matthew P Lungren, Jianfeng Gao, and Hoifung Poon. Biomedjourney: Counterfactual biomedical image generation by instruction-learning from multimodal patient journeys. *arXiv preprint arXiv:2310.10765*, 2023.

[22] Will Douglas Heaven. Hundreds of ai tools have been built to catch covid. none of them helped. *MIT Technology Review. Retrieved December 2023*, 2021.

[23] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2018.

[24] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-or. Prompt-to-prompt image editing with cross-attention control. In *The Eleventh International Conference on Learning Representations*, 2022.

[25] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. β-VAE: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, 2016.

[26] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance, 2022.

[27] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020.

[28] Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *The Journal of Machine Learning Research*, 23(1):2249–2281, 2022.

[29] Inbar Huberman-Spiegelglas, Vladimir Kulikov, and Tomer Michaeli. An edit friendly DDPM noise space: Inversion and manipulations, 2023.

[30] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison, 2019.

[31] Stefan Jaeger, Sema Candemir, Sameer Antani, Yì-Xiáng J Wáng, Pu-Xuan Lu, and George Thoma. Two public chest x-ray datasets for computer-aided screening of pulmonary diseases. *Quantitative imaging in medicine and surgery*, 4(6):475, 2014.

[32] Priyank Jaini, Kevin Clark, and Robert Geirhos. Intriguing properties of generative classifiers. *arXiv preprint arXiv:2309.16779*, 2023.

[33] Alistair E. W. Johnson, Tom J. Pollard, Seth J. Berkowitz, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih-ying Deng, Roger G. Mark, and Steven Horng. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6(1):317, 2019. ISSN 2052-4463. `https://doi.org/10.1038/s41597-019-0322-0`.

[34] Charles Jones, Daniel C Castro, Fabio De Sousa Ribeiro, Ozan Oktay, Melissa McCradden, and Ben Glocker. No fair lunch: A causal perspective on dataset bias in machine learning for medical imaging. *arXiv preprint arXiv:2307.16526*, 2023.

[35] Minguk Kang, Jun-Yan Zhu, Richard Zhang, Jaesik Park, Eli Shechtman, Sylvain Paris, and Taesung Park. Scaling up gans for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10124–10134, 2023.

[36] Priyatham Kattakinda, Alexander Levine, and Soheil Feizi. Invariant learning via diffusion dreamed distribution shifts. *arXiv preprint arXiv:2211.10370*, 2022.

[37] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes, 2022.

[38] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pages 5637–5664. PMLR, 2021.

[39] Ira Ktena, Olivia Wiles, Isabela Albuquerque, Sylvestre-Alvise Rebuffi, Ryutaro Tanno, Abhijit Guha Roy, Shekoofeh Azizi, Danielle Belgrave, Pushmeet Kohli, Alan Karthikesalingam, et al. Generative models improve fairness of medical classifiers under distribution shifts. *arXiv preprint arXiv:2304.09218*, 2023.

[40] Agostina J Larrazabal, Nicolás Nieto, Victoria Peterson, Diego H Milone, and Enzo Ferrante. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proceedings of the National Academy of Sciences*, 117(23):12592–12594, 2020.

[41] Choong Ho Lee and Hyung-Jin Yoon. Medical big data: promise and challenges. *Kidney research and clinical practice*, 36(1):3, 2017.

[42] Xiaodan Li, Yuefeng Chen, Yao Zhu, Shuhui Wang, Rong Zhang, and Hui Xue. ImageNet-e: Benchmarking neural network robustness via attribute editing, 2023.

[43] Wufeng Liu, Jiaxin Luo, Yan Yang, Wenlian Wang, Junkui Deng, and Liang Yu. Automatic lung segmentation in chest x-ray images using improved u-net. *Scientific Reports*, 12(1):8649, 2022.

[44] Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. *Advances in Neural Information Processing Systems*, 33:11525–11538, 2020.

[45] Lena Maier-Hein, Annika Reinke, Patrick Godau, Minu D. Tizabi, Florian Buettner, Evangelia Christodoulou, Ben Glocker, Fabian Isensee, Jens Kleesiek, Michal Kozubek, Mauricio Reyes, Michael A. Riegler, Manuel Wiesenfarth, A. Emre Kavur, Carole H. Sudre, Michael Baumgartner, Matthias Eisenmann, Doreen Heckmann-Nötzel, A. Tim Rädsch, Laura Acion, Michela Antonelli, Tal Arbel, Spyridon Bakas, Arriel Benis, Matthew Blaschko, M. Jorge Cardoso, Veronika Cheplygina, Beth A. Cimini, Gary S. Collins, Keyvan Farahani, Luciana Ferrer, Adrian Galdran, Bram van Ginneken, Robert Haase, Daniel A. Hashimoto, Michael M. Hoffman, Merel Huisman, Pierre Jannin, Charles E. Kahn, Dagmar Kainmueller, Bernhard Kainz, Alexandros Karargyris, Alan Karthikesalingam, Hannes Kenngott, Florian Kofler, Annette Kopp-Schneider, Anna Kreshuk, Tahsin Kurc, Bennett A. Landman, Geert Litjens, Amin Madani, Klaus Maier-Hein, Anne L. Martel, Peter Mattson, Erik Meijering, Bjoern Menze, Karel G. M. Moons, Henning Müller, Brennan Nichyporuk, Felix Nickel, Jens Petersen, Nasir Rajpoot, Nicola Rieke, Julio Saez-Rodriguez, Clara I. Sánchez, Shravya Shetty, Maarten van Smeden, Ronald M. Summers, Abdel A. Taha, Aleksei Tiulpin, Sotirios A. Tsaftaris, Ben Van Calster, Gaël Varoquaux, and Paul F. Jäger. Metrics reloaded: Recommendations for image analysis validation, 2023.

[46] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Guided image synthesis and editing with stochastic differential equations, 2022.

[47] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6038–6047, 2023.

[48] Gustav Müller-Franzes, Jan Moritz Niehues, Firas Khader, Soroosh Tayebi Arasteh, Christoph Haarburger, Christiane Kuhl, Tianci Wang, Tianyu Han, Sven Nebelung, Jakob Nikolas Kather, et al. Diffusion probabilistic models beat gans on medical images. *arXiv preprint arXiv:2212.07501*, 2022.

[49] Nick Pawlowski, Daniel C. Castro, and Ben Glocker. Deep structural causal models for tractable counterfactual inference. In *Advances in Neural Information Processing Systems*, volume 33, pages 857–869, 2020.

[50] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis, 2023.

[51] Viraj Prabhu, Sriram Yenamandra, Prithvijit Chattopadhyay, and Judy Hoffman. LANCE: Stress-testing visual models by generating language-guided counterfactual images, 2023.

[52] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

[53] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[54] Ali Razavi, Aaron Van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. *Advances in neural information processing systems*, 32, 2019.

[55] Jacob C. Reinhold, Aaron Carass, and Jerry L. Prince. A structural causal model for MR images of multiple sclerosis. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, volume 12905 of *LNCS*, pages 782–792, 2021. https://doi.org/10.1007/978-3-030-87240-3_75.

[56] Michael Roberts, Derek Driggs, Matthew Thorpe, Julian Gilbey, Michael Yeung, Stephan Ursprung, Angelica I Aviles-Rivero, Christian Etmann, Cathal McCague, Lucian Beer, et al. Common pitfalls and recommendations for using machine learning to detect and prognosticate for covid-19 using chest radiographs and ct scans. *Nature Machine Intelligence*, 3(3):199–217, 2021.

[57] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022.

[58] Johannes Rueckel, Lena Trappmann, Balthasar Schachtner, Philipp Wesp, Boj Friedrich Hoppe, Nicola Fink, Jens Ricke, Julien Dinkel, Michael Ingrisch, and Bastian Oliver Sabel. Impact of confounding thoracic tubes and pleural dehiscence extent on artificial intelligence pneumothorax detection in chest radiographs. *Investigative Radiology*, 55(12):792–798, July 2020. ISSN 0020-9996. https://doi.org/10.1097/rli.0000000000000707.

[59] Johannes Rueckel, Christian Huemmer, Andreas Fieselmann, Florin-Cristian Ghesu, Awais Mansoor, Balthasar Schachtner, Philipp Wesp, Lena Trappmann, Basel Munawwar, Jens Ricke, Michael Ingrisch, and Bastian O. Sabel. Pneumothorax detection in chest radiographs: optimizing artificial intelligence system for accuracy and confounding bias reduction using in-image annotations in algorithm training. *European Radiology*, 31(10):7888–7900, 2021. https://doi.org/10.1007/s00330-021-07833-w.

[60] Chitwan Saharia, William Chan, Huiwen Chang, Chris A. Lee, Jonathan Ho, Tim Salimans, David J. Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models, 2022.

[61] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Semantic foggy scene understanding with synthetic data. *International Journal of Computer Vision*, 126: 973–992, 2018.

[62] Pedro Sanchez, Antanas Kascenas, Xiao Liu, Alison Q O'Neil, and Sotirios A Tsaftaris. What is healthy? generative counterfactual diffusion for lesion localiza-

tion. In *MICCAI Workshop on Deep Generative Models*, pages 34–44. Springer, 2022.

[63] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[64] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of gans for semantic face editing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9243–9252, 2020.

[65] Junji Shiraishi, Shigehiko Katsuragawa, Junpei Ikezoe, Tsuneo Matsumoto, Takeshi Kobayashi, Ken-ichi Komatsu, Mitate Matsui, Hiroshi Fujita, Yoshie Kodera, and Kunio Doi. Development of a digital image database for chest radiographs with and without a lung nodule: receiver operating characteristic analysis of radiologists' detection of pulmonary nodules. *American Journal of Roentgenology*, 174(1):71–74, 2000.

[66] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

[67] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015.

[68] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models, 2022.

[69] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.

[70] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1921–1930, 2023.

[71] Paul Upchurch, Jacob Gardner, Geoff Pleiss, Robert Pless, Noah Snavely, Kavita Bala, and Kilian Weinberger. Deep feature interpolation for image content changes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7064–7073, 2017.

[72] Boris Van Breugel, Nabeel Seedat, Fergus Imrie, and Mihaela van der Schaar. Can you rely on your model evaluation? improving model evaluation with synthetic test data, 2023.

[73] Maria de la Iglesia Vayá, Jose Manuel Saborit, Joaquim Angel Montell, Antonio Pertusa, Aurelia Bustos, Miguel Cazorla, Joaquin Galant, Xavier Barber, Domingo Orozco-Beltrán, Francisco García-García, Marisa Caparrós, Germán González, and Jose María Salinas. BIMCV COVID-19+: a large annotated dataset of RX and CT images from COVID-19 patients, 2020. version: 3.

[74] I von Borzyskowski, A Mazumder, B Mateen, and M Wooldridge. Data science and ai in the age of covid-19, 2021.

[75] 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 *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3462–3471, 2017. https://doi.org/10.1109/CVPR.2017.369.

[76] Laure Wynants, Ben Van Calster, Gary S Collins, Richard D Riley, Georg Heinze, Ewoud Schuit, Elena Albu, Banafsheh Arshi, Vanesa Bellou, Marc MJ Bonten, et al. Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. *bmj*, 369, 2020.

[77] Anna Zawacki, Carol Wu, George Shih, Julia Elliott, Mikhail Fomitchev, Mohannad ParasLakhani Hussain, Phil Culliton, and Shunxing Bao. Siim-acr pneumothorax segmentation, 2019.