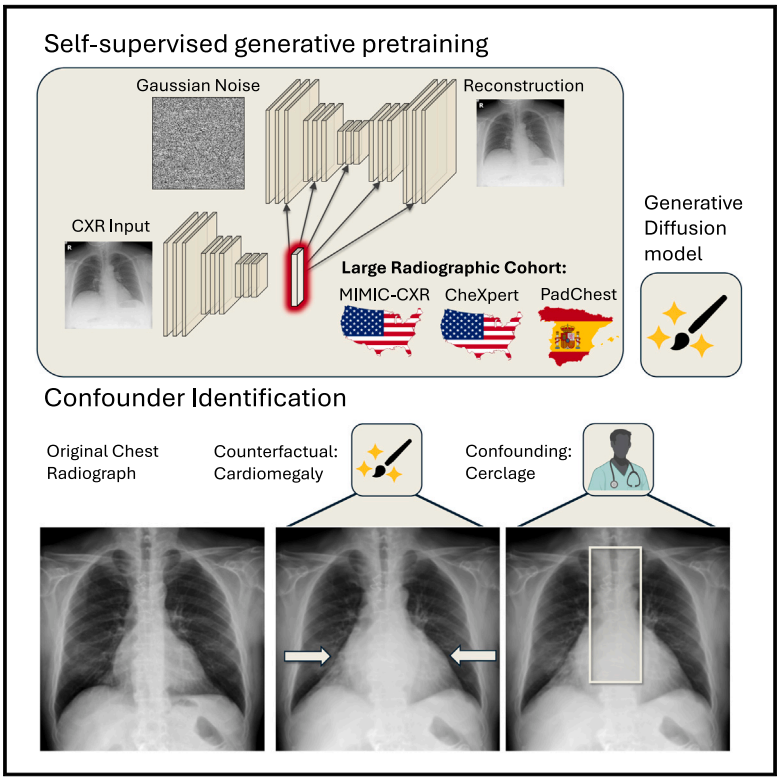
Cell Reports
Medicine

Article

Reconstruction of patient-specific confounders in AI-based radiologic image interpretation using generative pretraining

Graphical abstract



Highlights

- Conditioned diffusion models capture confounders in radiology, boosting reliability
- High diagnostic accuracy with minimal labeled data via generative pretraining
- Generated visual explanations enhance clinician agreement on complex disease grading

Authors

Tianyu Han, Laura Žigutytė, Luisa Huck, ..., Sven Nebelung, Jakob Nikolas Kather, Daniel Truhn

Correspondence

than@ukaachen.de

In brief

Han et al. combine generative diffusion and classification models to enhance AI-driven medical diagnosis. Their method improves explainability, identifies hidden confounders, and optimizes data use, addressing clinical adoption challenges. Validated on more than 500,000 chest radiographs, it offers patient-specific visual explanations and broader diagnostic potential.

Han et al., 2024, Cell Reports Medicine *5*, 101713
September 17, 2024 © 2024 The Author(s). Published by Elsevier Inc.
https://doi.org/10.1016/j.xcrm.2024.101713

CellPress

Cell Reports Medicine

CellPress
OPEN ACCESS

Article

Reconstruction of patient-specific confounders in AI-based radiologic image interpretation using generative pretraining

Tianyu Han,[1,9,*] Laura Žigutytė,[2] Luisa Huck,[1] Marc Sebastian Huppertz,[1] Robert Siepmann,[1] Yossi Gandelsman,[3] Christian Blüthgen,[4,5] Firas Khader,[1] Christiane Kuhl,[1] Sven Nebelung,[1] Jakob Nikolas Kather,[2,6,7,8] and Daniel Truhn[1,8]

[1]Department of Diagnostic and Interventional Radiology, University Hospital RWTH Aachen, 52074 Aachen, Germany
[2]Else Kroener Fresenius Center for Digital Health, Technical University Dresden, 01307 Dresden, Germany
[3]Department of Electrical Engineering and Computer Science, UC Berkeley, Berkeley, CA, USA
[4]Institute for Diagnostic and Interventional Radiology, University Hospital Zurich, 8006 Zurich, Switzerland
[5]Center for Artificial Intelligence in Medicine and Imaging (AIMI), Stanford University, Stanford, CA, USA
[6]Department of Medicine I, University Hospital Dresden, 01307 Dresden, Germany
[7]Medical Oncology, National Center for Tumor Diseases (NCT), University Hospital Heidelberg, 69120 Heidelberg, Germany
[8]These authors contributed equally
[9]Lead contact
*Correspondence: than@ukaachen.de
https://doi.org/10.1016/j.xcrm.2024.101713

SUMMARY

Reliably detecting potentially misleading patterns in automated diagnostic assistance systems, such as those powered by artificial intelligence (AI), is crucial for instilling user trust and ensuring reliability. Current techniques fall short in visualizing such confounding factors. We propose DiffChest, a self-conditioned diffusion model trained on 515,704 chest radiographs from 194,956 patients across the US and Europe. DiffChest provides patient-specific explanations and visualizes confounding factors that might mislead the model. The high inter-reader agreement, with Fleiss' kappa values of 0.8 or higher, validates its capability to identify treatment-related confounders. Confounders are accurately detected with 10%–100% prevalence rates. The pretraining process optimizes the model for relevant imaging information, resulting in excellent diagnostic accuracy for 11 chest conditions, including pleural effusion and heart insufficiency. Our findings highlight the potential of diffusion models in medical image classification, providing insights into confounding factors and enhancing model robustness and reliability.

INTRODUCTION

Confounding factors in medical datasets often lead to spurious correlations, negatively impacting model predictions.[1–4] For instance, neural networks trained on radiological data for detecting pneumothorax—a pathological air collection in the pleural space—may inadvertently learn to associate the presence of a chest tube with the disease, rather than identifying pneumothorax itself.[5] This is problematic since the medical value of such a model would be detecting patients suffering from pneumothorax before treatment begins, i.e., before the chest tube is inserted. Consequently, the presence of a chest tube acts as a confounder in pneumothorax detection.

Extensive research in explainable artificial intelligence (AI) underscores the prevalence of confounders in clinical image interpretation.[6–9] For example, metal tokens in radiographs, used to indicate body sides, can confound the presence of diseases like COVID-19.[8,9] Identifying and visualizing these confounders at a patient-specific level remains a significant challenge.

In recent years, diffusion-based models have gained significant attention in medical imaging due to their remarkable performance in tasks such as medical content generation,[10–12] image reconstruction,[13,14] image registration,[15] image segmentation,[16–18] and data synthesis with diseases such as tumors and lesions.[19,20] These models perturb original data toward a tractable posterior distribution, often Gaussian, and use neural networks to generate samples from this posterior.[21] However, compared to variational autoencoders and generative adversarial networks, diffusion models struggle to maintain semantically meaningful data representations in the latent space due to its high dimensionality and the nature of the forward diffusion process.[22,23]

Several works have explored the role of latent representations in image editing, demonstrating that images can be semantically edited based on natural language guidance.[24,25] For instance, Su et al.[26] acquire latent encodings from source images using the source diffusion model and denoising diffusion implicit model (DDIM) inversion. These encodings are subsequently decoded by the target diffusion model to create an image conditioned on the target text. Similarly, in SDEdit,[27] Meng et al. introduce intermediate noise to an image and then denoise it using a diffusion process conditioned on the intended edit. Nevertheless,

This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

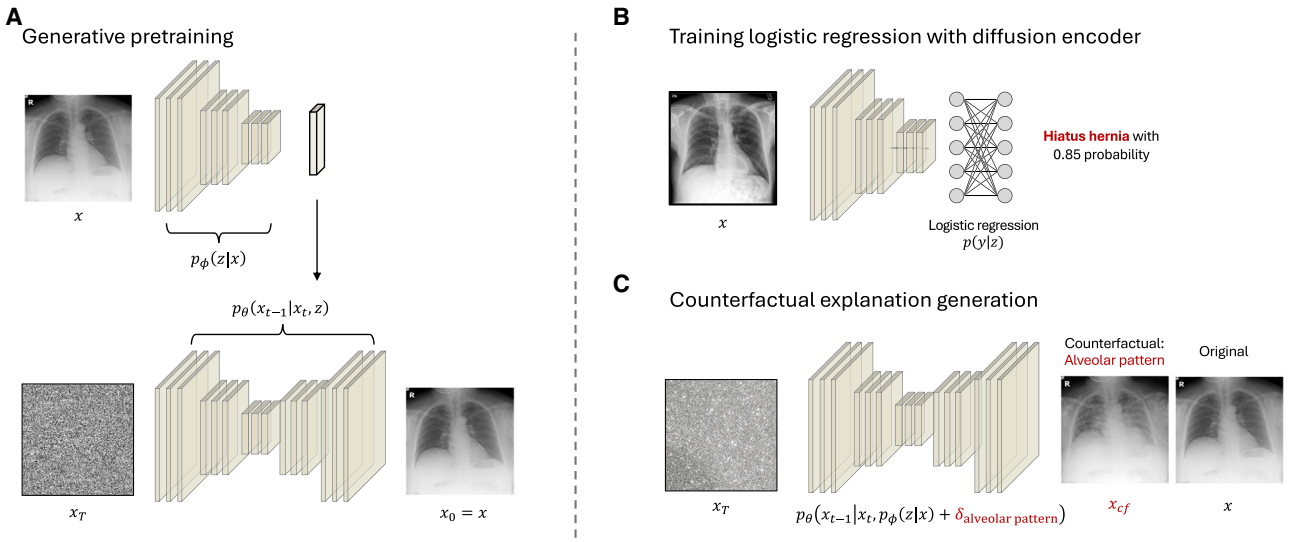


Figure 1. Schematic of DiffChest workflows

(A) The DiffChest model was pretrained on a collection of 497,215 chest radiographs (CXRs) from California, Massachusetts, and Spain. No labels were used to pretrain the model.

(B) The model's classification head, consisting of a single logistic regression layer, underwent fine-tuning using a selected subset of 18,489 radiographs from the PadChest dataset, which clinicians had manually annotated.

(C) Our method is both discriminative and generative. We use encoder-extracted features to train a classifier and employ the diffusion model to generate visual explanations for prediction classes. The latter is particularly crucial for clinicians to accept AI-assisted diagnosis, where interpretability is critical. We perturb the latent feature toward a target class and utilize the diffusion model to generate a visual explanation for the target class while preserving the original information (Equation 6).

these methods heavily depend on manipulating the original latent space of the diffusion model, which may be suboptimal due to its high dimensionality and noisy perturbations in the forward diffusion process.[24,28,29]

In this work, we introduce DiffChest, a self-supervised pretraining framework that utilizes diffusion models to generate patient-specific explanations for the classification of pulmonary pathologies. By pretraining on unlabeled 515,704 chest radiographs (Figure S1), DiffChest produces image representations rich in semantic information, enhancing disease classification capabilities. Our model combines generative and classification abilities, enabling the synthesis of high-quality radiographs and the identification of a wide range of confounders in training data, thereby evaluating model biases and failures. We demonstrate that optimizing the diffusion objective is equivalent to maximizing the mutual information between the input and output of the feature extractor. Our approach effectively leverages large volumes of unlabeled or noisy labeled data, advancing efficient and interpretable model-assisted diagnosis in clinical settings.

RESULTS

Comprehensive pulmonary disease detection

We investigate the effectiveness of the DiffChest model in diagnosing pulmonary diseases, comparing its performance to the state-of-the-art CheXzero model. We collected a testing cohort of 7,943 radiographs from 7,272 patients that expert clinicians labeled, see Figure S1. After generative pretraining (Figure 1A), DiffChest's features were linearly probed (Figure 1B) for diagnostic accuracy and benchmarked against CheXzero, which previously demonstrated high performance on the PadChest dataset.[30] CheXzero adopts contrastive learning, utilizing a dataset of more than 370,000 chest radiographs and radiology report pairs to learn a unified embedding space, enabling multi-label classification. DiffChest uses much less pretraining data (0.5 million images) than CheXzero does (400 million), demonstrating efficient data utilization. It is fine-tuned with 18,489 image-label pairs via supervised learning, while CheXzero uses 377,110 image-report pairs in a self-supervised framework. A detailed comparison between DiffChest and CheXzero can be found in Table S1.

To ensure a fair comparison, we trained a linear classifier on the latent space of CheXzero using the same labels as DiffChest. Our analysis, focusing on 73 imaging findings with over 30 entries in the PadChest testing cohort, showed that DiffChest performed on par with CheXzero in 46 findings; p values indicate no significance (p value > 0.05, Figure 2). In the remaining 27 findings, DiffChest outperformed CheXzero in 6, while CheXzero excelled in 21. Detailed results, including p values and area under the receiver operating characteristic curve values, are presented in Table S2. DiffChest performs worse than CheXzero on some findings such as pulmonary mass, consolidation, and lobar atelectasis. CheXzero leverages OpenAI's CLIP model, which was pretrained on 400 million image-text pairs and fine-tuned on 377,100 radiograph-report pairs.[30] In contrast, our DiffChest model was trained from scratch using less than 0.5 million unlabeled images and then fine-tuned on only 18,489 labeled images. In our fine-tuning dataset, the number of samples for pulmonary mass,

CellPress
OPEN ACCESS

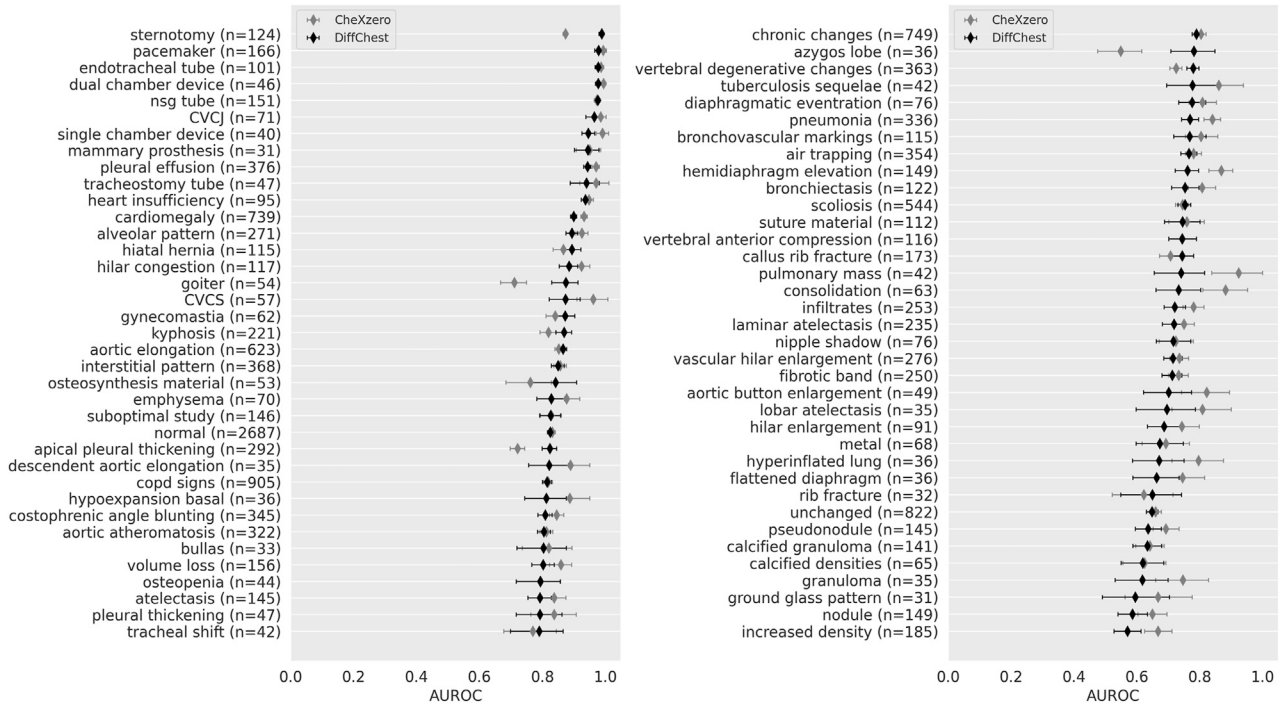


Figure 2. Detection of imaging findings by DiffChest and the CheXzero foundation model

The mean AUC and 95% confidence interval (CI) are shown for each imaging finding with more than 30 entries in the PadChest testing cohort. Findings are sorted by the mean AUC of DiffChest. The top 36 imaging findings are shown in left, and the rest imaging findings are shown in right. Among the total 73 evaluated imaging findings, DiffChest achieves a mean AUC of greater than 0.900 for 11 findings and greater than 0.700 for 59 findings. n in this plot refers to the number of positive examinations in the PadChest test set. CVCJ, central venous catheter via jugular vein; CVCS, central venous catheter via subclavian vein; nsg tube, nasogastric tube; copd, chronic obstructive pulmonary disease.

consolidation, and lobar atelectasis is limited to 98, 173, and 67, respectively. We believe that the performance of DiffChest can be further improved by incorporating more labeled data and pre-training the model on a larger dataset.

DiffChest is data efficient

Learning from insufficiently annotated datasets poses a significant challenge in medical AI, as the predictive performance of models generally improves with larger training datasets,[31] and high-quality medical data for training purposes are scarce. To address this issue, we investigated the potential of diffusion pre-training to mitigate performance losses when using only a small subset of radiographs from the CheXpert dataset for model tuning (the data flow is given in Figure S2). We employed CheXpert for our study due to the availability of high-quality testing labels and annotations from three board-certified radiologists.[32] Our findings demonstrate that DiffChest's performance remains stable and almost on par with radiologists when fine-tuning the model using only 10% of the total data (Figure S3). However, it is worth noting that all radiologists outperformed DiffChest on the classification of supporting devices. We attribute this to the presence of label noise within the CheXpert dataset, as DiffChest achieved an almost perfect area under the curve (AUC) when classifying device labels such as single-chamber device, dual-chamber device, and pacemaker in the PadChest dataset (Figure 2). Even when downsampled to only 3% of the total data, DiffChest maintains high AUC scores of at least 0.800 for no finding, edema, supporting devices, and pleural effusion.

Generating patient-specific explanations

The integration of AI in medicine necessitates explainable models to gain clinical acceptance.[33,34] While approaches like feature attribution,[35] gradient saliency,[9] class activation mapping (CAM),[36] and Grad-CAM[37] provide insights by highlighting regions or pixels in an image that influence the classification, they fall short of explaining the specific attributes within those regions that are relevant to the classification. They do not clarify whether the classification of medical images is influenced by factors such as the radiographic density of an organ or its shape. Here, we provide a visual explanation for DiffChest's classification decisions by starting from the representation of the original image in the semantic latent space $\mathbf{z}$ and then modifying this representation such that the image synthesized by the diffusion model has a higher probability of being classified as exhibiting the pathology (see Figure 1C).

Our method involves adding adversarial noise to the latent variable $\mathbf{z}$ in the reverse diffusion kernel $p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{z}^*)$:

$$\mathbf{z}^* = \mathbf{z} - \epsilon \cdot \nabla_{\mathbf{z}} \mathcal{L}\left(h(\mathbf{z}; \theta_{lr}), y^t\right), \qquad \text{(Equation 1)}$$

where $\mathcal{L}$ is the cross-entropy loss, $h(\mathbf{z}; \theta_{lr})$ is the logistic regression layer, and y^t is the target label. Unlike adversarial

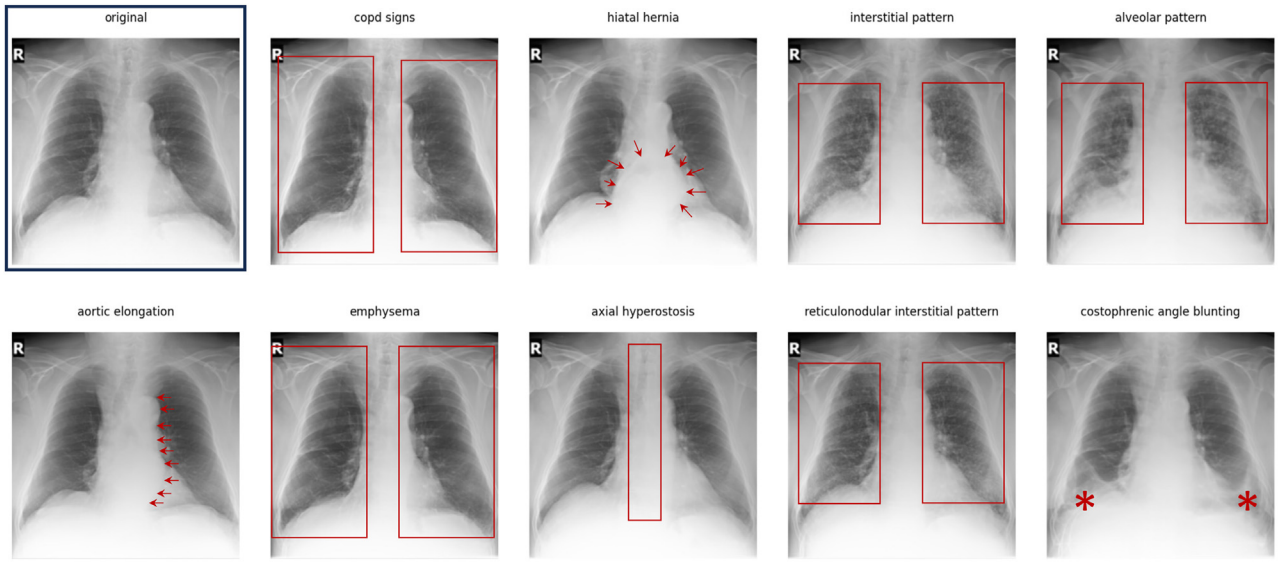


Figure 3. Visual explanations of patient CXRs

The original radiograph is highlighted by a black frame and was acquired from a male patient born in 1928. It was labeled as normal in the PadChest dataset. To demonstrate that DiffChest incorporates the image characteristics of a multitude of pathologies, we synthesize nine different pathologies, namely chronic obstructive pulmonary disease signs, hiatal hernia, interstitial pneumonia, alveolar pneumonia, aortic elongation, emphysema, axial hyperostosis, reticulonodular interstitial pneumonia, and costophrenic angle blunting for this patient. The generated explanations exhibit relevant pathological signatures, which are highlighted with arrows, stars, and frames by a board-certified radiologist with 12 years of experience.

examples,[38] which are generated by perturbing the input image, our approach generates counterfactual examples by perturbing the latent code **z**, which is then used to generate an image containing the desired clinical attributes.[39] Figures 3 and S4–S6 illustrate our generated explanations. Figure S7 further showcases the "progressive" and "regressive" visual explanations for a patient with a pathological condition, e.g., cardiomegaly, heart insufficiency, hilar congestion, and infiltrates. Moreover, our model is capable of distinguishing between chest posterior anterior (PA) and anterior posterior (AP) projections, as shown in Figure S8B. The heart size looks slightly larger in the generated AP projection compared to the generated PA projection due to projection differences. When training DiffChest, we only normalized the images and did not apply any data augmentation techniques such as flipping or rotation. After unsupervised pretraining, DiffChest achieved an AUC of 0.999 in distinguishing between AP and PA projections. We demonstrate the model's ability to generate AP/PA counterfactuals for the same patient in Figure S8A.

The quality of generated visual explanations is crucial for the reliability of the model's performance. We, therefore, conducted a quantitative assessment of the quality of the generated images using precision and recall metrics as described in the section quality assessment. As shown in Figure S9, our generated counterfactuals exhibit high precision values except for the class "endotracheal tube." These results indicate that the generated counterfactuals are realistic compared to the original samples. The recall values are lower compared to the precision values, indicating that the generated counterfactuals are less diverse than the original samples. This is expected as the visual explanations are (attribute) edited versions of the original radiographs.

This paradigm is important for the introduction of AI models in medicine as it allows users to understand the model's decision-making by getting direct visual feedback on which image characteristics lead to the model classifying a radiograph as exhibiting a certain pathology. Examples visualized in Figure 3 comprise overinflated lungs for patients suffering from chronic obstructive pulmonary disease, or characteristic changes of the lung parenchyma for patients suffering from interstitial pneumonic infiltration. These visualizations might also allow the clinical expert to make more nuanced diagnoses by emphasizing different pneumonic patterns (compare interstitial pattern, alveolar pattern, and reticulonodular interstitial pattern) that would not have been possible with traditional methods of AI explainability such as CAM.

Confounders in the training cohort can be identified

AI models can exploit subtle image characteristics for their diagnosis. While this may help achieve super-human performance, it comes with a problem: when the AI model is trained to differentiate between two groups (e.g., patients with pneumonia and without pneumonia), it may focus on differences in the group that are not directly related to the pathology but rather are present as spurious correlations. For instance, patients suffering from pneumonia might more often have a catheter inserted, which is not directly related to pneumonia but rather to the treatment of pneumonia. Therefore, an AI model that is trained to detect pneumonia may learn to detect the presence of a catheter, which is a confounder.[6–9] DiffChest's visual explanations enable the identification of such confounders by analyzing synthetic images for these spurious correlations.

CellPress
OPEN ACCESS

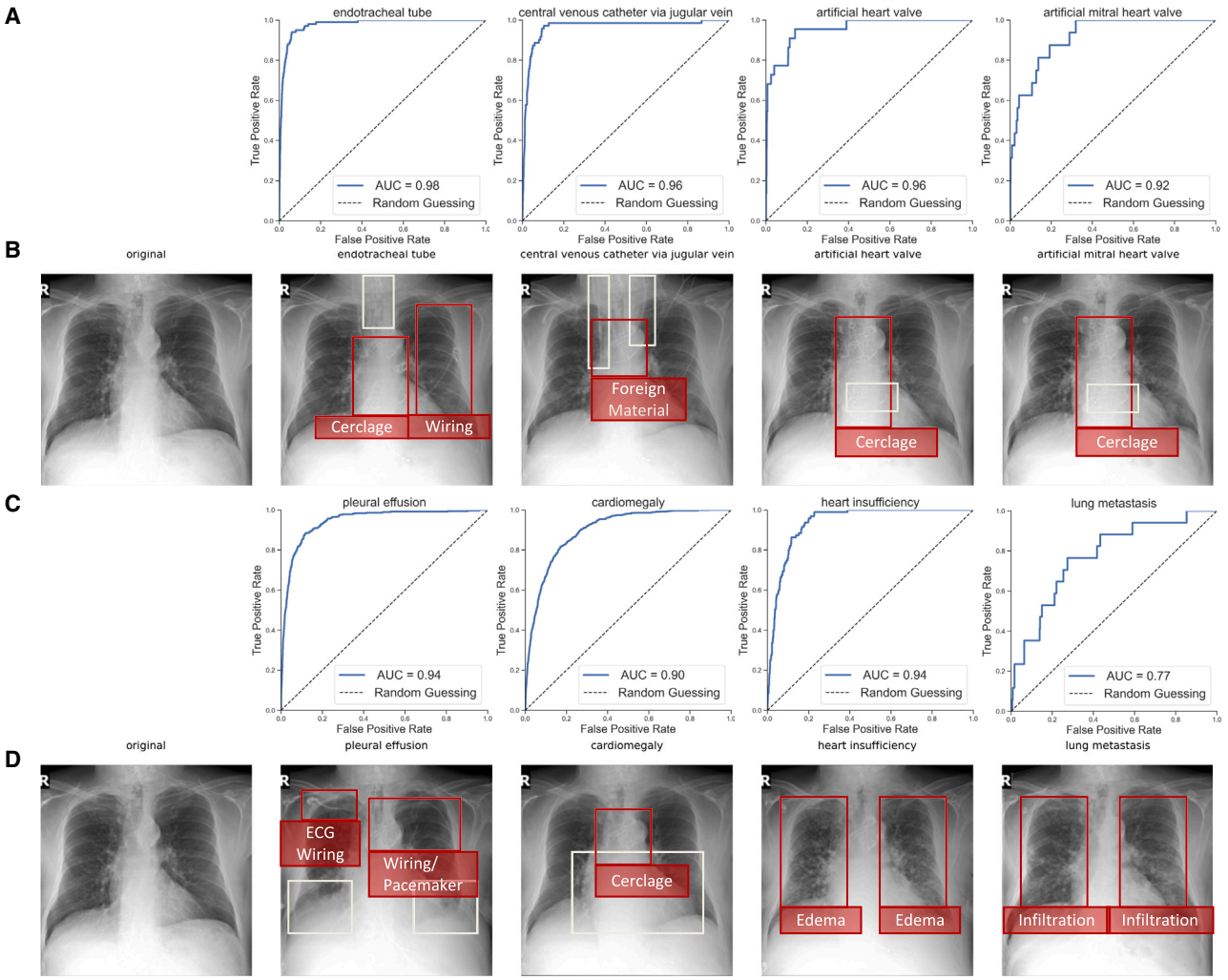


Figure 4. Confounders as synthesized by DiffChest

(A) Strong performance of DiffChest in detecting foreign materials.

(B) Visual explanations by DiffChest showcasing a specific pathology on a non-pathological radiograph. The target pathology is delineated with a white frame, while confounders such as sternal wire cerclages, drainage tubes, device wirings, and catheters are marked in red.

(C) High accuracy in pathology detection, evidenced by elevated AUC scores.

(D) Capability of DiffChest to generate visual explanations that illustrate co-morbidities.

To illustrate this, we conducted reader experiments where DiffChest generated synthetic radiographs exhibiting specific pathologies based on images that do not exhibit this pathology (Figure S10). A board-certified radiologist with 12 years of experience reviewed these images for confounders (Figure S10 step 1). Confounders were classified as either being treatment related (such as the presence of a catheter for patients suffering from pneumonia, Figure 4B) or as being related to the pathology (such as the presence of lung congestion for patients suffering from cardiomegaly, Figure 4D). Treatment-related and physiology-induced confounders are summarized in Tables S3 and S4.

To further validate the model's ability to detect confounding biases in the original data, we asked three radiologists to annotate the presence of confounders in real radiographs. The study design is detailed in Figure S10 in step 2 and step 3. Tables 1 and 2 show the number of real radiographs containing at least one treatment-related or physiology-related confounder. We illustrated the discordance between the radiologists when identifying treatment-related and physiology-related confounders in Figures S11 and S12. Our findings confirm that the confounders identified in synthetic images also appear in a significant portion of real radiographs, demonstrating DiffChest's capability to detect confounders in data.

Following the generation of counterfactuals for the presence of cardiomegaly (Figure S13A), the predicted probabilities of the reference patient images and the generated counterfactuals for the presence of confounding sternotomy were calculated (Figure S13B). Our DiffChest model was able to detect the confounding sternotomy in the generated counterfactuals.

Table 1. Presence of treatment-related confounders in radiographs

"Diagnosis" in radiograph	Radiographs (*n*)	Confounder present	Ratio (%)	Fleiss' kappa
Artificial heart valve	30	30	100.0	1.000
Central venous catheter via jugular vein	30	29	96.7	0.900
Endotracheal tube	30	28	93.3	0.867
Heart insufficiency	30	24	80.0	0.731
Artificial aortic heart valve	24	22	91.7	1.000
Pleural effusion	30	12	40.0	0.967
Lung metastasis	30	6	20.0	0.967
Pneumothorax	30	5	16.7	1.000
Cardiomegaly	30	4	13.3	0.932
Calcified densities	30	3	10.0	0.861
Catheter	9	1	11.1	1.000

In total, 303 radiographs were analyzed by three radiologists for the presence of a subset of confounders (Table S3). The subset of confounders that these radiographs were tested for had previously been identified with the help of DiffChest. Fleiss' kappa was computed to quantify inter-reader agreement.

Table 2. Presence of physiology-related confounders in radiographs

"Diagnosis" in radiograph	Radiographs (*n*)	Confounder present	Ratio (%)	Fleiss' kappa
Central venous catheter via jugular vein	30	20	66.7	0.551
Pleural effusion	30	20	66.7	0.660
Cardiomegaly	30	17	56.7	0.766
Calcified densities	30	14	46.7	0.466
Lung metastasis	30	11	36.7	0.358
Catheter	9	2	22.2	1.000

In total, 303 radiographs were analyzed by three radiologists for the presence of a subset of confounders (Table S4). The subset of confounders for which these radiographs were tested had previously been identified with the help of DiffChest. Fleiss' kappa was computed to quantify inter-reader agreement.

Additionally, the model can identify multiple confounders for a single diagnosis. For example, in cases of hilar congestion, the model generated hypothetical images featuring obesity, as verified by one radiologist (Figure S14).

Visual explanations improve congestion grading

To highlight the clinical relevance of our study, we investigated if the generated visual explanations could offer assistance in complex disease grading, specifically hilar congestion. Hilar congestion is a critical clinical finding that can signify underlying cardiovascular or pulmonary disorders, such as heart failure, pulmonary hypertension, or lymphadenopathy, and requires thorough investigation to diagnose and manage potential life-threatening conditions.

In our experiment, shown in Figure 5A, two board-certified radiologists were asked to grade the severity of hilar congestion in the reference patient image with and without the assistance of the counterfactuals (Figure 5B). Due to the lack of a gold standard for grading disease severity, personal perception and experience may influence the radiologists' grading (Figure 5C). When visual explanations were provided, the radiologists' grading was more consistent, as reflected by the increase in Cohen's kappa values from 0.43 to 0.47. This improvement was statistically significant ($p = 2.85 \times 10^{-38}$), indicating that the visual explanations enhanced the radiologists' agreement in identifying the findings. In Figure 5C, we show the distribution of the radiologists' grading scores with and without the visual explanations. Notably, the class indicating that the radiologist was not sure about the presence of congestion ("(+)") was not selected after the visual explanations were provided, indicating that the visual explanations helped the radiologists make more confident decisions. Both radiologists' labeling files are available online for further inspection.

DISCUSSION

In this work, we used a diffusion pretraining method for medical image interpretation that allows training classifiers without requiring large amounts of annotated data while also giving detailed insights into the model's decision-making process. The model development involved large-scale pretraining on 497,215 unlabeled chest radiographs followed by a subsequent fine-tuning step on a small subset of annotated data. Our method maintains high performance even when fine-tuning on only 3% of the total fine-tuning data, demonstrating its data efficiency.

We demonstrated that the diffusion part of the model can be used in conjunction with the classification part to synthesize high-quality radiographs. This capability can serve two purposes: first, it can explain the model's reasoning to its users, and second, it can be used to identify confounders in the training data. This is important since medical practitioners need to have confidence in the rationale behind an AI model's predictions before they will consider delegating aspects of their workload to the system. The ability to view a synthesized image that highlights the features upon which the model focuses can facilitate a quick assessment by physicians to determine whether the AI model's diagnosis aligns with their expert judgment. Moreover, the issue of confounding variables poses a significant challenge in the development of AI models for medical applications, leading to potential model biases.[40] The generation of radiographs by our model allows the direct identification of these confounders within the training dataset, thereby offering an avenue to enhance data quality and build more robust machine learning models. We substantiated this claim through a reader study, wherein we demonstrated that the confounders identified by DiffChest were consistently present in a large subset of real radiographic images. Even though we found that visual explanations help radiologists identify confounding factors, such generated visual explanations might sometimes be hard to understand, particularly with complex or subtle medical images.

CellPress
OPEN ACCESS

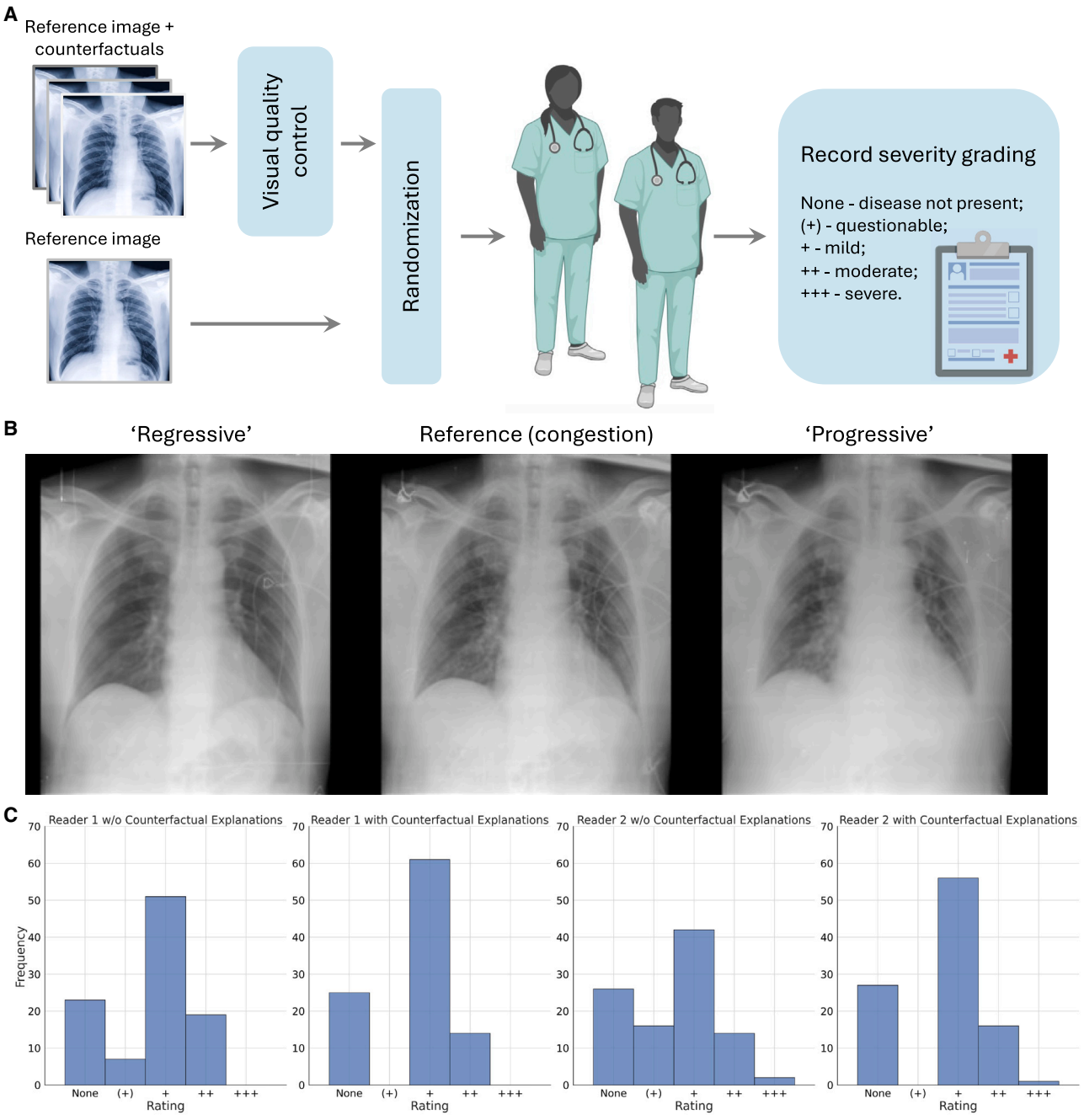


Figure 5. Visual explanations can assist radiologists in complex disease grading

(A) Overview of the joint expert and counterfactual disease grading procedure. Two radiologists were asked to grade the severity of hilar congestion in the reference patient image with and without the assistance of the counterfactuals.

(B) Example of a reference patient image with hilar congestion. Both "progressive" and "regressive" visual explanations were generated for the reference patient image.

(C) Distribution of the radiologists' grading scores with and without the (counterfactual) visual explanations.

We also investigated the clinical utility of visual explanations in the grading of hilar congestion, a critical clinical finding indicative of various cardiovascular or pulmonary disorders. Two board-certified radiologists graded the severity of hilar congestion in patient images with and without the assistance of counterfactual visual explanations. The inclusion of visual explanations led to increased consistency in grading, as evidenced by the rise in Cohen's kappa values from 0.43 to 0.47 ($p < 0.001$). Additionally, the visual explanations eliminated uncertainty in the radiologists' assessments, as indicated by the absence of the "not sure"

classification after visual explanations were provided. These findings highlight the practical benefits of visual explanations in aiding diagnostic confidence and consistency, particularly in complex cases where personal perception and experience significantly influence grading. To further validate the clinical utility of visual explanations, we recommend more extensive studies involving larger cohorts of clinicians and incorporating their feedback on the practical application of these tools in real-world settings.

Our method builds on the foundation of diffusion autoencoder (AE) and introduces several advancements that differentiate it from the previous approach. We provide rigorous theoretical proofs (section extended derivations) demonstrating that the latent encoding **z** contains rich semantic information. This theoretical underpinning is a contribution that substantiates the semantic richness of our latent representations. While diffusion AE was primarily aimed at image editing, our work focuses on the representation learning capability of diffusion-based generative pretraining. Through the manipulation of the semantic latent **z**, we detected and visually reconstructed various radiological confounders in medical image analysis. This capability enhances the interpretability and diagnostic utility of our method. Lastly, we have validated our model on extensive medical datasets with more than 500,000 medical images, whereas diffusion AE primarily relies on 70,000 images from Flickr-Faces-HQ and 30,000 images from CelebA-HQ.[28]

In contrast to currently prevailing contrastive learning approaches,[30,41] which rely on the information noise contrastive estimation loss to maximize the mutual information between the input and the context vector,[42–45] we introduced a desirable information bottleneck[46,47] within the denoising U-Net architecture, which helps to learn a more compact and semantically meaningful latent space. However, persistently maximizing mutual information between the input image and its encoded latent code may affect the model's downstream classification performance.[47] Future work can add a regularization term in the objective function to further boost the model classification performance.

Our model's ability to manipulate attributes of real clinical images raises ethical concerns, like DeepFake.[48] This potential may, in principle, be exploited for the generation of fake medical data, which may be leveraged for insurance fraud. This is a problem since the detectability of manipulated samples is challenging as compared to standard adversarial approaches. A potential solution to this problem is the addition of another neural network for the detection of fake images, as suggested by Preechakul et al.[28] Our work has limitations in that the resolution of the synthesized images is lower than what is normally used in clinical practice. This may be solved by integrating progressive growing techniques as demonstrated by Ho et al.[49] and Karras et al.[50] However, implementing these techniques may require substantial hardware resources, which may not always be readily available in clinical settings. Exploring resource-efficient alternatives to improve generative capabilities while maintaining practical feasibility will therefore be an important future research direction.

The identification of confounders is a core outcome of our research. The next steps that need to be undertaken by future research are the mitigation of these confounders through causality-inspired methods. For instance, Deng et al.[51] developed a confounder-aware representation learning system to improve immunotherapy response prediction, and Ouyang et al.[52] employed a causality-inspired data augmentation approach to enhance domain robustness in medical image segmentation. Future work could integrate the identification of confounders more directly into the model's learning process. Building upon a multi-label dataset, the model could be trained to automatically flag potential confounders in the training data, thereby simplifying the identification process. Additionally, future work can explore using the pretrained CLIP encoder as a starting point for training DiffChest, which may improve the model's performance and generalization capabilities.

In conclusion, this study demonstrates the potential of generative pretraining for developing interpretable models in medical image analysis. The model's ability to identify confounders in the training data and analyze model biases and failures is a significant step toward ensuring reliable and accountable AI applications in healthcare. Beyond chest radiography, our proof-of-concept study highlights the broader potential of self-supervised learning in conjunction with diffusion models, offering a pathway to informed predictions for medical personnel.

Limitations of the study

Despite the promising results, our study has some limitations in the context of diffusion-based pretraining for radiograph classification and model interpretation. Although our approach is data efficient, the absence of web-scale radiographic datasets containing billions of samples constrains the potential of our model. The full capacity of our approach may remain untapped until such large-scale datasets become available, suggesting that further enhancements could be realized with more extensive data. Additionally, while our method for image-based confounder identification successfully highlights potential biases in the training data, it does not inherently provide a means to correct these biases. Addressing such biases requires specialized statistical and causal modeling approaches, which were beyond the scope of this study. Lastly, the resolution of the synthesized images produced by our method is lower than that of real radiographs. This limitation may reduce the practical applicability of the generated images in clinical settings, where high-resolution imaging is often critical.

RESOURCE AVAILABILITY

Lead contact

Requests for further information on software and resources should be directed and will be fulfilled by the lead contact, Tianyu Han (than@ukaachen.de).

Materials availability

This study did not generate any unique reagents.

Data and code availability

All datasets used in this study are publicly available: The chest radiography datasets from MIMIC-CXR, CheXpert, and PadChest datasets can be requested from the following URLs: MIMIC-CXR: https://physionet.org/content/mimic-cxr/2.0.0/(requires credential access for potential users); CheXpert: https://stanfordmlgroup.github.io/competitions/chexpert/; PadChest: https://bimcv.cipf.es/bimcv-projects/padchest/. Testing labels and radiologists' annotations

of CheXpert can be downloaded from https://github.com/rajpurkarlab/cheXpert-test-set-labels. The code and pretrained models used in this study are made fully publicly available under https://github.com/peterhan91/diffchest. Additionally, our demo for generating visual explanations is accessible at https://colab.research.google.com/drive/1gHWCQxreE1Olo2uQiXfSFSVInmiX85Nn. Any additional information required to reanalyze the data reported in this work paper is available from the Lead Contact upon request.

ACKNOWLEDGMENTS

J.N.K. is supported by the German Cancer Aid (DECADE, 70115166), the German Federal Ministry of Education and Research (PEARL, 01KD2104C; CAMINO, 01EO2101; SWAG, 01KD2215A; TRANSFORM LIVER, 031L0312A; TANGERINE, 01KT2302 through ERA-NET Transcan; Come2Data, 16DKZ2044A; DEEP-HCC, 031L0315A), the German Academic Exchange Service (SECAI, 57616814), the German Federal Joint Committee (TransplantKI, 01VSF21048), the European Union's Horizon Europe and innovation programme (ODELIA, 101057091; GENIAL, 101096312), the European Research Council (ERC; NADIR, 101114631), the National Institutes of Health (EPICO, R01 CA263318), and the National Institute for Health and Care Research (NIHR, NIHR203331) Leeds Biomedical Research Centre. D.T. is funded by the German Federal Ministry of Education and Research (TRANSFORM LIVER, 031L0312A), the European Union's Horizon Europe and innovation programme (ODELIA, 101057091), and the German Federal Ministry of Health (SWAG, 01KD2215B).

AUTHOR CONTRIBUTIONS

T.H., J.N.K., S.N., and D.T. devised the concept of the study. D.T., L.H., M.S.H., and R.S. performed the reader tests. T.H. wrote the code and conducted the performance studies. T.H. and D.T. did the statistical analysis. T.H., J.N.K., and D.T. wrote the draft of the manuscript. All authors contributed to correcting the manuscript.

DECLARATION OF INTERESTS

J.N.K. declares consulting services for Owkin, France; DoMore Diagnostics, Norway; Panakeia, UK; and Scailyte, Basel, Switzerland. Furthermore, J.N.K. holds shares in Kather Consulting, Dresden, Germany and StratifAI GmbH, Dresden, Germany and has received honoraria for lectures and advisory board participation by AstraZeneca, Bayer, Eisai, MSD, BMS, Roche, Pfizer, and Fresenius. D.T. received honoraria for lectures by Bayer and holds shares in StratifAI GmbH, Germany.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - Ethics statement
- METHOD DETAILS
  - Dataset description
  - Image preprocessing
  - Model selection
  - Mutual information maximization for free
  - DiffChest workflow
  - Model architecture
  - Implementation of the DiffChest approach
  - Input encoding
  - Model interpretability and visualization
  - Classifier selection
  - Quality assessment
  - Design of reader studies
  - Extended derivations
  - Simplified diffusion objective
  - Mutual information maximized pretraining
  - Manipulating latent representations
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.xcrm.2024.101713.

Received: March 17, 2024
Revised: June 29, 2024
Accepted: August 13, 2024
Published: September 5, 2024

REFERENCES

1. Castro, D.C., Walker, I., and Glocker, B. (2020). Causality matters in medical imaging. Nat. Commun. *11*, 3673.
2. Glocker, B., Musolesi, M., Richens, J., and Uhler, C. (2021). Causality in digital medicine. Nat. Commun. *12*.
3. Zeng, J., Gensheimer, M.F., Rubin, D.L., Athey, S., and Shachter, R.D. (2022). Uncovering interpretable potential confounders in electronic medical records. Nat. Commun. *13*, 1014.
4. Mukherjee, P., Shen, T.C., Liu, J., Mathai, T., Shafaat, O., and Summers, R.M. (2022). Confounding factors need to be accounted for in assessing bias by machine learning algorithms. Nat. Med. *28*, 1159–1160.
5. Rueckel, J., Trappmann, L., Schachtner, B., Wesp, P., Hoppe, B.F., Fink, N., Ricke, J., Dinkel, J., Ingrisch, M., and Sabel, B.O. (2020). Impact of confounding thoracic tubes and pleural dehiscence extent on artificial intelligence pneumothorax detection in chest radiographs. Invest. Radiol. *55*, 792–798.
6. Zhao, Q., Adeli, E., and Pohl, K.M. (2020). Training confounderfree deep learning models for medical applications. Nat. Commun. *11*, 6010.
7. De Sousa Ribeiro, F., Xia, T., Monteiro, M., Pawlowski, N., and Glocker, B. (2023). High Fidelity Image Counterfactuals with Probabilistic Causal Models. Preprint at arXiv. https://doi.org/10.48550/arXiv.2306.15764.
8. Zech, J.R., Badgeley, M.A., Liu, M., Costa, A.B., Titano, J.J., and Oermann, E.K. (2018). Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a crosssectional study. PLoS Med. *15*, e1002683.
9. DeGrave, A.J., Janizek, J.D., and Lee, S.I. (2021). AI for radiographic COVID19 detection selects shortcuts over signal. Nat. Mach. Intell. *3*, 610–619.
10. Moghadam, P.A., Van Dalen, S., Martin, K.C., Lennerz, J., Yip, S., Farahani, H., and Bashashati, A. (2023). A morphology focused diffusion probabilistic model for synthesis of histopathology images. In Proceedings of the IEEE/CVF winter conference on applications of computer vision, pp. 2000–2009.
11. Kim, B., and Ye, J.C. (2022). Diffusion deformable model for 4D temporal medical image generation. In International Conference on Medical Image Computing and ComputerAssisted Intervention (Springer), pp. 539–548.
12. Dorjsembe, Z., Odonchimed, S., and Xiao, F. (2022). Threedimensional medical image synthesis with denoising diffusion probabilistic models. Medical Imaging with Deep Learning.
13. Jalal, A., Arvinte, M., Daras, G., Price, E., Dimakis, A.G., and Tamir, J. (2021). Robust compressed sensing mri with deep generative priors. Adv. Neural Inf. Process. Syst. *34*, 14938–14954.
14. Chung, H., and Ye, J.C. (2022). Scorebased diffusion models for accelerated MRI. Med. Image Anal. *80*, 102479.
15. Kim, B., Han, I., and Ye, J.C. (2022). Diffusemorph: Unsupervised deformable image registration using diffusion model. In European conference on computer vision (Springer), pp. 347–364.

16. Kim, B., Oh, Y., and Ye, J.C. (2022). Diffusion Adversarial Representation Learning for Selfsupervised Vessel Segmentation. In The Eleventh International Conference on Learning Representations.

17. Heidari, M., Kazerouni, A., Soltany, M., Azad, R., Aghdam, E.K., CohenAdad, J., and Merhof, D. (2023). Hiformer: Hierarchical multiscale representations using transformers for medical image segmentation. In Proceedings of the IEEE/CVF winter conference on applications of computer vision, pp. 6202–6212.

18. Azad, R., Heidari, M., Wu, Y., and Merhof, D. (2022). Contextual attention network: Transformer meets unet. In International Workshop on Machine Learning in Medical Imaging (Springer), pp. 377–386.

19. Chen, Q., Chen, X., Song, H., Xiong, Z., Yuille, A., Wei, C., and Zhou, Z. (2024). Towards generalizable tumor synthesis. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11147–11158.

20. Zhang, H., Yang, J., Wan, S., and Fua, P. (2024). Lefusion: Synthesizing myocardial pathology on cardiac mri via lesionfocus diffusion models. Preprint at arXiv. https://doi.org/10.48550/arXiv.2403.14066.

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

22. Kazerouni, A., Aghdam, E.K., Heidari, M., Azad, R., Fayyaz, M., Hacihaliloglu, I., and Merhof, D. (2023). Diffusion models in medical imaging: A comprehensive survey. Med. Image Anal. *88*, 102846.

23. Yang, L., Zhang, Z., Song, Y., Hong, S., Xu, R., Zhao, Y., Zhang, W., Cui, B., and Yang, M.H. (2023). Diffusion models: A comprehensive survey of methods and applications. ACM Comput. Surv. *56*, 1–39.

24. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. (2022). Highresolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 10684–10695.

25. Kawar, B., Zada, S., Lang, O., Tov, O., Chang, H., Dekel, T., Mosseri, I., and Irani, M. (2023). Imagic: Textbased real image editing with diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6007–6017.

26. Su, X., Song, J., Meng, C., and Ermon, S. (2022). Dual diffusion implicit bridges for imagetoimage translation. Preprint at arXiv. https://doi.org/10.48550/arXiv:2203.08382.

27. Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J.Y., and Ermon, S. (2021). Sdedit: Guided image synthesis and editing with stochastic differential equations. Preprint at arXiv. https://doi.org/10.48550/arXiv:2108.01073.

28. Preechakul, K., Chatthee, N., Wizadwongsa, S., and Suwajanakorn, S. (2022). Diffusion autoencoders: Toward a meaningful and decodable representation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10619–10629.

29. Vahdat, A., Kreis, K., and Kautz, J. (2021). Scorebased generative modeling in latent space. Adv. Neural Inf. Process. Syst. *34*, 11287–11302.

30. Tiu, E., Talius, E., Patel, P., Langlotz, C.P., Ng, A.Y., and Rajpurkar, P. (2022). Expertlevel detection of pathologies from unannotated chest Xray images via selfsupervised learning. Nat. Biomed. Eng. *6*, 1399–1406.

31. Lu, M.Y., Williamson, D.F.K., Chen, T.Y., Chen, R.J., Barbieri, M., and Mahmood, F. (2021). Dataefficient and weakly supervised computational pathology on wholeslide images. Nat. Biomed. Eng. *5*, 555–570.

32. Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciureallcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., et al. (2019). Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. Proc. AAAI Conf. Artif. Intell. *33*, 590–597.

33. Kundu, S. (2021). AI in medicine must be explainable. Nat. Med. *27*, 1328.

34. Singla, S., Pollack, B., Chen, J., and Batmanghelich, K. (2019). Explanation by progressive exaggeration. Preprint at arXiv. https://doi.org/10.48550/arXiv:1911.00483.

35. Sundararajan, M., Taly, A., and Yan, Q. (2017). Axiomatic attribution for deep networks. In International conference on machine learning (PMLR), pp. 3319–3328.

36. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., and Torralba, A. (2016). Learning deep features for discriminative localization. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2921–2929.

37. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). Gradcam: Visual explanations from deep networks via gradientbased localization. In Proceedings of the IEEE international conference on computer vision, pp. 618–626.

38. Han, T., Nebelung, S., Pedersoli, F., Zimmermann, M., SchulzeHagen, M., Ho, M., Haarburger, C., Kiessling, F., Kuhl, C., Schulz, V., and Truhn, D. (2021). Advancing diagnostic performance and clinical usability of neural networks via adversarial training and dual batch normalization. Nat. Commun. *12*, 4315.

39. Han, T., Kather, J.N., Pedersoli, F., Zimmermann, M., Keil, S., SchulzeHagen, M., Terwoelbeck, M., Isfort, P., Haarburger, C., Kiessling, F., et al. (2022). Image prediction of disease progression for osteoarthritis by stylebased manifold extrapolation. Nat. Mach. Intell. *4*, 1029–1039.

40. SeyyedKalantari, L., Zhang, H., McDermott, M.B.A., Chen, I.Y., and Ghassemi, M. (2021). Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in underserved patient populations. Nat. Med. *27*, 2176–2182.

41. Azizi, S., Culp, L., Freyberg, J., Mustafa, B., Baur, S., Kornblith, S., Chen, T., Tomasev, N., Mitrović, J., Strachan, P., et al. (2023). Robust and dataefficient generalization of selfsupervised machine learning for diagnostic imaging. Nat. Biomed. Eng. *7*, 756–779.

42. Oord, A.v. d., Li, Y., and Vinyals, O. (2018). Representation learning with contrastive predictive coding. Preprint at arXiv. https://doi.org/10.48550/arXiv:1807.03748.

43. He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. (2020). Momentum contrast for unsupervised visual representation learning. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 9729–9738.

44. Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. (2020). A simple framework for contrastive learning of visual representations. In International conference on machine learning (PMLR), pp. 1597–1607.

45. Bachman, P., Hjelm, R.D., and Buchwalter, W. (2019). Learning representations by maximizing mutual information across views. Adv. Neural Inf. Process. Syst. *32*.

46. Tishby, N., Pereira, F.C., and Bialek, W. (2000). The information bottleneck method. Preprint at arXiv. https://doi.org/10.48550/physics/0004057.

47. Tishby, N., and Zaslavsky, N. (2015). Deep learning and the information bottleneck principle. In 2015 ieee information theory workshop (itw) (IEEE), pp. 1–5.

48. Nguyen, T.T., Nguyen, Q.V.H., Nguyen, D.T., Nguyen, D.T., HuynhThe, T., Nahavandi, S., Nguyen, T.T., Pham, Q.V., and Nguyen, C.M. (2022). Deep learning for deepfakes creation and detection: A survey. Comput. Vis. Image Understand. *223*, 103525.

49. Ho, J., Saharia, C., Chan, W., Fleet, D.J., Norouzi, M., and Salimans, T. (2022). Cascaded diffusion models for high fidelity image generation. J. Mach. Learn. Res. *23*, 2249–2281.

50. Karras, T., Aila, T., Laine, S., and Lehtinen, J. (2017). Progressive growing of gans for improved quality, stability, and variation. Preprint at arXiv. https://doi.org/10.48550/arXiv:1710.10196.

51. Deng, J., Yang, J., Hou, L., Wu, J., He, Y., Zhao, M., Ni, B., Wei, D., Pfister, H., Zhou, C., et al. (2022). Genopathomic profiling identifies signatures for immunotherapy response of lung adenocarcinoma via confounderaware representation learning. iScience *25*, 105382.

52. Ouyang, C., Chen, C., Li, S., Li, Z., Qin, C., Bai, W., and Rueckert, D. (2023). Causalityinspired singlesource domain generalization for medical image segmentation. IEEE Trans. Med. Imag. *42*, 1095–1106.

53. Johnson, A.E.W., Pollard, T.J., Berkowitz, S.J., Greenbaum, N.R., Lungren, M.P., Deng, C.Y., Mark, R.G., and Horng, S. (2019). MIMICCXR, a de-identified publicly available database of chest radiographs with freetext reports. Sci. Data *6*, 317.

54. Peng, Y., Wang, X., Lu, L., Bagheri, M., Summers, R., and Lu, Z. (2018). NegBio: a highperformance tool for negation and uncertainty detection in radiology reports. AMIA Summits on Translational Science Proceedings *2018*, 188.

55. Bustos, A., Pertusa, A., Salinas, J.M., and de la Iglesia-Vayá, M. (2020). Padchest: A large chest xray image dataset with multilabel annotated reports. Med. Image Anal. *66*, 101797.

56. Higgins, I., Matthey, L., Pal, A., Burgess, C.P., Glorot, X., Botvinick, M.M., Mohamed, S., and Lerchner, A. (2017). betavae: Learning basic visual concepts with a constrained variational framework. ICLR (Poster) *3*.

57. Donahue, J., and Simonyan, K. (2019). Large scale adversarial representation learning. Adv. Neural Inf. Process. Syst. *32*.

58. Han, T., Nebelung, S., Haarburger, C., Horst, N., Reinartz, S., Merhof, D., Kiessling, F., Schulz, V., and Truhn, D. (2020). Breaking medical data sharing boundaries by using synthesized radiographs. Sci. Adv. *6*, eabb7973.

59. Kim, G., Kwon, T., and Ye, J.C. (2022). Diffusionclip: Textguided diffusion models for robust image manipulation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2426–2435.

60. Kwon, M., Jeong, J., and Uh, Y. (2022). Diffusion models already have a semantic latent space. Preprint at arXiv. https://doi.org/10.48550/arXiv:2210.10960.

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

62. Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I., and Abbeel, P. (2016). Infogan: Interpretable representation learning by information maximizing generative adversarial nets. Adv. Neural Inf. Process. Syst. *29*.

63. Dhariwal, P., and Nichol, A. (2021). Diffusion models beat gans on image synthesis. Adv. Neural Inf. Process. Syst. *34*, 8780–8794.

64. Ronneberger, O., Fischer, P., and Brox, T. (2015). Unet: Convolutional networks for biomedical image segmentation. Medical Image Computing and ComputerAssisted Intervention– MICCAI 2015: 18th International Conference, Munich, Germany, October 59, 2015, Proceedings, Part III 18 (Springer), pp. 234–241.

65. Brock, A., Donahue, J., and Simonyan, K. (2018). Large scale GAN training for high fidelity natural image synthesis. Preprint at arXiv. https://doi.org/10.48550/arXiv:1809.11096.

66. Wang, X., Girshick, R., Gupta, A., and He, K. (2018). Nonlocal neural networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 7794–7803.

67. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. (2019). Pytorch: An imperative style, highperformance deep learning library. Adv. Neural Inf. Process. Syst. *32*.

68. Song, J., Meng, C., and Ermon, S. (2020). Denoising Diffusion Implicit Models. In International Conference on Learning Representations.

69. Sajjadi, M.S., Bachem, O., Lucic, M., Bousquet, O., and Gelly, S. (2018). Assessing generative models via precision and recall. Adv. Neural Inf. Process. Syst. *31*.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
trained models	GitHub: https://github.com/peterhan91/diffchest	https://doi.org/10.5281/zenodo.13283726
reader study files	GDrive ID: 1NkDfBPHV71ZFAW2WrKk8eT9js3YDkDvz	this manuscript
Software and Algorithms		
Python (3.9.17)	https://www.python.org/	RRID:SCR_008394
NVIDIA CUDA (11.7)	https://developer.nvidia.com/cuda-downloads	N/A
NVIDIA cuDNN (8.5)	https://developer.nvidia.com/cudnn	N/A
PyTorch (2.0.1)	https://pytorch.org	RRID:SCR_018536
Pytorch Lightning (1.4.5)	https://lightning.ai/docs/pytorch/stable/	N/A
Other		
GPU RTX A6000	Nvidia Corp., Santa Clara, California.	N/A

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Ethics statement

This study was carried out in accordance with the Declaration of Helsinki and local institutional review board approval was obtained (EK028/19). This study does not include any experimental models (animals, human subjects, plants, microbe strains, cell lines, primary cell cultures).

METHOD DETAILS

Dataset description

Our method was trained on a diverse set of three publicly available datasets: the MIMIC-CXR dataset, the CheXpert dataset, and the PadChest dataset. The demographics of the combined dataset are shown in Table S5. The MIMIC-CXR dataset comprises an extensive collection of 377,110 frontal and lateral chest radiographs from 227,835 radiological studies.[53] To ensure the inclusion of patient information, we selected only the frontal radiographs, specifically the anteroposterior (AP) or posteroanterior (PA) views, for model pretraining, in cases where multiple radiographs were available for the same patient. Within the MIMIC-CXR dataset, each radiograph has been meticulously labeled using an NLP labeling tool[54] to indicate the presence of 14 different pathological conditions. These conditions include Atelectasis, Cardiomegaly, Consolidation, Edema, Enlarged Cardiomediastinum, Fracture, Lung Lesion, Lung Opacity, No Finding, Pleural Effusion, Pleural Other, Pneumonia, Pneumothorax, and Support Devices.

The CheXpert dataset, a large and publicly available dataset, contributed 224,316 chest radiographs obtained from 65,240 patients.[32] Like the MIMIC-CXR dataset, the radiographs in the official training split of CheXpert have also been annotated using an NLP labeling tool[32] to identify the presence of the same 14 pathological conditions as in the MIMIC-CXR dataset. In the official test split of CheXpert, radiological ground truth was manually annotated by five board-certified radiologists using a majority vote.[32] Manually annotated CheXpert radiographs were used to evaluate the data efficiency of our model. The study flow of CheXpert is shown in Figure S2.

Another dataset used in our study is PadChest, a publicly available dataset comprising 160,861 chest radiographs from 67,625 patients at Hospital San Juan, Spain [55]. The radiographs in PadChest were extensively annotated, covering 193 different findings, including 174 distinct radiographic findings and 19 differential diagnoses.[55] Notably, 27% of the images were manually annotated by trained physicians, while the remaining 73% were automatically labeled using a recurrent neural network.[55] We divided the PadChest dataset into two subsets based on patients: The subset, comprising 43,008 patients, contained only automatically labeled images, while the second subset, consisting of 24,239 patients, contained only physician-labeled images (see Figure S1). For evaluating the performance of our model, DiffChest, we further partitioned 30% of patients with manually labeled studies from the second subset to form a testing set, resulting in a total of 7,272 patients dedicated to model evaluation.

CellPress
OPEN ACCESS

Image preprocessing

Each of the radiographs used in this study was resized to 256 × 256 and zero-padded before training and testing. In DiffChest, each image was then normalized to the range of [-1, 1].

Model selection

The generation of diagnostic-level medical images is crucial for the clinical utility of the model. Prior models, such as beta-VAE, learn disentangled visual concepts with a constrained variational objective.[56] Despite its ability to learn interpretable latent representations, beta-VAE struggles to generate high-quality images due to the trade-off between reconstruction quality and latent space disentanglement. GAN-based models, such as BigBiGAN and StyleGAN, generate high-quality medical images but require adversarial training and additional approximations to recover the latent code from the image.[39,57,58] In this work, we selected the self-conditioned diffusion model for its ability to generate high-quality images while automatically learning a semantic latent space without the need for adversarial training or additional variational approximations.

Mutual information maximization for free

Diffusion models are a class of latent variable models expressed as $p_\theta(\mathbf{x}_0) = \int p_\theta(\mathbf{x}_{0:T})d\mathbf{x}_{1:T}$. In this formulation, $p_\theta(\mathbf{x}_i)$ refers to a neural network responsible for denoising a latent variable $\mathbf{x}_i$ at time t = i, as shown in Figure 1A. The sequence of latent variables $\mathbf{x}_1$, $\mathbf{x}_2$, …, $\mathbf{x}_T$ shares the same dimensionality as the observed data $\mathbf{x}_0$, sampled from the data distribution $p(\mathbf{x}_0)$[21]. However, such a high dimensional latent space poses challenges in finding meaningful directions, i.e., directions that are identified by human experts as relevant to the disease. This property is, however, essential for representation control.[59,60] In DiffChest, we explicitly introduce an additional 512-dimensional latent space $p_\phi(\mathbf{z}|\mathbf{x}_0)$, which conditions the reverse diffusion process $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{z})$, to obtain a latent space in which we can identify meaningful directions. We will refer to this latent space as semantic latent space.[28]

Following the approach described in [21] and,[28] DiffChest was trained using a simplified diffusion objective:

$$L_{\text{simple}} = \sum_{t=1}^{T} \mathbb{E}_{\mathbf{x}_0,\epsilon_t}\left[\|\epsilon_\theta(\mathbf{x}_t, t, \mathbf{z}) - \epsilon_t\|_2^2\right]. \quad \text{(Equation 2)}$$

In Section extended derivations, we prove that optimizing Equation 2 is equivalent to maximizing the likelihood of a joint distribution between $\mathbf{x}_0$,$\mathbf{z}$, i.e., $\log p_{\theta,\phi}(\mathbf{x}_0,\mathbf{z})$. This joint probability can be further decomposed into the product of the likelihood of the data $p_\theta(\mathbf{x}_0)$ and a second term $p_{\theta,\phi}(\mathbf{z}|\mathbf{x}_0)$, which is the posterior distribution of $\mathbf{z}$ given $\mathbf{x}_0$. Here, the second term $\log p_{\theta,\phi}(\mathbf{x}_0, \mathbf{z})$ is also known as a lower bound of mutual information in Variational Information Maximization (VIM)[61,62]:

$$I(\mathbf{z}; \mathbf{x}_0) \geq \mathbb{E}_{(\mathbf{z},\mathbf{x}_0)\sim p(\mathbf{z},\mathbf{x}_0)}\left[\log p_{\theta,\phi}(\mathbf{z}|\mathbf{x}_0)\right] + H(\mathbf{z}), \quad \text{(Equation 3)}$$

We denote $I(\mathbf{z}; \mathbf{x}_0)$ as the mutual information between $\mathbf{z}$ and $\mathbf{x}_0$, while $H(\mathbf{z})$ represents the entropy of $\mathbf{z}$. Therefore, we can interpret the training of DiffChest as maximizing both the data likelihood $\log p_\theta(\mathbf{x}_0)$ and the lower bound of the mutual information between the input image $\mathbf{x}_0$ and its encoded latent code $\mathbf{z}$. Once trained, DiffChest excels at generating high-quality samples while also preserving the semantic manipulation of the input data.

DiffChest workflow

The DiffChest model was generatively pretrained on a large dataset of 497,215 chest X-rays from the U.S. and Spain, without the use of labels (Figure 1A). This is followed by a fine-tuning stage, where the model's classifier, composed of a logistic regression layer, is optimized using 18,489 radiographs from the clinically annotated PadChest dataset (Figure 1B). The model's methodology is both discriminative, utilizing encoder-extracted features for classification (Figure 1B), and generative, employing a diffusion model to generate visual explanations for its predictions (Figure 1C). Details on visual explanation generation are explained in the section model interpretability and visualization.

Model architecture

Our model comprises a self-conditioned diffusion model, $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{z})$, used for both diffusion pretraining and image generation, and a feature extractor, $p_\phi(\mathbf{z}|\mathbf{x}_0)$. To facilitate image generation, we adopted an enhanced convolutional U-Net architecture.[63] This U-Net architecture can handle input resolutions of 256 × 256. We leveraged the efficacy of BigGAN residual blocks[64,65] and global attention layers,[66] which were integrated into the U-Net at multiple resolution stages. These additions enhance the model's ability to capture intricate details and global context, resulting in improved image generation. Adaptive group normalization layers were used in the U-Net to facilitate self-conditioning.[28,63] To achieve self-conditioning, we made changes to the U-Net's normalization layers. Instead of using traditional normalization techniques like batch or group normalization, we adopted adaptive group normalization layers.[28] The feature extractor $p_\phi(\mathbf{z}|\mathbf{x}_0)$ shares the same architecture as the encoder part of the previously mentioned denoising U-Net.

Implementation of the DiffChest approach

Our self-supervised model consists of a denoising U-Net and a feature encoder that we jointly train on 497,215 images. To prepare the data for pretraining, all extracted images are normalized and stored in a single LMDB file. For pretraining, we used an Adam optimizer with default $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1e{-}8$, and no weight decay to optimize the diffusion loss (Equation 2). The training progress was measured by the number of real images shown to DiffChest.[50] We trained our model with a fixed learning rate of $1e{-}4$ and a batch size of 12 until 200 million real radiographs were shown to the model. Our classification head consisted of a linear classifier, i.e., a logistic regression layer, which was trained on the latent $\mathbf{z}$ space in all experiments. During classifier fine-tuning, each latent vector was normalized using a sample mean and standard deviation of the entire fine-tune dataset. All computations were performed on a GPU cluster equipped with three Nvidia RTX A6000 48 GB GPUs (Nvidia, Santa Clara, Calif). When not otherwise specified, the code implementations were in-house developments based on Python 3.8 (https://www.python.org) and the software modules Numpy 1.24.3, Scipy 1.10.1, and Pytorch 2.0.[67]

Input encoding

The latent space of DiffChest consists of two parts: a 512-dimensional latent code $\mathbf{z}$ and a 256×256 noise map $\mathbf{x}_T$. To generate visual explanations, we first need to encode the input image $\mathbf{x}_0$ into both latent spaces. Embedding $\mathbf{x}_0$ into $\mathbf{z}$ space is accomplished using our trained feature extractor $p_\phi(\mathbf{z}|\mathbf{x}_0)$. Similarly, we encode $\mathbf{x}_0$ into $\mathbf{x}_T$ space by running the generative process in our diffusion model backward. This process is deterministic as we utilized a DDIM sampler for sample generation:

$$\mathbf{x}_t = \sqrt{\alpha_t}\mu_{\theta,\phi}(\mathbf{x_{t-1}}, t-1, \mathbf{z}) + \sqrt{1-\alpha_t}\epsilon_{\theta,\phi}(\mathbf{x}_{t-1}, t-1, \mathbf{z}). \quad \text{(Equation 4)}$$

Here, the mean $\mu_{\theta,\phi}(\mathbf{x_{t-1}}, t, \mathbf{z})$ is defined in Equation 11. In our experiments, we set the number of encoding steps to 250. After encoding, we can use both the encoded $\mathbf{z}$ and $\mathbf{x}_T$ (as shown in Figure S15) to reconstruct the input image. As depicted in Figure S15, sharp and high-fidelity reconstructions can be achieved by using our DiffChest with just 100 sampling steps.

By design, the latent code $\mathbf{z}$ captures most high-level semantics, whereas the DDIM latent $\mathbf{x}_t$ is mainly responsible for the control of low-level stochastic variations. As shown in Figure S16, we begin by computing the semantic latent $\mathbf{z} = p_\phi(\mathbf{z}|\mathbf{x}_0)$ from an input image $\mathbf{x}_0$. Instead of extracting the stochastic subcode $\mathbf{x}_T$ directly from the input, we opt to sample it multiple times from a normal distribution. We then decode these samples to produce several outputs, each shown in Figure S16B to f. While high-dimensional latent codes do not inherently lack interpretability, the interpretability of the latent codes generated by DDIM inversion is limited by the non-smooth transitions observed in the latent space.[28] In contrast, DiffChest generates smooth and interpretable counterfactuals by leveraging the diffusion model's ability to generate realistic images.

Model interpretability and visualization

Visual explanations are obtained by extrapolating the latent code from $\mathbf{z}$ to $\mathbf{z}^*$ linearly, moving along the targeted adversarial direction, as represented by Equation 1. The closed-form solution for Equation 1 is given by:

$$\mathbf{z}^* = \mathbf{z} + \epsilon \cdot \left(\mathbf{y}^t - \sigma(\mathbf{w}^\top \mathbf{z} + \mathbf{b})\right)\mathbf{w}^\top, \quad \text{(Equation 5)}$$

where $\mathbf{w}$ and $\mathbf{b}$ are the weight and bias of the logistic regression layer, respectively, and σ is the sigmoid function. Derivations of Equation 5 are provided in the supplemental material. The term $\mathbf{y}^t\mathbf{w}^\top$ increases the probability of the target class $\mathbf{y}^t$, while the term $-\sigma(\mathbf{w}^\top\mathbf{z} + \mathbf{b})\mathbf{w}^\top$ decreases the probability of the original class. To preserve the original input information, we update only the latent code $\mathbf{z}$ toward the target class, while keeping the original image attributes unchanged, i.e.,

$$\mathbf{z}^* = \mathbf{z} + \epsilon \cdot \mathbf{y}^t\mathbf{w}^\top. \quad \text{(Equation 6)}$$

In our experiments, we set the factor ϵ to 0.3. The manipulated latent code $\mathbf{z}^*$ was then used to condition the reverse diffusion process $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{z}^*)$, allowing us to generate an image $\mathbf{x}^t$. For this purpose, our conditional diffusion model takes inputs $(\mathbf{z}^*, \mathbf{x}_T)$, where $\mathbf{x}_T$ is an encoded noisy image used to initialize the diffusion process (see section input encoding). To expedite sample generation, we adopt a non-Markovian DDIM sampler with 200 sampling steps proposed by Song et al..[68]

Classifier selection

Our main objective is to demonstrate that our self-supervised pretraining technique effectively learns a linearly separable space for different lung conditions. We selected logistic regression, as it is a linear classification acting on this space and as it is simple and interpretable. Importantly, our approach's ability to generate high-quality counterfactual examples through simple linear operations (Equation 6) highlights the robustness and interpretability of our model.

Quality assessment

To measure whether the generated images converged to realistic radiographs, we measured the precision and recall of generated visual explanations.[69] Precision and recall values plotted in Figure S9 characterize the consistency between the target and the model distribution. Precision denotes how realistic the generated samples are, and recall compares the coverage between the generative model and the original training manifold. Both values were computed based on the distribution-wise precision and recall curve (PRC) proposed in [69]. Unlike the operating points selected from standard PRCs, the x and y axes in Figure S9 are F_8 and $F_{1/8}$ defined as:

$$F_\beta = (1+\beta^2)p \cdot \frac{r}{\beta^2 p + r},$$

where p and r are precision and recall values in PRC. The hyperparameter β was chosen to be 8 and 1/8 as suggested by.[69]

In detail, we first generated 100 visual explanation samples for each class in our reader study (Table 1) except for the class "Catheter" due to the lack of samples (real samples with the class "Catheter" are fewer than 100 in our test set). In total, 1,000 visual explanation samples were generated. We then randomly collected 100 real radiographs from the same class and calculated the precision and recall values based on the generated and real samples. Finally, we visualized the precision and recall values for each class in our reader study in Figure S9.

Design of reader studies

Radiological confounders, as reconstructed by DiffChest, expose potential biases that are specific to the data-collecting institutions. To validate the accuracy and reliability of the identified confounding factors, we enlisted the expertise of four radiologists. We proceeded as follows: First, we provided the radiologists with a list of potential confounders that were pre-selected based on their clinical relevance by a radiologist with 12 years of experience. Then we asked the radiologists to annotate the signals solely using data generated by DiffChest (step 1 in Figure S10). Then we chose a test set of real radiographs (step 2 in Figure S10) with which we verified these confounding elements in the radiologists' reader study (step 3 in Figure S10).

Extended derivations

Mutual information

Mutual information is a general way to measure dependency between two random variables. In information theory, the mutual information can be defined as the difference of two entropies:

$$I(\mathbf{z};\mathbf{x}) = H(\mathbf{z}) - H(\mathbf{z}|\mathbf{x}) = H(\mathbf{x}) - H(\mathbf{x}|\mathbf{z}). \quad \text{(Equation 7)}$$

$I(\mathbf{z};\mathbf{x})$ is the reduction of uncertainty in $\mathbf{x}$ when $\mathbf{z}$ is observed. In self-supervised learning, we hope to find a compressed and informative representation of the input data, i.e., maximizing the mutual information between the input and the latent representation. However, it is hard to compute the mutual information directly due to the intractable posterior:

$$I(\mathbf{z};\mathbf{x}) = H(\mathbf{z}) - \mathbb{E}_{(\mathbf{z},\mathbf{x})\sim p(\mathbf{z},\mathbf{x})}[-\log p(\mathbf{z}|\mathbf{x})]. \quad \text{(Equation 8)}$$

In VIM,[62] we can estimate the posterior by introducing a variational distribution $p_\phi(\mathbf{z}|\mathbf{x})$:

$$\begin{aligned} I(\mathbf{z};\mathbf{x}) &= H(\mathbf{z}) - \mathbb{E}_{(\mathbf{z},\mathbf{x})\sim p(\mathbf{z},\mathbf{x})}[-\log p(\mathbf{z}|\mathbf{x})] \\ &= H(\mathbf{z}) + \mathbb{E}_{(\mathbf{z},\mathbf{x})\sim p(\mathbf{z},\mathbf{x})}\left[\log p(\mathbf{z}|\mathbf{x}) - \log p_\phi(\mathbf{z}|\mathbf{x}) + \log p_\phi(\mathbf{z}|\mathbf{x})\right] \\ &= H(\mathbf{z}) + \mathbb{E}_{(\mathbf{z},\mathbf{x})\sim p(\mathbf{z},\mathbf{x})}\left[D_{\mathrm{KL}}\left(p(\mathbf{z}|\mathbf{x}) \,\|\, p_\theta(\mathbf{z}|\mathbf{x})\right) + \log p_\phi(\mathbf{z}|\mathbf{x})\right] \\ &\geq H(\mathbf{z}) + \mathbb{E}_{(\mathbf{z},\mathbf{x})\sim p(\mathbf{z},\mathbf{x})}\left[\log p_\phi(\mathbf{z}|\mathbf{x})\right]. \end{aligned} \quad \text{(Equation 9)}$$

The entropy $H(\mathbf{z})$ is typically treated as a constant, and the lower bound of mutual information is given by $\mathbb{E}_{(\mathbf{z},\mathbf{x})\sim p(\mathbf{z},\mathbf{x})}[\log p_\phi(\mathbf{z}|\mathbf{x})]$ (Equation 3).

Simplified diffusion objective

Next, we will derive the simplified diffusion objective from the variational lower bound (VLB) on negative log joint likelihood $-\log p_{\theta,\phi}(\mathbf{x}_0,\mathbf{z})$ in DiffChest. Given the forward diffusion process as $q(\mathbf{x}_{1:T}|\mathbf{x}_0)$ and our designed backward process as $p_{\theta,\phi}(\mathbf{x}_{1:T}|\mathbf{x}_0,\mathbf{z})$, we can derive the VLB as:

$$
\begin{aligned}
-\log p_{\theta,\phi}(\mathbf{x}_0,\mathbf{z}) &\leq -\log p_{\theta,\phi}(\mathbf{x}_0,\mathbf{z}) + D_{\mathrm{KL}}\left(q(\mathbf{x}_{1:T}|\mathbf{x}_0)\,\|\,p_{\theta,\phi}(\mathbf{x}_{1:T}|\mathbf{x}_0,\mathbf{z})\right)\\
&= -\log p_{\theta,\phi}(\mathbf{x}_0,\mathbf{z}) + \mathbb{E}_{\mathbf{x}_{1:T}\sim q(\mathbf{x}_{1:T}|\mathbf{x}_0)}\left[\log\frac{q(\mathbf{x}_{1:T}|\mathbf{x}_0)}{p_{\theta,\phi}(\mathbf{x}_{1:T}|\mathbf{x}_0,\mathbf{z})}\right]\\
&= \mathbb{E}_{(\mathbf{x}_{0:T},\mathbf{z})\sim q(\mathbf{x}_{0:T},\mathbf{z})}\left[\log\frac{q(\mathbf{x}_{1:T}|\mathbf{x}_0)}{p_{\theta,\phi}(\mathbf{x}_{0:T},\mathbf{z})}\right]\\
&= \mathbb{E}_q\left[\log\frac{\prod_{t=1}^{T} q(\mathbf{x}_t|\mathbf{x}_{t-1})}{p(\mathbf{x}_T,\mathbf{z})\prod_{t=1}^{T} p_{\theta,\phi}(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{z})}\right]\\
&= \mathbb{E}_q\left[-\log p(\mathbf{x}_T,\mathbf{z}) + \sum_{t=1}^{T}\log\frac{q(\mathbf{x}_t|\mathbf{x}_{t-1})}{p_{\theta,\phi}(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{z})}\right] \qquad \text{(Equation 10)}\\
&= \mathbb{E}_q\left[-\log p(\mathbf{x}_T,\mathbf{z}) + \sum_{t=2}^{T}\log\frac{q(\mathbf{x}_t|\mathbf{x}_{t-1})}{p_{\theta,\phi}(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{z})} + \log\frac{q(\mathbf{x}_1|\mathbf{x}_0)}{p_{\theta,\phi}(\mathbf{x}_0|\mathbf{x}_1,\mathbf{z})}\right]\\
&= \mathbb{E}_q\left[-\log p(\mathbf{x}_T,\mathbf{z}) + \sum_{t=2}^{T}\log\left(\frac{q(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{x}_0)}{p_{\theta,\phi}(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{z})}\cdot\frac{q(\mathbf{x}_t|\mathbf{x}_0)}{q(\mathbf{x}_{t-1}|\mathbf{x}_0)}\right) + \log\frac{q(\mathbf{x}_1|\mathbf{x}_0)}{p_{\theta,\phi}(\mathbf{x}_0|\mathbf{x}_1,\mathbf{z})}\right]\\
&= \mathbb{E}_q\left[\underbrace{\log\frac{q(\mathbf{x}_T|\mathbf{x}_0)}{p(\mathbf{x}_T,\mathbf{z})}}_{L_T} + \underbrace{\sum_{t=2}^{T}\log\frac{q(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{x}_0)}{p_{\theta,\phi}(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{z})}}_{L_t} - \underbrace{\log p_{\theta,\phi}(\mathbf{x}_0|\mathbf{x}_1,\mathbf{z})}_{L_0}\right].
\end{aligned}
$$

The VLB is now factorized into three terms, namely L_0, L_t, and L_T. We may ignore L_T since it has no learnable parameters and $\mathbf{x}_T$ is by definition a Gaussian noise. Following,[28] we model L_0 using a deterministic decoder derived from $\mathcal{N}(\mathbf{x}_0;\mu_{\theta,\phi}(\mathbf{x}_1,1,\mathbf{z}),\mathbf{0})$. In DDIM's formulation, the mean of the reverse kernel $\mu_{\theta,\phi}(\mathbf{x}_t,t,\mathbf{z})$ is parameterized as

$$
\mu_{\theta,\phi}(\mathbf{x}_t,t,\mathbf{z}) = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \sqrt{1-\alpha_t}\,\epsilon_{\theta,\phi}(\mathbf{x}_t,t,\mathbf{z})\right), \qquad \text{(Equation 11)}
$$

where α_t is the diffusion parameter used in [68].

Equation 2 is derived from L_t. Ho et al. proved the forward kernel $q(x_{t-1}|x_t,x_0)$ follows Gaussian distributions.[21] The backward kernel $p_{\theta,\phi}(x_{t-1}|x_t,z)$ is also Gaussian since the step size is small. L_t is $D_{\mathrm{KL}}(q(x_{t-1}|x_t,x_0)\|p_{\theta,\phi}(x_{t-1}|x_t,z))$, since both distributions are Gaussian, we can derive the closed-form solution for L_t:

$$
\begin{aligned}
L_t &= \mathbb{E}_{\mathbf{x}_0,\epsilon}\left[\frac{1}{2\,\|\,\Sigma_{\theta,\phi}(\mathbf{x}_t,t,\mathbf{z})\,\|_2^2}\,\|\,\mu_t(\mathbf{x}_t,\mathbf{x}_0) - \mu_{\theta,\phi}(\mathbf{x}_t,t,\mathbf{z})\|^2\right]\\
&= \mathbb{E}_{\mathbf{x}_0,\epsilon}\left[\frac{(1-\alpha_t)^2}{2\alpha_t(1-\bar{\alpha}_t)\,\|\,\Sigma_{\theta,\phi}\,\|_2^2}\,\|\,\epsilon_t - \epsilon_{\theta,\phi}(\mathbf{x}_t,t,\mathbf{z})\|^2\right],
\end{aligned}
\qquad \text{(Equation 12)}
$$

where $\Sigma_{\theta,\phi}(\mathbf{x}_t,t,\mathbf{z})$ is the variance and ϵ_t is the Gaussian noise added at step t. We can get the simplified diffusion objective (Equation 2) by neglecting the weighting term in Equation 12.

Mutual information maximized pretraining

We have proven in the previous section that optimizing our diffusion pretraining loss, i.e., Equation 2, is equivalent to maximizing the VLB of the likelihood $\log p_{\theta,\phi}(\mathbf{x}_0,\mathbf{z}) = \log p_\theta(\mathbf{x}_0) + \log p_{\theta,\phi}(\mathbf{z}|\mathbf{x}_0)$. The first term, $\log p_\theta(\mathbf{x}_0)$, represents the log likelihood of the input data, while the second term, $\log p_{\theta,\phi}(\mathbf{z}|\mathbf{x}_0)$, corresponds to the VLB of the mutual information between the input and the latent representation (Equation 9). Therefore, optimizing our diffusion pretraining loss is equivalent to maximizing both the data likelihood and the mutual information between the input $\mathbf{x}$ and its latent representation $\mathbf{z}$.

Manipulating latent representations

In Equation 1, we showed, in general, visual explanations can be obtained by moving the encoded latent ($\mathbf{z}$) of a patient radiograph along the target adversarial direction $\nabla_{\mathbf{z}}\mathcal{L}(h(\mathbf{z};\theta),\mathbf{y})$. Here, $\mathcal{L}$ is the cross-entropy loss, $h(\mathbf{z};\theta)$ can be an arbitrary classifier with parameters θ, and $\mathbf{y}$ is the target label. Given the logistic regression model $\widehat{\boldsymbol{y}} = \sigma(\mathbf{w}^{\top}\mathbf{z}+b)$ we can derive its gradient concerning input as

$$\frac{\partial L}{\partial \mathbf{z}} = \frac{\partial L}{\partial \mathbf{k}}\cdot\frac{\partial \mathbf{k}}{\partial \mathbf{z}}. \quad \text{(Equation 13)}$$

Given $L(\mathbf{y},\widehat{\boldsymbol{y}}) = -\mathbf{y}\log(\widehat{\boldsymbol{y}}) - (1-\mathbf{y})\log(1-\widehat{\boldsymbol{y}})$, $\mathbf{k} = \mathbf{w}^{\top}\mathbf{z}+b$, and $\sigma(\mathbf{k}) = 1/(1+\exp(-\mathbf{k}))$, we can derive

$$\begin{aligned}\frac{\partial L}{\partial \mathbf{k}} &= \frac{\partial L}{\partial \widehat{\mathbf{y}}}\cdot\frac{\partial \widehat{\mathbf{y}}}{\partial \mathbf{k}}\\ &= \left(-\frac{\mathbf{y}}{\widehat{\mathbf{y}}}+\frac{1-\mathbf{y}}{1-\widehat{\mathbf{y}}}\right)\cdot\widehat{\mathbf{y}}(1-\widehat{\mathbf{y}})\\ &= -(\mathbf{y}-\widehat{\mathbf{y}}).\end{aligned} \quad \text{(Equation 14)}$$

In summary, the gradient of the loss for the input $\mathbf{z}$ is $-\partial L/\partial \mathbf{z} = (\mathbf{y}-\widehat{\boldsymbol{y}})\cdot\mathbf{w}$ (Equation 5). Our manipulation is done by moving $\mathbf{z}$ linearly along the target direction $\mathbf{w}$, founded by training a logistic regression classifier.

QUANTIFICATION AND STATISTICAL ANALYSIS

For each of the experiments, we calculated the ROC-AUC on the test set. If not otherwise stated, we extracted 95% CI using bootstrapping with 1,000 redraws. The difference in ROC-AUC was defined as the Δmetric. To perform bootstrapping, we built models for the total number of $N = 10{,}000$ bootstrapping by randomly permuting predictions of two classifiers, and then computed metric differences Δmetric_i from their respective scores. We obtained the two-tailed p-value of individual metrics by counting all Δmetric_i above the threshold Δmetric. Statistical significance was defined as $p < 0.001$. For the clinical reader experiments, we used Fleiss' kappa to calculate inter-reader agreement between the three radiologists.