

Image-Conditioned Diffusion for Privacy-Preserving Synthetic Medical Images

Doron Yaya-Stupp^{1,2}, Guy Lutsker¹, Or Spiegel-Yerushalmi¹, Eran Segal^{1,3}

1. Department of Computer Science and Applied Mathematics, Weizmann Institute of Science, Rehovot, Israel
2. The Gertner Institute for Epidemiology and Health Policy Research, Sheba Medical Center, Ramat Gan, Israel
3. Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE

Abstract

Medical imaging models depend on large, shareable datasets, yet privacy constraints limit data dissemination. Current text-conditioned diffusion models fail to preserve subtle, distributed clinical signals, such as continuous physiological biomarkers, rendering synthetic data insufficient for robust downstream physiological modeling. Here, we evaluate image-to-image (I2I) diffusion as a tunable, privacy-preserving transformation that produces a synthetic counterpart of real images while preserving downstream-relevant information. We fine-tune Stable Diffusion with low-rank adapters on retinal fundus photographs and chest radiographs, assessing fidelity, clinical signal preservation, cross-site transfer, and empirical re-identification risk. I2I consistently outperforms text-to-image generation in image fidelity and in preserving biomarker information. In cross-cohort transfer to an external retinal dataset from the UK Biobank, pretraining on I2I synthetic data performs comparably to real-image pretraining and surpasses it in the smallest fine-tuning sets. Varying I2I strength reveals that the privacy-utility tradeoff is highly modality-dependent: while retinal images achieve practical de-identification, chest X-rays exhibit structural combinatorics that leave them substantially re-identifiable even at high noise strengths, exposing critical boundaries for diffusion-based anonymization. These results position image-conditioned diffusion as a practical approach for generating shareable medical images with tunable de-identification.

Introduction

Medical image models require the curation and annotation of large image datasets. Recent image foundation models for medicine promise to reduce labeled-data requirements by enabling efficient fine-tuning on smaller task-specific datasets. However, privacy constraints and concerns raised about

the potential re-identification of patients from images used for training may hamper their development and dissemination¹. Synthetic images provide a potentially privacy-preserving alternative to training strong foundation models with a low risk of patient re-identification. These generative models and the predictive models trained using generated synthetic data have shown promise in recent years for chest X-ray (CXR)^{2,3}, retina fundus images⁴, and other imaging modalities⁵⁻⁸. These developments motivate approaches that simultaneously quantify (i) downstream clinical utility—such as enabling cross-institutional transfer learning without sharing restricted source images—and (ii) re-identification risk.

However, current approaches, mostly focused on text- or label-conditioned generation, can only generate images contained in the given prompt. In medical imaging, clinically relevant signals are often subtle, distributed, and correlated across phenotypes, and may not be captured by a small set of explicit conditioning variables. Subtle physiological signals and correlated variation can thus be lost unless explicitly mentioned, especially when seldom known in advance. Hence, previous works have typically focused on binary disease labels (e.g., pneumonia², diabetic retinopathy grade⁴) and baseline characteristics (age, sex). This limitation is particularly important for continuous biomarkers (e.g., hemoglobin level, blood pressure, etc.) whose imaging correlates may be weak and not directly interpretable, yet still measurable by learned models^{9,10}.

Text and image-based conditioning of diffusion models is an established approach in computer vision, where a noised version of an existing image is used as the starting point for the diffusion generative process¹¹. This approach, henceforth referred to as image-to-image - I2I, in contrast to text-to-image - T2I, can thus portray attributes in the original image that were not explicitly mentioned in the generation prompt. Moreover, it offers the opportunity to adjust and quantify the tradeoff between privacy and image quality. Modulating the strength of noise injected into the source image provides a natural knob that interpolates between near-identity transformations and unconstrained generation, enabling a controllable privacy–utility spectrum rather than a binary de-identification decision.

In this paper, we present an approach that uses text- and image-conditioned diffusion to generate medical images that preserve the semantic attributes of the original image when creating a synthetic, anonymized copy. We treat I2I generation as a stochastic, privacy-preserving transformation that produces a synthetic counterpart of real images while aiming to preserve downstream-relevant signals. We show that this approach better preserves both disease classes and physiological attributes in retina images from the Human Phenotype Project (HPP)^{12,13} and CXR images from the NIH CXR8 dataset¹⁴ using several complementary metrics: (i) pixel and perceptual similarity, (ii) embedding-space proximity, and (iii) agreement of trained clinical biomarker estimators between real images and their synthetic counterparts. We further show that the resulting synthetic images can be used to train biomarker estimator models using synthetic-only data with emphasis on the low-data regime. We additionally evaluate whether mixing real and synthetic images improves performance relative to real-only training under controlled subsampling, and we test cross-site transfer to an external cohort (UK Biobank¹⁵).

Finally, we characterize the tradeoff between similarity to the original image and preservation of semantic attributes when varying conditioning strength on the original image, providing empirical deidentification risk estimates.

Results

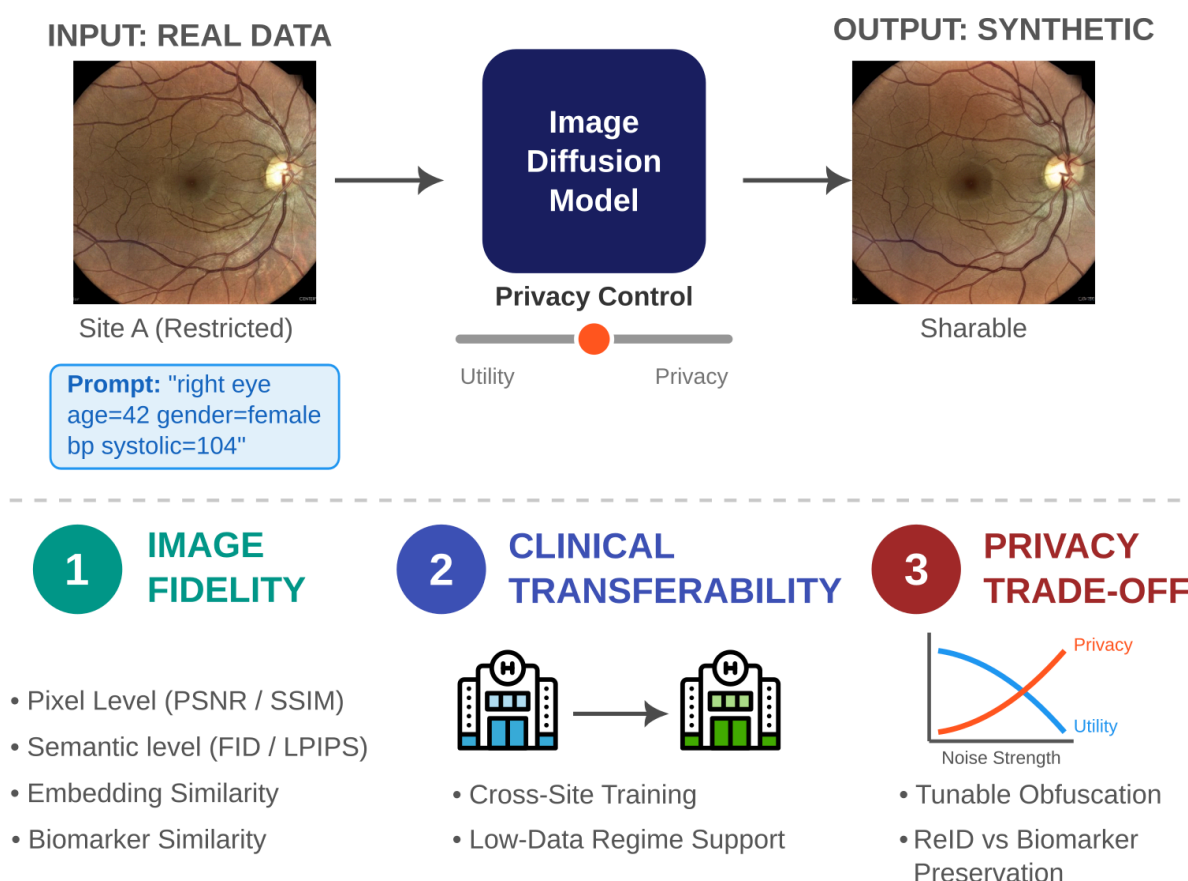


Figure 1: An overview of the study workflow: we fine-tune image-diffusion models on clinical imaging cohorts and generate synthetic images using either text-only (T2I) or image-and-text conditioning (I2I). We evaluate (1) image fidelity, clinical signal preservation, and downstream estimator performance across training regimes; (2) cross-site training performance focused on the low-data regime; and (3) the privacy-utility trade-off of image-conditioned diffusion - trading re-identification risk with biomarker preservation.

Synthetic Data Preserves Semantic Attributes

We trained Low-rank adapters (LoRA)¹⁶ for Stable Diffusion¹⁷ to generate retinal fundus photographs from the HPP^{12,13} and chest x-ray (CXR) radiographs from the NIH CXR8 dataset¹⁴ using a structured prompt (Supplementary Text 1), see methods for details. We then generated images using either

text-only conditioning (T2I) or text-and-image-based conditioning (I2I), which produces a synthetic counterpart of a specific source image while injecting controlled noise. Unless otherwise specified, all I2I-generated images were generated with strength=0.5 to balance attribute preservation and de-identification (see methods).

Generated images showed consistently greater similarity to their corresponding source images under I2I than under T2I, both in pixel space (Figure 2B, C - PSNR, SSIM) and in semantic latent measures (Figure 2D, E - LPIPS, FID). Representative fundus examples illustrate that I2I better retains clinically salient structure (e.g., vessel geometry and optic disc appearance) while allowing controlled variation, whereas T2I generations more frequently drift in global appearance (Figure 2A).

To assess whether improved fidelity translates into preservation of clinically relevant information, we compared the agreement of biomarker estimators between real images and their synthetic counterparts. Estimator outputs on I2I images were more strongly correlated with outputs on the original images than those on T2I images across multiple attributes (Figure 2F, G for retina and CXR images, respectively). Estimator predictions on I2I images also showed greater agreement with ground-truth labels than T2I predictions (Supplementary Figure 3), consistent with better preservation of downstream-relevant signal rather than superficial visual similarity alone. Notably, the advantage extends to biomarkers not included in the text-conditioning prompt, such as hemoglobin, creatinine, and waist circumference, indicating I2I preserves subtle signals encoded in the images. Finally, when embedding images using external self-supervised visual features (see methods), the generated I2I spreads across a larger manifold (Figure 2H, I for retina and CXR, respectively), supporting the interpretation of I2I as a controllable, local transformation.

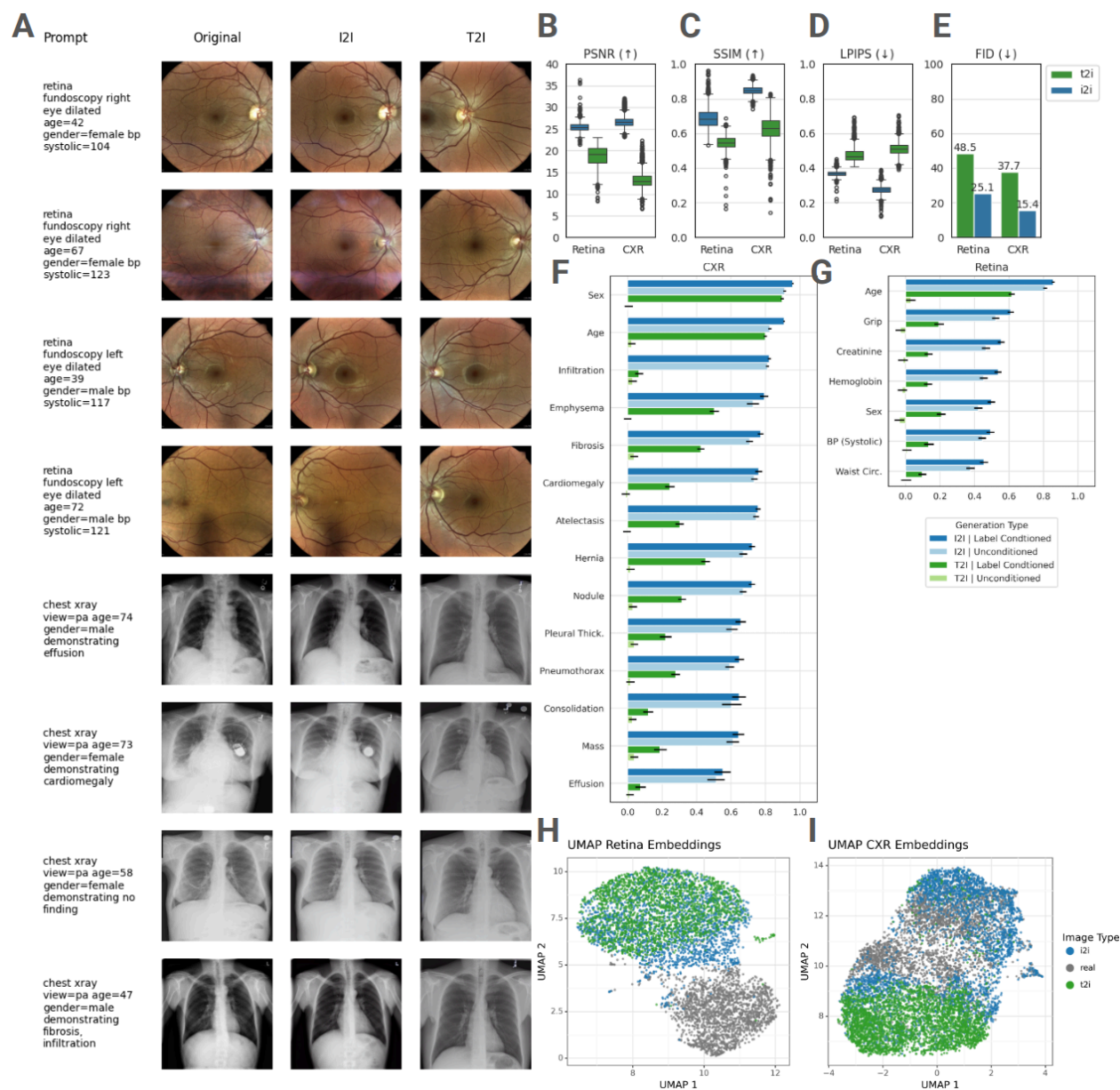


Figure 2: Comparison of image-to-image and text-to-image synthesis fidelity and attribute preservation. a, Representative examples of retinal fundus photographs (4 top rows) and chest X-rays (4 bottom rows) generated using the diffusion pipeline. Columns display the structured text prompt used for conditioning (left), the original real image, the image-to-image (I2I) synthetic counterpart generated with partial noise, and the text-to-image (T2I) synthetic image generated from pure noise using the same prompt. I2I retains source-specific anatomical structures (e.g., vessel geometry, disc and macula appearance, medical devices, and overall color) that are lost in T2I. b–e, Quantitative evaluation of image fidelity for Retina and Chest X-ray (CXR) datasets comparing I2I (blue) and T2I (green). Metrics include: b, Peak Signal-to-Noise Ratio (PSNR; higher is better); c, Structural Similarity Index Measure (SSIM; higher is better); d, Learned Perceptual Image Patch Similarity (LPIPS; lower is better); and e, Fréchet Inception Distance (FID; lower is better). Box plots in b–c indicate median (center line), interquartile range (box), and whiskers (1.5× IQR). f, g, Preservation of clinically relevant signals, measured by the Pearson correlation

coefficient between biomarker estimators applied to real images and their synthetic counterparts. f, Physiological biomarkers (e.g., Age, Hemoglobin, Blood Pressure) predicted from retinal images. g, physiological biomarkers and pathology classification in chest X-rays. I2I (blue bars) consistently yields higher agreement with source attributes than T2I (green bars). Error bars denote 95% confidence intervals derived via bootstrapping. Unconditioned refers to text conditioning only on the image modality (light blue and green for i2i and t2i, respectively), while label conditioned refers to prompts containing label information, as in panel a (darker blue and green, see Supplementary Text 1). h, i, UMAP visualizations of the image embedding space derived from a self-supervised encoder. h, For retina, local structure highlighting that I2I generations cluster more closely to their specific real source counterparts and vary more compared to T2I generations. i, For CXR, global distribution showing that I2I images (blue) occupy a similar manifold as real images (gray), whereas T2I images (green) exhibit a distribution shift.

Training biomarker estimators using synthetic images

We next compared training estimator models for multiple physiological attributes using (i) real images only, (ii) real images with standard augmentations, (iii) synthetic-only datasets generated with I2I or T2I, and (iv) mixtures of real and synthetic images at varying ratios. While models trained exclusively on synthetic images performed worse than those trained on real images, I2I-derived synthetic training consistently yielded higher predictive performance than T2I-derived synthetic training across all reported continuous phenotypes in terms of Pearson correlation (Figure 3A, B). In 6 out of 7 biomarkers, I2I synthetic-only training approached the performance of real-only training more closely than T2I, suggesting that the stronger source-consistency of I2I carries clinically informative variation that is not reliably preserved under text-only generation.

We further evaluated whether mixing synthetic images with real images improves performance when labeled data are limited. Across subsampled training sizes, adding synthetic images provided the largest relative benefit at small sample sizes, with diminishing gains as the amount of real training data increased. Training with a mix of 1:2 real to synthetic images improved performance when emulating a small training set for both I2I and T2I images in some attributes (age, Figure 3C, sex, Figure 3D, and the rest in Supplementary Figure 4). We observed no clear trend when training with mixes of 1:1-1:10 real:generated images (Supplementary figure 5).

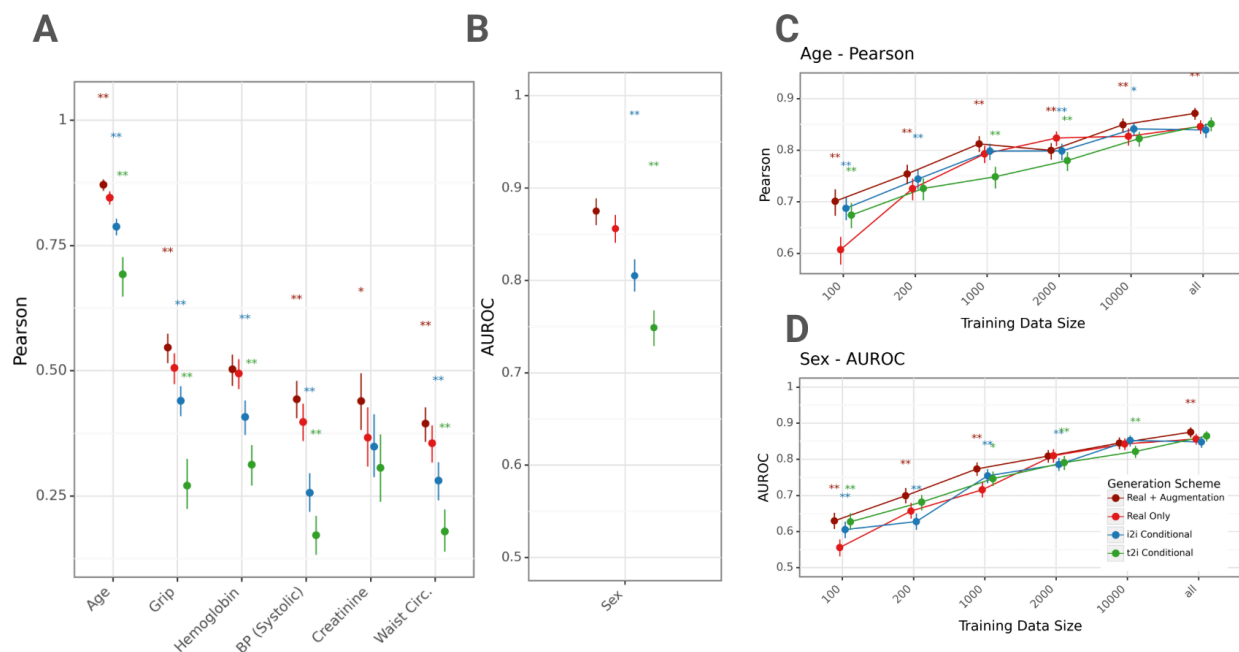


Figure 3: Classifier Performance by Training Data Mix

a,b, Comparison of training on real data (red), real data with augmentations (dark red), using only synthetic I2I images (blue), and only T2I synthetic images (green) on the HPP data, conditioned on age, sex, and systolic blood pressure. a, Pearson correlation for continuous outcomes, b, area under the receiver operating characteristic curve (AUROC) for binary outcomes. c,d, scaling data curves for training biomarker estimator models on the HPP data, age (c), and sex (d) using the same generation schemes. Asterisks denote statistical significance at the 0.001 (***), 0.01 (**), and 0.05 (*) levels against real only (red) after FDR correction.

Cross-Site Transfer Learning with Synthetic Data

Cross-site transfer, i.e., adapting models trained on the data of one cohort to another, is a major bottleneck in deploying medical image models^{3,6}. Although privacy-preserving learning frameworks such as differential privacy offer a privacy-preserving solution, they can be difficult to implement in practice and may introduce additional complexity or performance trade-offs. We thus wanted to test whether diffusion-generated synthetic images could serve as a practical alternative substrate for pretraining and transfer, as previously proposed^{6,18}.

Using real and synthetic counterpart fundus images from the HPP, we trained image estimators and used them as a base model for transfer learning the same attributes on retina images from the UKBiobank (UKB)¹⁵. The two cohorts differed in acquisition conditions and population characteristics (Supplementary Table 1-2), providing a clinically relevant domain shift. Models trained on a mix of real

and synthetic images or only synthetic I2I images performed comparably to transfer learning from either real or real images with traditional augmentations (Figure 4A-B).

This pattern was stable across UKB fine-tuning set sizes, with the clearest separations observed in the smallest training regimes (Figure 4C for age, Supplementary Figure 6 for other labels). These findings suggest that synthetic pretraining, particularly with I2I, which preserves source-specific structure, can support cross-cohort transfer at a level similar to real-image pretraining, while enabling workflows that reduce dependence on sharing sensitive source images. Notably, the benefit is most clinically relevant in low-data settings where acquiring labeled target-site data is challenging.

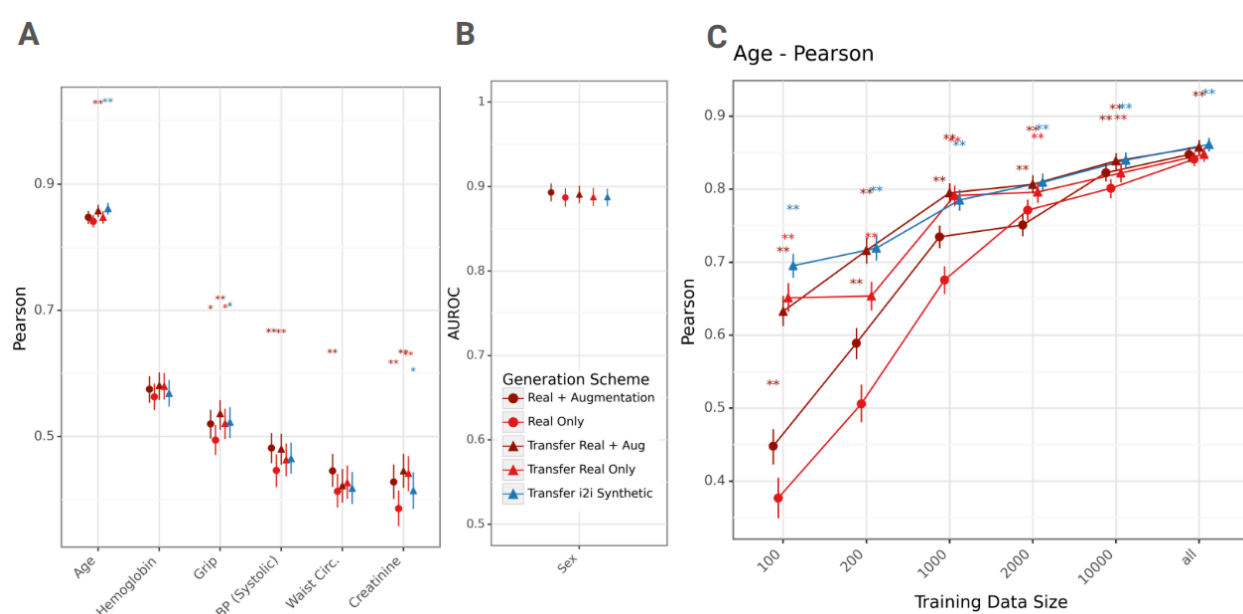


Figure 4: Cross-Site Transfer Learning Results

a,b Comparison of training on UKB real data (red), real data with augmentations (dark red), transfer learning from only synthetic I2I-based images, or from models trained on real HPP data or real HPP data with augmentations. a, Pearson correlation for continuous outcomes, b, area under the receiver operating characteristic curve (AUROC) for binary outcomes. c, scaling data curves for training age estimator models on the UKB data, using the same generation schemes. Asterisks denote statistical significance at the 0.001 (***) , 0.01 (**), and 0.05 (*) levels against real only (red) after FDR correction.

Privacy-Preserving Synthetic Data Generation

I2I-generated images show better fidelity but at the potential cost of being more similar to the source image and potentially revealing private information. We therefore quantified a privacy-utility tradeoff that is intrinsic to image-conditioned generation: stronger conditioning preserves more clinically

relevant information, but may increase reidentification risk. We sought to quantify the tradeoff between privacy and semantic attribute preservation with respect to the strength of image-conditioning. This strength translates to the amount of random noise added to the image before diffusion, where a strength of 0 is the real image (no noise), and 1 is equivalent to T2I generation (starting from pure noise). In practical terms, increasing strength progressively shifts I2I generation from a near-identity transformation toward a fully synthetic sample, providing a continuous control knob rather than a binary “de-identified/not de-identified” setting.

As the strength increases, images become more dissimilar to the original (Figure 5A) and preserve less of the original attributes (Figure 5B). For example, vessel branching structure and disc features become more dissimilar. Consistent with this, agreement between clinical estimator outputs on the source image and the generated image decreased monotonically with strength across multiple phenotypes. For example, the Pearson correlation for hemoglobin drops from 0.83 at strength 0.1 to 0.52 at 0.5 and 0.32 at 1.0 (T2I equivalent). For privacy estimation, we simulated a re-identification attack of 300 synthetic images from the test set against the full test set (Figure 5C). Specifically, for each synthetic image, we retrieved its nearest neighbor among candidate real images and evaluated whether the true source image was recovered as the top-1 match. While at strength 0.1, both structural (as measured using SSIM) and semantic distances (as measured using DINOv2 embeddings cosine distance) identify the original image as the nearest image in 100% and 95%, respectively. This drops to 31%, 15% at 0.5, and 2%, 2.7% at 0.8 for SSIM and embedding cosine distances, respectively. Notably, the drop in top-1 recovery with increasing strength mirrored the decline in attribute preservation, highlighting the operational privacy-utility tradeoff for downstream use cases. A similar analysis for CXR images (Supplementary Figure 7) shows that these images remain re-identifiable even at higher noise strengths, with top-1 accuracy of 63% at 0.9 and 17.3% at 1.0 (equivalent to T2I) for SSIM.

A distance to closest record analysis reveals that the distribution of distances between synthetic images and their real counterparts is similar to the distribution of distances between real images for strengths 0.5 and higher for retinal images (Supplementary Figure 8). This pattern is consistent with reduced re-identifiability at moderate-to-high strengths, where synthetic-to-real similarity becomes comparable to baseline real-to-real similarity within the cohort. We emphasize that these evaluations provide empirical estimates of re-identification risk under specific similarity-based attacks, rather than formal privacy guarantees, but they offer practical guidance for selecting strength values that balance de-identification and clinical utility.

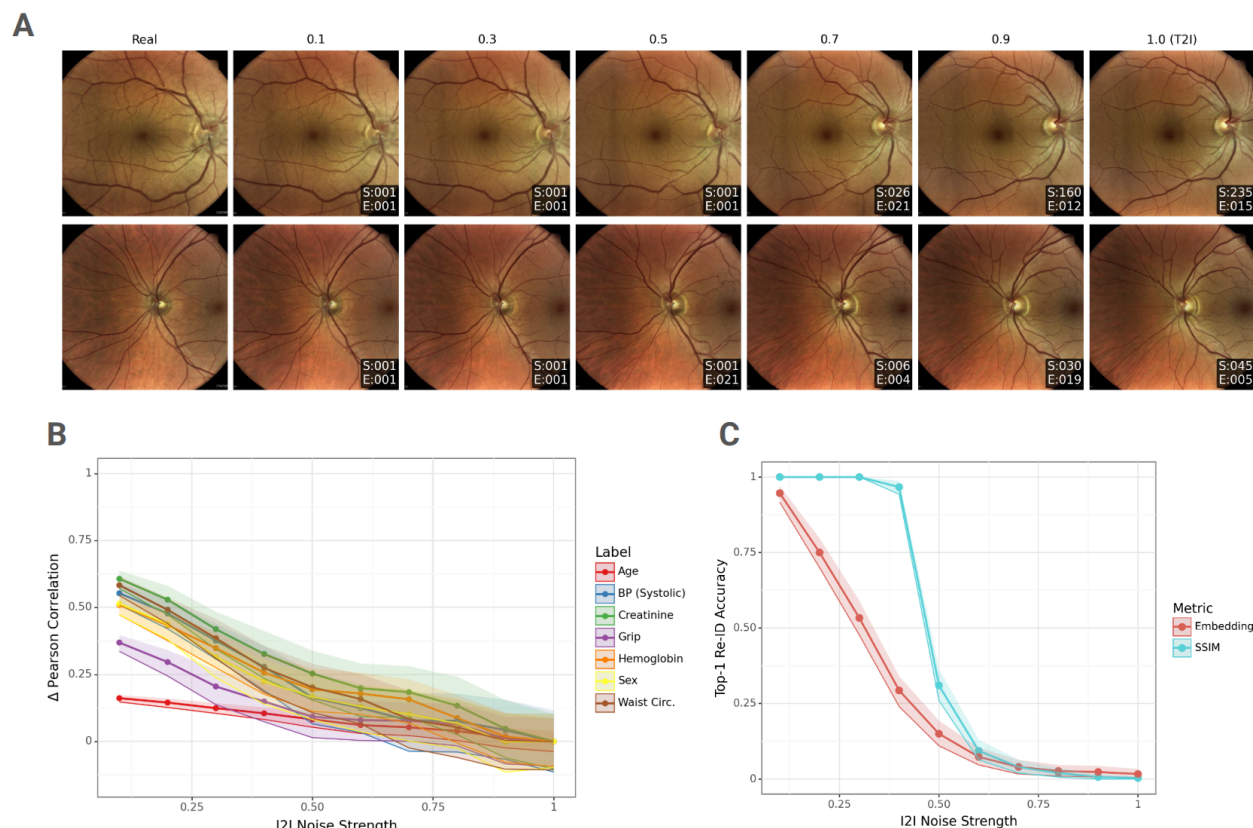


Figure 5: Empirical Privacy-Utility Tradeoff

a, Two example retina images with I2I generated using a progressive strength schedule from 0.1-1.0 where 1.0 is equivalent to T2I generation using the same prompt. Insets denote the rank of similarity to the original image in SSIM (S) or embedding cosine similarity (E). b, difference between Pearson correlation of predictions (numeric or log odds) with the same model predictions on the real image and a similar correlation calculated on the strength 1.0 (T2I equivalent) image. Higher values indicate that the synthetic image retains more original structural information than captured by the text prompt alone. c, Risk of identifying the real image by SSIM (blue) or embedding cosine similarity (red) among the test set from the generated image at that strength level (x-axis).

Discussion

In this paper, we describe the use of diffusion models for the synthesis of privacy-preserving medical images. We show that these models, particularly using image-and-text-based conditioning (I2I), produce high-fidelity images that preserve semantic attributes of the original image both with and without text-conditioning on the specific attributes. We apply this method to both retinal funduscopy images and chest X-rays and show that it enables training estimator models on a dataset of purely synthetic images with performance approaching training on real images. We also show that these synthetic images can be used for cross-site transfer learning and quantify the privacy-utility tradeoff with relation to the strength of noise applied to the conditioned-upon image. Together, these findings

position image-conditioned diffusion as a practical mechanism for generating shareable, de-identified imaging derivatives while retaining downstream utility.

Image-based conditioning methods are prevalent in computer vision since the introduction of SDEdit¹¹ and subsequent works¹⁹. However, medical image generation has thus far focused mostly on text-conditioning. As we show in this paper, extending diffusion generation to image-conditioned generation enables the preservation of subtle features in the image, leading to better preservation of attributes such as age and hemoglobin level. Similarly, while text-based conditioning on continuous attributes is possible and can be represented well (Supplementary Figure 2), we observed that these features, such as age and systolic blood pressure, are better preserved when the generation is anchored to a specific source image via I2I conditioning.

We present a systematic evaluation of training models utilizing synthetic images, both in-domain and out-of-domain, and with a range of data sizes. Similar to previous works^{1,4,20}, we observe that synthetic data can supplement real images, specifically in the low data regime. However, in this work, synthetic data mixing provides no significant added benefit when compared with traditional image augmentations on the same data. This finding suggests that, for institutions that already have access to the underlying real images, diffusion-based augmentation may not universally outperform well-established augmentation pipelines, and its benefits may be concentrated in specific phenotypes or data-scarce settings. This remains true in all but one of the cross-domain transfer experiments (age, 100 images). The use of strictly synthetic data can provide an opportunity for data sharing across institutions. Complementing the finding from above, cross-domain transfer learning using a model trained exclusively on I2I synthetic images performs comparably with a model trained on real images with traditional augmentations. This synthetic-only pretraining scenario is particularly relevant when source images cannot be shared due to governance restrictions, but model initialization or synthetic derivatives may be permissible.

We show that I2I synthetic images can have a low re-identification risk with a controllable trade-off for fidelity and privacy risk. Previous works considered even T2I-generated images re-identifiable²¹. This difference can be explained by a more relaxed definition of re-identification, which considers all images more similar than expected as memorized. While it is clear that even T2I images are sometimes memorized verbatim, our results indicate that the re-identification risk, as measured by nearest neighbor retrieval, is low even in moderate strength I2I images at the structural and semantic level in retinal images. It is important to emphasize that our privacy analysis focuses on whether a synthetic image can be matched back to its source record under similarity-based retrieval, rather than establishing formal privacy guarantees as provided, for example, by differential privacy frameworks. However, formal differential privacy approaches may require noise levels during training that would degrade subtle clinical signals this work aims to preserve. This empirical risk estimate can offer a practical basis for data-sharing decisions navigating the privacy-utility tradeoff.

Chest X-ray images tell a different story, in which re-identifiability is high in all I2I strengths and even at T2I generation. We speculate there are two main contributing factors. First, the combinatorial nature of CXR labels creates patient subsets that are identifiable by their structured data alone. Specifically, the unique combination of biological sex and radiographic findings uniquely identifies 3% of the test set and 11% of the set used for reidentification analysis. Second, unlike biomarkers and identifying features in retinal images, which are diffuse throughout the image, many of the labels in chest X-rays correspond to localized findings. This, in turn, makes random noise less effective at masking these image features, even at high noise levels.

This study is limited by its choice of modalities and labels for each modality. First, extensions to other modalities, including volumetric imaging modalities, are required to span the full breadth of medical imaging. Second, our evaluation emphasizes phenotype prediction and similarity-based re-identification-style retrieval. Future work should examine additional privacy threats (e.g., membership inference, stronger linkage attacks) and clinical endpoints (e.g., downstream diagnostic classification and calibration). Third, diffusion generation may preserve site-specific acquisition signatures in ways that are not fully captured by the presented metrics, and further auditing across devices and protocols will be important for deployment. Fourth, we used a single base model for image generation (SD 1.4) and estimator models (BiT-ResNet v2 101). For the generative model, as both retinal and CXR images are significantly out of the pretraining distribution of natural images, we expect little qualitative difference with newer models. For the estimator model, while hyperparameter tuning showed improved performance over VIT-based models for estimation on real images, the robustness of the results for synthetic training remains to be verified. Additionally, as healthcare data sharing is subject to regulatory oversight, it remains to be seen how synthetic data is treated by data governance standards.

Image-conditioned diffusion provides a pragmatic path to generating de-identified medical image derivatives that retain clinically relevant information for downstream modeling. By making the privacy-utility tradeoff explicit and tunable through conditioning strength, this approach supports both low-data learning and cross-cohort transfer in settings where sharing raw images is infeasible. Future work should extend these evaluations to additional modalities and clinical tasks, and integrate stronger privacy auditing frameworks to guide adoption under real-world governance constraints.

Methods

Datasets

We used 3 distinct datasets throughout this work.

Human Phenotype Project

The Human Phenotype Project (HPP) is a large-scale, prospective cohort encompassing deep clinical and molecular phenotyping of 27,916 participants in Israel^{12,13}. Here, we utilized the retinal fundus images for 10,010 patients together with clinical data recorded up to or at the same visit as retina funduscopy. We extracted their age at the visit, sex, heart rate, blood pressure, hip and waist circumference, grip strength, and body mass index (BMI). We additionally incorporated comprehensive laboratory blood panels, specifically: hemoglobin, platelets, mean corpuscular volume (mcv), white blood cell count (wbc), glucose, hba1c%, creatinine, ferritin, albumin, ldl-, hdl-, non-hdl, and total-cholesterol, and triglycerides. We included the first and subsequent visits, which included a retinal funduscopy for a total of 23,259 images, with 2.3 (0.8) images and 1.2 (0.4) visits per person (presented as mean (SD)). Patients were split into a train and test sets by percentage - 90%, 10% respectively, such that the same patient images appeared in only one of the splits. 5-fold cross-validation splits were derived from the training set, with patients appearing in only the train or validation fold, similarly.

UK Biobank

The UK Biobank (UKB) includes over 95,000 patients with retinal images. Data from the UK Biobank was accessed under application ID 28784. Data was downloaded in 08/2022. We included 11,030 patients from the first instance (0), including only the first visit for each patient, for a total of 13,191 images. In accordance with the data processing detailed above for the HPP, we included the most recent value for each attribute up to the funduscopy visit. We included the attributes age, sex, hemoglobin, blood pressure, grip strength, waist circumference, and creatinine.

NIH CXR-8

NIH CXR-8 is a large public dataset of chest X-rays with 108,948 frontal-view X-ray images of 32,717 unique patients with 14 text-derived image labels - Atelectasis, Consolidation, Infiltration, Pneumothorax, Edema, Emphysema, Fibrosis, Effusion, Pneumonia, Pleural_thickening, Cardiomegaly, Nodule, Mass, and Hernia¹⁴. Hernia and Pneumonia labels were removed due to their low frequency. The dataset was retrieved from the NIHCC box - <https://nihcc.app.box.com/v/ChestXray-NIHCC/file/220660789610> at 25/04/2025. We filtered the data by taking the first image per person, keeping only PA view positions, and downsampling images with no findings by taking only 30% of these samples. Images were split into a train and test set by percentage - 80%, 20% respectively, as only one image was taken per patient.

Data Processing

Images were resized to (512,512) and normalized to the range (-1,1). We encoded categorical attributes using one-hot encoding, and continuous attributes using min-max scaling according to the distribution of the HPP training set for both HPP and UKB.

Generative Model Architecture and Training

Stable Diffusion 1.4 is an established high-fidelity diffusion image generation model. It enables the use of both text and images when generating images. We finetuned the model on images from the train set of each of the datasets using low-rank adapters (LoRA) through the HuggingFace Diffusers library²². The LoRA config was rank=64, alpha=32, and was applied to the UNet transformer blocks. We trained adapters on 20,000 steps using the AdamW optimizer, a constant learning rate of 1e-4, a batch size of 12, mixed fp16 precision, and no gradient accumulation.

LoRA fine-tuning procedure. We fine-tuned Stable Diffusion 1.4¹⁷ using Low-Rank Adaptation (LoRA)¹⁶ applied to the UNet attention modules, training a small set of adapter parameters while keeping the base model weights frozen. We used a fixed image resolution of 512×512 throughout. LoRA rank was set to 64 with $\alpha=32$, and adapters targeted the attention projection layers (query/key/value and output projections) following standard practice. Training was performed using mixed precision (fp16) to reduce memory footprint, and checkpoints were saved at regular intervals together with validation generations to monitor fidelity and conditioning adherence. All LoRA runs were executed with the same data split boundaries as the downstream estimator experiments to ensure that synthetic generation was performed only from training images (for I2I) and that evaluation sets were never used to tune generation.

Estimator models

All estimator models were based on a ResNet v2 101 pretrained using the BiT recipe²³, as available on TIMM²⁴. We added model heads for each attribute, using cross-entropy loss for classification and mean-squared error loss for regression heads. Losses were masked where the attribute value was unavailable. We trained for 20,000 steps using early stopping on validation loss (with a patience of 10), AdamW as the optimizer, with constant learning rates separate for the pretrained base and heads. Configurations, as well as hyperparameter tuning specifications on 5-fold cross-validation, are found in supplementary methods. Baseline performance for models trained on the full training set is shown in Supplementary Figure 1.

Synthetic Data Generation

Synthetic images were generated from the fine-tuned diffusion models using two complementary regimes: text-to-image (T2I) and image-and-text conditioned generation (I2I). In T2I, images were sampled from noise conditioned only on the structured prompt describing modality and available attributes. In I2I, each synthetic image was generated as a stochastic transformation of a specific real source image: we first perturbed the source image by adding Gaussian noise and then performed conditional denoising guided by the same structured prompt. The strength parameter controls how much of the diffusion process is re-applied to the source image. At some strength s , the source is corrupted to an intermediate point in the forward noise schedule and then denoised from that point. Lower strengths preserve more of the original structure; higher strengths allow greater deviation. At $s = 0$, the output is the source image itself; at $s = 1$, the generation starts from pure noise (equivalent to T2I). Unless otherwise specified, all I2I results use $s = 0.5$. All generated images were stored alongside generation metadata (prompt text, conditioning variables, strength, and source image identifier) to support reproducibility, downstream training, and privacy auditing.

Evaluation Metrics

Estimator models were evaluated using Pearson correlation for numerical attributes modeled with regression heads, and area under the receiver operating characteristic (auROC) for classification models. Metrics were calculated in bootstrap samples of the test set to construct 95% confidence intervals using the percentile bootstrap approach.

Image generation was evaluated on 2 levels: pixel-level and attribute-level. Pixel-level evaluation utilized both classic fidelity metrics - peak signal to noise ratio (PSNR), structural similarity index measure (SSIM)²⁵; as well as neural-network based metrics - Fréchet inception distance (FID)²⁶, learned perceptual image patch similarity (LPIPS, using a VGG backbone)²⁷, all were calculated using pyiqa version 0.14²⁸.

Uncertainty estimation and statistical testing. For estimator performance metrics (Pearson r for regression targets; AUROC for binary classification), we report uncertainty using bootstrap resampling of the held-out test set to construct 95% confidence intervals via the percentile bootstrap. For comparisons between training regimes (e.g., I2I vs T2I, synthetic-only vs real-only, and mixed vs augmented), we used paired bootstrap procedures whenever predictions were available for the same test examples across conditions, enabling direct estimation of the distribution of performance differences and the corresponding bootstrap hypothesis test. For multi-phenotype analyses, we report both nominal p -values and multiple-comparison-adjusted results, using Benjamini-Hochberg false discovery rate, where appropriate.

Prompting and conditioning protocol (T2I vs I2I)

For both fundus and CXR modalities, we generated prompts using a structured template that encodes modality- and view-specific descriptors together with available clinical attributes (Supplementary Text 1). For retinal images, prompts included laterality (left/right) together with the conditioning variables used in a given experiment (e.g., age+sex+systolic blood pressure; or no labels for unconditional models). For CXR images, prompts included view information and, in label-conditioned experiments, the presence of radiographic findings using a standardized list. Text-to-image (T2I) generations were produced solely from these prompts. Image-and-text conditioning (I2I) additionally used a real image as the starting point, which was perturbed by injecting Gaussian noise according to a user-specified strength parameter; we interpret strength=0 as the unmodified real image and strength=1 as fully noise-initialized sampling, equivalent to T2I generation. Unless otherwise specified, we generated multiple stochastic variations per source image by varying the random seed while keeping the prompt and conditioning settings fixed.

Image generation hyperparameters, sampling, and deduplication safeguards

For each LoRA model, we generated synthetic images using both T2I and I2I pipelines implemented in the Hugging Face diffusers framework. Unless otherwise specified, we used the DDIM scheduler and low classifier-free guidance (guidance scale ≈ 2) based on generation-parameter optimization sweeps (Supplementary Methods). For I2I generation, we evaluated a range of noise strengths spanning low to high transformation, and for each strength, we produced multiple stochastic variants per source image by varying the random seed. To reduce trivial duplicates and to support downstream auditing, we stored all generation parameters (prompt, strength). Generated images were written with accompanying metadata to enable full reproducibility of each synthetic record and to support linkage analyses.

Synthetic dataset construction and relabeling protocol

To enable controlled comparisons across training regimes, we constructed synthetic datasets by generating a fixed number of synthetic images per real training image, storing generation metadata (prompt, strength and source image identifier where applicable) alongside the resulting file paths. All synthetic-only and mixing experiments were derived from the same underlying train/validation/test partitioning strategy as the corresponding real datasets, and we ensured that synthetic images derived from a given subject were assigned exclusively to the same split as that subject's real images to prevent leakage.

Training regimes, mixing experiments, and cross-site transfer setup

We evaluated estimator training under four primary regimes: real-only; real with conventional augmentations (rotation, photometric perturbations, and cropping; Supplementary Methods, Supplementary Figure 9); synthetic-only (T2I or I2I); and real:synth mixtures. For mixing, we constructed datasets with predefined synthetic-to-real ratios (e.g., 1x, 2x, 5x, 10x synthetic per real) by sampling synthetic images uniformly per real image to maintain balanced coverage across subjects and phenotypes. To characterize performance in the low-data regime, we additionally trained on nested subsamples of increasing size, where smaller training sets were strict subsets of larger ones, enabling fair scaling comparisons across conditions. For cross-site transfer, we pretrained estimators on HPP under each regime (real-only, real+augmentation, synthetic-only, and mixed) and then fine-tuned on UKB using matched phenotype definitions and preprocessing, evaluating performance as a function of UKB training set size. Unless otherwise specified, all comparisons used identical model architectures, optimization settings, early stopping criteria, and evaluation pipelines across regimes.

Quantifying structural and semantic preservation beyond pixel metrics

We quantified higher-level “semantic” similarity using self-supervised visual embeddings (e.g., DINOv2²⁹): images were embedded with a frozen encoder, and we computed cosine distances between real images and between real images and synthetic samples. These embeddings also served for UMAP visualization to compare real vs synthetic manifold overlap and to probe whether I2I remains a local transformation relative to T2I.

Privacy evaluation

We evaluated privacy risk using similarity-based linkage attacks that attempt to match a synthetic image back to its originating real image. For each synthetic image, we retrieved its nearest neighbor among candidate real images using multiple similarity measures (pixel-space as SSIM; and embedding-space distances using the frozen encoder), and recorded whether the true source was recovered as top-1 under each measure. To contextualize these results, we compared synthetic-to-real distance distributions against a real-to-real baseline (“distance to closest record”) computed within the cohort, and repeated the analysis across I2I strengths to directly map the privacy–utility curve. As basic safeguards against trivial duplicates, we stored full generation metadata (prompt, strength, and source identifier for I2I), performed exact-duplicate checks (e.g., filename hashing).

Computational Resources

All training, synthetic images generation and evaluation was performed on the Weizmann GPU cluster consisting of L40 and A100 NVIDIA GPUs. We used Python 3.11, CUDA 11.8, and Pytorch 2.7 for all code executed.

Ethics Declarations

The Human Phenotype Project (HPP) cohort study is conducted according to the principles of the Declaration of Helsinki and was approved by the Institutional Review Board (IRB) of the Weizmann Institute of Science (reference number 1719-1). All HPP participants signed an informed consent form to be included in the cohort, and all identifying details of the participants were removed prior to computational analysis. Data from the UK Biobank was accessed under application ID 28784, complying with the UK Biobank's generic ethical approval from the NHS National Research Ethics Service. The NIH CXR-8 dataset is a fully de-identified, publicly available dataset and is therefore exempt from additional IRB review.

Data Availability

- **Human Phenotype Project (HPP):** Data from the HPP include personal health information and, in compliance with institutional review board regulations. The data can be made accessible to researchers from universities and other research institutions via a secure cloud-based platform at <https://humanphenotypeproject.org/>. Interested bona fide researchers should contact info@pheno.ai to obtain instructions for accessing the data.
- **UK Biobank:** Data are available to authorized researchers upon application to the UK Biobank (<https://biobank.ndph.ox.ac.uk/>). The data used in this study were accessed under application ID 28784.
- **NIH CXR-8:** The dataset is publicly available and can be downloaded from the NIH Clinical Center at <https://nihcc.app.box.com/v/ChestXray-NIHCC>.
- **Generated Images:** 2500 unique HPP patient i2i generated retinal fundus images complete with original image labels, are available on HuggingFace for downstream use at <https://huggingface.co/datasets/doronys/synthmed-retina-hpp-sample>

Code Availability

Code to train LoRA adapters and biomarker estimator models, as well as to reproduce the analyses in figures 2 and 6, is available at <https://github.com/dst1/SynthMed-demo>. Trained LoRA adapters can be found on HuggingFace - <https://huggingface.co/doronys/synthmed-loras>.

Bibliography

1. Sun, Y. *et al.* A data-efficient strategy for building high-performing medical foundation models. *Nat. Biomed. Eng.* **9**, 539–551 (2025).

2. Bluethgen, C. *et al.* A vision–language foundation model for the generation of realistic chest X-ray images. *Nat. Biomed. Eng.* **9**, 494–506 (2025).
3. Han, T. *et al.* Breaking medical data sharing boundaries by using synthesized radiographs. *Sci. Adv.* **6**, eabb7973 (2020).
4. Wang, J. *et al.* Self-improving generative foundation model for synthetic medical image generation and clinical applications. *Nat. Med.* <https://doi.org/10.1038/s41591-024-03359-y> (2024) doi:10.1038/s41591-024-03359-y.
5. Diaz-Asper, C., Bongo, L. A. & Elvevåg, B. Navigating the tradeoff between personal privacy and data utility in speech anonymization for clinical research. *Npj Digit. Med.* **8**, 616 (2025).
6. DuMont Schütte, A. *et al.* Overcoming barriers to data sharing with medical image generation: a comprehensive evaluation. *Npj Digit. Med.* **4**, 141 (2021).
7. Han, C. *et al.* GAN-based synthetic brain MR image generation. in *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)* 734–738 (IEEE, Washington, DC, 2018). doi:10.1109/ISBI.2018.8363678.
8. Ktena, I. *et al.* Generative models improve fairness of medical classifiers under distribution shifts. *Nat. Med.* **30**, 1166–1173 (2024).
9. Lang, O. *et al.* Using generative AI to investigate medical imagery models and datasets. *eBioMedicine* **102**, 105075 (2024).
10. Poplin, R. *et al.* Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nat. Biomed. Eng.* **2**, 158–164 (2018).
11. Meng, C. *et al.* SDEdit: Guided Image Synthesis and Editing with Stochastic Differential Equations. Preprint at <https://doi.org/10.48550/arXiv.2108.01073> (2022).
12. Reicher, L. *et al.* Deep phenotyping of health–disease continuum in the Human Phenotype Project. *Nat. Med.* **31**, 3191–3203 (2025).
13. Shilo, S. *et al.* 10 K: a large-scale prospective longitudinal study in Israel. *Eur. J. Epidemiol.* **36**, 1187–1194 (2021).
14. Wang, X. *et al.* 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)* 3462–3471 (2017). doi:10.1109/CVPR.2017.369.

15. Sudlow, C. *et al.* UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.* **12**, e1001779 (2015).
16. Hu, E. J. *et al.* LoRA: Low-Rank Adaptation of Large Language Models. Preprint at <https://doi.org/10.48550/ARXIV.2106.09685> (2021).
17. Rombach, R., Blattmann, A., Lorenz, D., Esser, P. & Ommer, B. High-Resolution Image Synthesis with Latent Diffusion Models. Preprint at <https://doi.org/10.48550/arXiv.2112.10752> (2022).
18. Liu, C. *et al.* Can Medical Vision-Language Pre-training Succeed with Purely Synthetic Data? Preprint at <https://doi.org/10.48550/arXiv.2410.13523> (2025).
19. Pérez-García, F. *et al.* RadEdit: stress-testing biomedical vision models via diffusion image editing. in vol. 15070 358–376 (2025).
20. Kim, J. M., Bader, J., Alaniz, S., Schmid, C. & Akata, Z. DataDream: Few-shot Guided Dataset Generation. Preprint at <http://arxiv.org/abs/2407.10910> (2024).
21. Dar, S. U. H. *et al.* Unconditional latent diffusion models memorize patient imaging data. *Nat. Biomed. Eng.* 1–15 (2025) doi:10.1038/s41551-025-01468-8.
22. von Platen, P. *et al.* Diffusers: State-of-the-art diffusion models. (2026).
23. Kolesnikov, A. *et al.* Big Transfer (BiT): General Visual Representation Learning. Preprint at <https://doi.org/10.48550/arXiv.1912.11370> (2020).
24. Wightman, R. PyTorch Image Models. <https://doi.org/10.5281/zenodo.4414861> (2026).
25. Wang, Z., Bovik, A. C., Sheikh, H. R. & Simoncelli, E. P. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.* **13**, 600–612 (2004).
26. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B. & Hochreiter, S. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. in *Advances in Neural Information Processing Systems* vol. 30 (Curran Associates, Inc., 2017).
27. Zhang, R., Isola, P., Efros, A. A., Shechtman, E. & Wang, O. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. Preprint at <https://doi.org/10.48550/arXiv.1801.03924> (2018).
28. Chen, C. & Mo, J. IQA-PyTorch: PyTorch Toolbox for Image Quality Assessment. (2022).
29. Oquab, M. *et al.* DINOv2: Learning Robust Visual Features without Supervision. Preprint at <https://doi.org/10.48550/arXiv.2304.07193> (2024).