Overcoming Barriers to Data Sharing with Medical Image Generation: A Comprehensive Evaluation

August DuMont Schütte[1,5,*], Jürgen Hetzel[2,3], Sergios Gatidis[4], Tobias Hepp[4,5], Benedikt Dietz[1], Stefan Bauer[5,6], and Patrick Schwab[7]

[1] *ETH Zurich, Switzerland*
[2] *Department of Medical Oncology and Pneumology, University Hospital of Tübingen, Germany*
[3] *Department of Pneumology, Kantonsspital Winterthur, Switzerland*
[4] *Department of Radiology, University Hospital of Tübingen, Germany*
[5] *Max Planck Institute for Intelligent Systems, Tübingen, Germany*
[6] *CIFAR Azrieli Global Scholar*
[7] *F. Hoffmann-La Roche Ltd, Basel, Switzerland*
[*] *Corresponding author, augustschdmnt@gmail.com*

December 8, 2020

1 Abstract

Privacy concerns around sharing personally identifiable information are a major practical barrier to data sharing in medical research. However, in many cases, researchers have no interest in a particular individual's information but rather aim to derive insights at the level of cohorts. Here, we utilize Generative Adversarial Networks (GANs) to create derived medical imaging datasets consisting entirely of synthetic patient data. The synthetic images ideally have, in aggregate, similar statistical properties to those of a source dataset but do not contain sensitive personal information. We assess the quality of synthetic data generated by two GAN models for chest radiographs with 14 different radiology findings and brain computed tomography (CT) scans with six types of intracranial hemorrhages. We measure the synthetic image quality by the performance difference of predictive models trained on either the synthetic or the real dataset. We find that synthetic data performance disproportionately benefits from a reduced number of unique label combinations and determine at what number of samples per class overfitting effects start to dominate GAN training. Our open-source benchmark findings also indicate that synthetic data generation can benefit from higher levels of spatial resolution. We additionally conducted a reader study in which trained radiologists do not perform better than random on discriminating between synthetic and real medical images for both data modalities to a statistically significant extent. Our study offers valuable guidelines and outlines practical conditions under which insights derived from synthetic medical images are similar to those that would have been derived from real imaging data. Our results indicate that synthetic data sharing may be an attractive and privacy-preserving alternative to sharing real patient-level data in the right settings.

2 Introduction

Sharing sensitive data under strict privacy regulations remains a crucial challenge in advancing medical research [1]. By accessing large amounts of collected data, there have been impressive research results in a range of medical fields such as genetics [2], radiomics [3, 4], neuroscience [5], diagnosis [6, 7, 8], patient outcome prediction [9, 10] or drug discovery [11, 12]. Particularly deep learning systems, composed of millions of trainable parameters, require large amounts of data to learn meaningful representations robustly [13]. Aside from quantity, the quality of the available patient-level data is particularly essential for medical research [14]. Highly diverse and well-curated training data empowers researchers to produce generalisable insights and reduces the risk of biased predictions when applied in practice.

It is especially difficult to share and distribute medical data due to privacy concerns and the potential abuse of personal information [15]. To overcome these privacy concerns there has been an impressive number of large-scale research collaborations to pool and curate de-identified medical data for open-source research purposes [16, 17, 18]. Nevertheless, most medical data is still isolated and locally stored in hospitals and laboratories due to the concerns associated with sharing patient data [19]. In many countries, privacy laws inhibit medical data sharing [20], and potentially available de-identification methods lack guarantees as de-identified data can, in some cases, be linked back to individuals [21, 22].

In medical research, information is often analysed at the level of cohorts rather than individuals. A potential solution to the medical data sharing bottleneck, is therefore, the generation of synthetic patient data that, in aggregate, has similar statistical properties to those of a source dataset without revealing sensitive private information about individuals. While synthetic data can be generated for all kinds of data modalities, we focus on the particularly important medical imaging domain in this work.

Recently, new generative machine-learning approaches, such as Generative Adversarial Networks (GANs), have demonstrated the capability to generate realistic, high-resolution image datasets [23]. In GANs, two neural networks play an adversarial game against each other. The generator (G) tries to learn the real data distribution while the discriminator (D) estimates the probability of a sample belonging to the real training set, as opposed to having been generated by G [24]. If training is stable, the model converges to a point where D can no longer discriminate between real and synthetic data [25]. When each neural

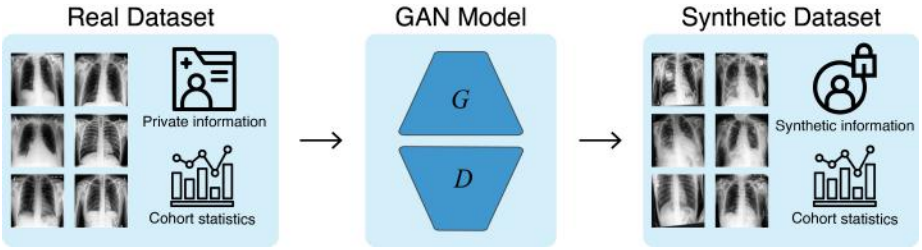

Figure 1: **Synthetic medical imaging dataset generation to overcome data sharing barriers.** We train our GAN models with real medical imaging data, to generate the corresponding synthetic images. The synthetic dataset ideally no longer contains private information about individual patients while in aggregate, maintaining the real training cohort's statistical properties.

network is composed of a convolutional neural network (CNN), GANs have demonstrated state of the art image generation capabilities [26, 27].

Within the medical imaging domain, there are several works demonstrating the generation of realistic synthetic data, among others, retinal images [28, 29], skin lesions [30, 31, 32], hematoxylin and eosin (H&E) stained breast cancer tissue in digital pathology [33], x-ray mammographs [34], chest radiographs [35, 36] and brain tumor magnetic resonance imaging (MRI) [37]. Other works such as [38] focus on utilizing GANs for image-to-image translations within the medical domain.

To the best of our knowledge, there is no work aimed at providing a comprehensive benchmark analysis for the generation of synthetic medical images across different GAN architectures and data modalities. We offer guidelines for the use of GAN models to fully synthesise realistic datasets as a potentially viable approach to privacy-preserving data sharing, and make the following contributions:

- We develop an open benchmark to analyse the generation of synthetic medical images when varying the number of label combinations, the number of samples per label combination, and the spatial resolution level present in the dataset.

- We present valuable guidelines for the effective generation of medical image datasets by evaluating our open-source benchmark on a reference GAN model and our newly proposed GAN architecture for two different data modalities.[1]

- We additionally analyse privacy considerations, assess the causal contributions of predictive models trained on the synthetic datasets, and finally conduct a large-scale reader study in which trained radiologists discriminate between real and synthetic medical images.

[1] The computational cost of our medical imaging benchmark amounts to approximately $16,280$ GPU-hours (678 GPU-days) on NVIDIA's Pascal P100 GPU.

3 Results

3.1 Overview of approach

Both datasets consist of binary multi-label classes. Each chest x-ray image can have a combination of the following 13 labels: Enlarged cardiomediastinum, cardiomegaly, lung opacity, lung lesion, edema, consolidation, pneumonia, atelectasis, pneumothorax, pleural effusion, pleural other, fracture, support device or the no finding class. The brain CT scans can consist of a combination of five different hemorrhage types: Epidural, subarachnoid, subdural, intraparenchymal and intraventricular or the no finding class. We randomly split each patient cohort into training, validation and test set within strata of radiology findings. We developed all GAN models on the training datasets and stopped GAN training when the quality between real and synthetic images converged. Next, we generated synthetic data for the train and validation folds by conditioning on the respective labels. In all settings, we used a pre-trained densenet-121 CNN as a predictive model, with the mean area under the receiver operating characteristics curve over all labels ($\overline{AUC}$) as the evaluation metric. For each classifier, we stopped training when the validation $\overline{AUC}$ converged. After the real and synthetic predictive models are fully developed, we evaluated both on the separate, real data test fold to compute the difference in performance: $\overline{AUC}_{\text{real}} - \overline{AUC}_{\text{syn}}$. We repeated all experiments multiple times with varying random initialisation of the deep learning systems, allowing us to perform statistical tests on whether the distribution of $\overline{AUC}_{\text{real}} - \overline{AUC}_{\text{syn}}$ scores differs at different benchmark settings. Additionally, we compared the predictive models' feature importance when trained on either real or synthetic datasets. We addressed privacy concerns by analysing differences between synthetic images and the most closely matching nearest neighbour images from the entire training dataset. Finally, we performed a large-scale reader study in which we asked trained radiologists to label a mixture of real and generated images.

3.2 Model performance

To accurately assess the potential of synthetic data, we analysed two model architectures across two different datasets for our benchmark. The prog-GAN model refers to the progressive GAN as a reference model, as it is still commonly used for medical image generation [32, 36]. The cpD-GAN refers to our novel and improved model that we specifically developed for this benchmark. To assess the generalisation capabilities, we did not fine-tune across different benchmark settings, only when increasing the resolution, we make the necessary changes to the network architectures.

The prog-GAN achieved an average $\overline{AUC}_{\text{real}} - \overline{AUC}_{\text{syn}}$ score of 0.0495 (±0.0276) across all settings on the chest radiograph dataset and 0.1367 (±0.0324) across all brain scans experiments. These scores were substantially improved with the cpD-GAN that achieves 0.0206 (±0.0100) on the chest x-ray settings and 0.0650 (±0.0198) on the brain hemorrhage dataset experiments.

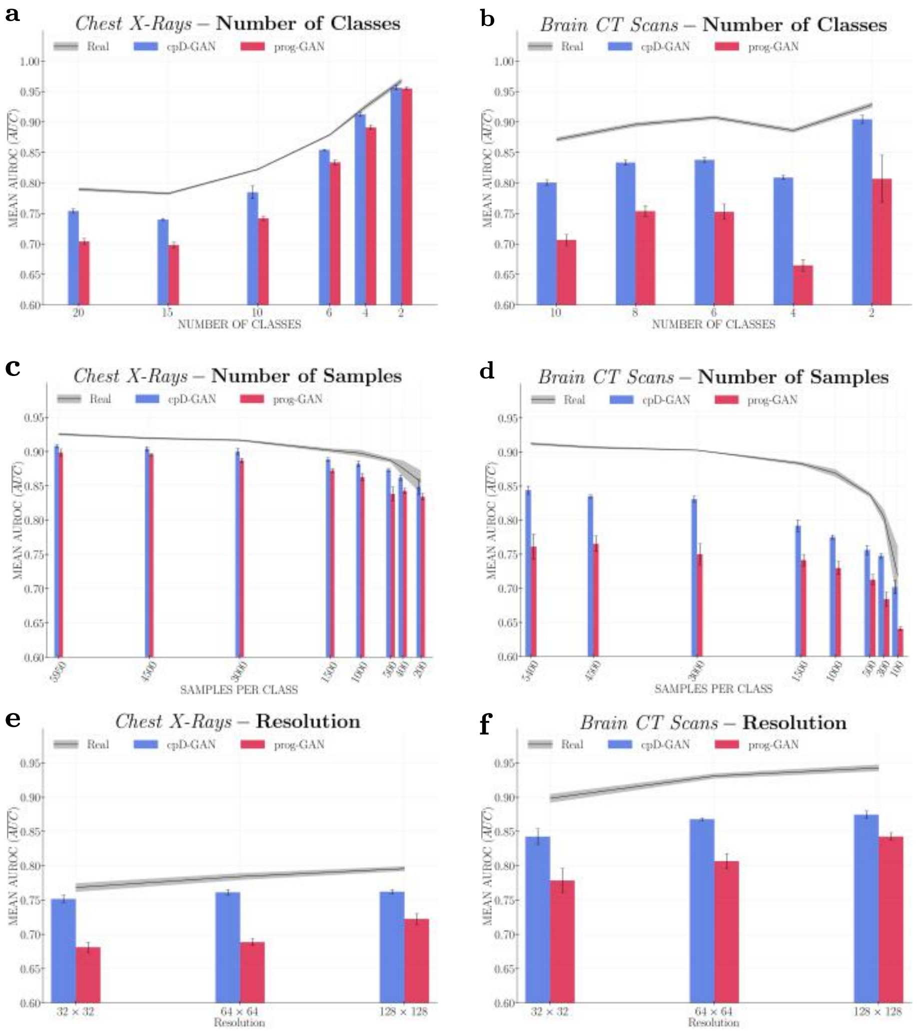

Figure 2: **Benchmark results.** In each figure, the mean area under the receiver operator characteristic curve scores obtained when training classifiers on real data ($\overline{AUC}_{\text{real}}$) are indicated by the black line with the shaded area representing the standard deviation across repeated experiments. The bar plots represent the mean AUC scores achieved when training classifiers on synthetic data ($\overline{AUC}_{\text{syn}}$) generated by the cpD-GAN (blue) and prog-GAN (red), while the error bars indicate the standard deviation. The subfigures show the changes in predictive performance observed when varying the number of classes (or label combinations) **a)** for chest radiographs and **b)** for brain computed tomography (CT) scans, the number of samples per class **c)** for chest radiographs and **d)** for brain CT scans, and the image resolution **e)** for chest radiographs and **f)** for brain CT scans.

3.3 Benchmark findings

We evaluated the model performance across three benchmark dimensions, detailed in Table S1. First, we varied the number of unique binary label combinations (which we also refer to as number of classes) included in the dataset. Next, we fixed the present classes and assessed how changes in the number of samples for each group of findings impacted performance. While we evaluated the first two benchmark settings at a resolution of 32×32 pixels, we finally analysed how increasing the resolution to 64×64 and 128×128 pixels affected our scores. We only performed changes across a single dimension at a time to ensure no confounding factors can impact training.

Impact of number of classes. The classification performance on both real and synthetic data increased when we lowered the number of unique present classes. We reason that the complexity of the predictive task decreases with fewer label combinations, resulting in higher $\overline{AUC}$ scores. However, as can be seen in Figure 2a and 2b, the differences in $\overline{AUC}_{\text{real}} - \overline{AUC}_{\text{syn}}$ scores also decreased when lowering the number of classes. For both datasets, the differences between the extreme cases (20 and 2 classes for the chest x-rays and 10 and 2 classes for the brain CT scans) for the cpD-GAN were statistically significant (p-values $<$ 0.0001). The relative performance increase was even more pronounced for the prog-GAN. The trend of improvement in classifier performance when trained on synthetic data versus the performance when trained on real data shows that GAN models and the generated data quality disproportionately benefited from a smaller label space, thereby confirming the significance of the class conditioning methods. One crucial difference between the two evaluated GAN models is the improved label conditioning mechanism used with the cpD-GAN. The improved label conditioning was partially responsible for the lower overall scores and also explains why the prog-GAN had a more significant relative performance improvement on chest radiographs: Due to its inferior conditioning, the prog-GAN model benefited from a lower class number to a greater extent.

Impact of number of samples per class. When we lowered the number of samples per label combination included in each dataset, the predictive performances obtained when training on real and synthetic data remained similar until approximately $3,000$ samples per class. These results indicate that GAN model performance may be stable when the training data consists of at least $3,000$ samples per class. Between $3,000$ and $1,500$ samples, both the $\overline{AUC}_{\text{real}}$ and $\overline{AUC}_{\text{syn}}$ scores started to decrease substantially. However, we also observed a relative performance improvement, meaning decreasing $\overline{AUC}_{\text{real}} - \overline{AUC}_{\text{syn}}$ scores for the cpD-GAN, when moving towards low numbers of samples. Despite the heightened variance in predictive performance, the difference in the scores between the extrema ($5,950$ and 200 samples per class for the chest x-rays and $5,400$ and 100 samples per class for the brain CT scans) was statistically significant (p-values < 0.001). The observed trend in performance in the low data regime indicates the growing effects of overfitting during GAN training:

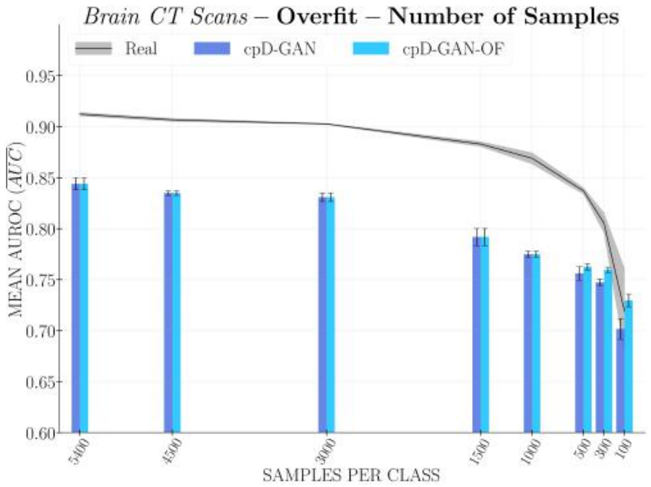

Figure 3: **Overfitting GAN.** The cpD-GAN was trained with early-stopping set to terminate training after the discriminator had seen 7M real images. Training of the cpD-GAN-OF was stopped after the discriminator had seen 14M real images at $500, 300$ and 100 samples to demonstrate the impact of overfitting in low data regimes. Due to more substantial overfitting effects the $\overline{AUC}_{\text{syn}}$ scores of classifiers trained on data generated by the cpD-GAN-OF improved further, leading to a higher predictive performance achieved when training with synthetic data compared to real data at 100 samples (not significant).

Given a low number of samples, the variation within real images becomes too low, and the generative model may resort to memorizing the training set instead of learning the real data distribution.

The overfitting problem is also apparent when analysing how the predictive performance of cpD-GAN was influenced by the available number of samples on the brain scan dataset: Below 500 samples per class, the FID scores calculated during GAN training no longer converged but monotonically decreased. The previously mentioned $\overline{AUC}_{\text{real}} - \overline{AUC}_{\text{syn}}$ score and p-value resulted from models where we performed early-stopping after the discriminator of each GAN had seen 7M real images, which approximately corresponded to the FID convergence point of the other experiments. However, when we continued GAN training to 14M real images, where the FID scores still had not converged, the overfitting effects grew stronger (see Figure 3). At 100 samples per class we observed a negative relative performance (not significant) and the statistical difference between the aforementioned extrema was more significant (p-value < 0.0001).

Impact of resolution. When increasing the resolution from 32×32 pixels to 128×128 pixels, all $\overline{AUC}$ scores improved, as shown in Figure 2e and 2f. However, in terms of relative performance we observed a different behaviour for the two GAN models. For the cpD-GAN, the predictive performance on real data increased disproportionately more, resulting in increased $\overline{AUC}_{\text{real}} - \overline{AUC}_{\text{syn}}$ scores on both datasets. For the prog-GAN, we observed a slight increase in $\overline{AUC}_{\text{real}} - \overline{AUC}_{\text{syn}}$ scores at a resolution of 64 × 64 pixels, with substantially lower scores at 128 × 128 pixels. Generally, GAN model training at higher resolutions is less stable and becomes more difficult due to the emergence of fine-scale details in the images. However, the improvement in the prog-GAN scores indicates that training stability at higher resolutions might not be the limiting factor for the generated images' performance. Instead, we hypothesize that because the prog-GAN model is better fine-tuned on a more considerable number of resolution settings, it can achieve a relative performance improvement when increasing the spatial resolution. Scaling the generation of synthetic medical images to even higher resolutions remains an area of active research ([33, 36]), and is an important direction for future studies.

3.4 Further evaluation

Image quality and privacy. In Figure 4, we show randomly sampled synthetic example images from the best performing cpD-GAN for both datasets at a resolution of 128 × 128 pixels. Below each synthetic image, we show the most similar real image (nearest neighbour) out of the entire training dataset. From a visual assessment, there appears to be no noticeable quality difference between the real and synthetic images. By comparing synthetics and nearest matching neighbours we demonstrate that the cpD-GAN model is not simply memorizing training data, and is therefore likely to preserve private, potentially sensitive information. For more example images from the cpD-GAN and the respective nearest neighbours see Figure S2 in the supplementary materials. From a visual inspection, the quality of the images generated by the prog-GAN appear to be only marginally worse than those generated by the cpD-GAN (Figure S4).

Feature importance. To gain more interpretability, we analysed the feature importance at a resolution of 128 × 128 pixels of real test images, estimated with the method of Schwab and Karlen [39]. Instead of randomly sampling real test images, we chose the real nearest neighbours from Figure 4. Figure 5 shows the nearest neighbours, the corresponding attribution maps of the predictive model trained on reals and the attribution maps for those trained on the synthetic images generated by the cpD-GAN and prog-GAN, respectively (for more examples see Figure S3). The observed results support the hypothesis that the predictive models trained on synthetic data from the cpD-GAN assign importance to similar image features as those trained on real data. In line with the observed larger $\overline{AUC}_{\text{real}} - \overline{AUC}_{\text{syn}}$ scores, the attributions assigned by the classifier trained on synthetic data generated by the prog-GAN appeared to be visually more dissimilar from those assigned by classifiers trained on real data. We note that none of the feature importance maps were identical, which we expected given that the observed difference in predictive performance between classifiers trained on real and synthetic data was greater than zero at this resolution.

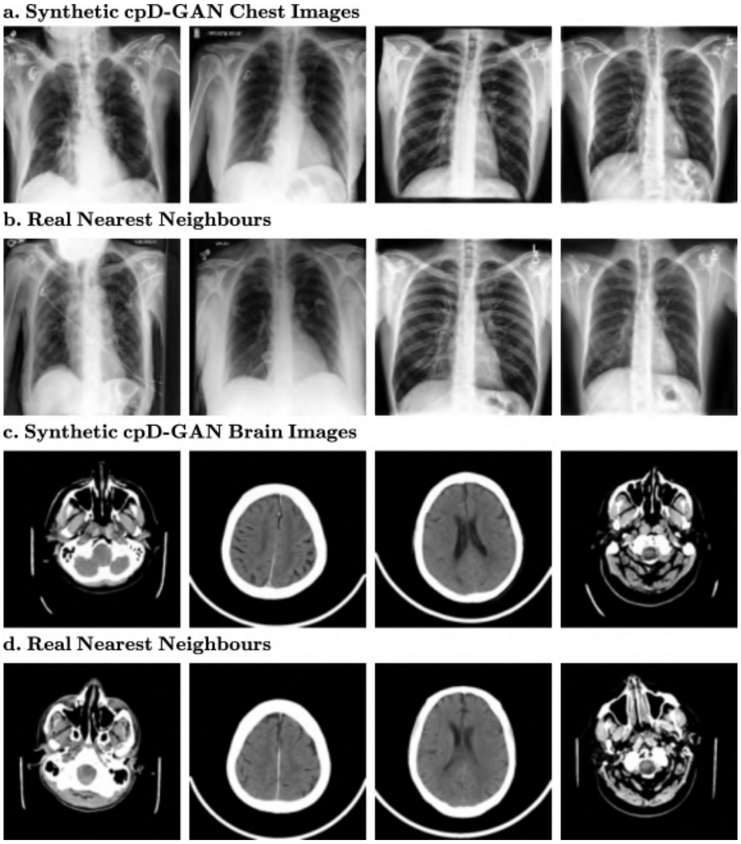
a. Synthetic cpD-GAN Chest Images
b. Real Nearest Neighbours
c. Synthetic cpD-GAN Brain Images
d. Real Nearest Neighbours

Figure 4: **Randomly sampled synthetic images generated by the cpD-GAN and real nearest neighbour images from the training fold at a resolution of** 128×128 **pixels.** **a)** Synthetic chest radiographs. **b)** Nearest matching real images found in the chest radiograph training set. **c)** Synthetic brain computed tomography (CT) scans. **d)** Nearest matching real images found in the brain CT training set.

Reader study. We additionally conducted a reader study in which we asked trained radiologists to label a mixed set of 100 images for both data modalities as real or synthetic (generated by the cpD-GAN) at a resolution of 128×128 pixels. In terms of results, we found that radiologists were unable to achieve a higher accuracy than a classifier assigning labels at random with an expected accuracy of 50% ($p < 0.05$ for chest radiographs, $p < 0.01$ for brain CT scans). The presented results indicate that trained clinicians cannot discriminate between real and synthetic images in the aforementioned setting, which further substantiates that both the general quality and label information in the synthetic images are realistic. Please refer to Section S.8 for details on the set-up and to Section S.10 for a presentation of the detailed results of the conducted reader study.

4 Discussion

In this study, we benchmarked the generation of synthetic medical image data to closely mimic the distribution-level statistical properties of a real source dataset. To do so, we evaluated two state-of-the-art GAN models, prog-GAN and cpD-GAN, on two real-world medical image corpora consisting of chest radiographs and brain CTs, respectively. We compared the difference in performance on real test data between a predictive model trained only on real or only synthetic images. As part of the conducted benchmark evaluation, we analysed the effects of changes in the number of label combinations, samples per class, and resolution. The presented results offer valuable guidelines for synthesising medical imaging datasets in practice. In addition, we analysed the difference in causal contributions of predictive models when trained on either the real or synthetic dataset and investigated the privacy-preservation in our generated medical images by comparing them to the most closely matching real training images. We found that synthetic medical images generated by the cpD-GAN enabled training of classifiers that closely matched the performance of classifiers trained on real data. Finally, we conducted a large-scale reader study in which we found that trained radiologists could not discriminate better than random between real and synthetic images, generated by the cpD-GAN, for both datasets at a resolution of 128 × 128 pixels.

Generalisation ability. We determined that both GAN models are stable across all benchmark dimensions, meaning that we did not observe anomalous $\overline{AUC}_{\text{real}} - \overline{AUC}_{\text{syn}}$ scores. While some GAN models that we trained were not robust across the experiments (see Section S.2.3), the prog-GAN and cpD-GAN did not collapse at any setting or choice of random initialisation. There were no hyper-parameter changes for the varying experiments or across the two datasets. Only when increasing the spatial resolution, we added the necessary convolutional blocks to both models. The observed results indicate that the presented GAN pipeline is robust to changes in the dataset-, and data modality, and that it may generate high quality synthetic medical images across various conditions with the desired statistical similarity compared to the training cohort. For synthetic data generation to work reliably in practice, the convergence of the generative models and the quality of the generated images must be robust across different cohorts where the number of available samples or classes might deviate. While we focused on 2 datasets, our generative methods and evaluation protocols can be easily extended to different settings and are not limited to the chest radiographs and brain CT scans. We believe that our findings show that sharing synthetic medical imaging datasets may be an attractive and privacy-preserving alternative to sharing real patient-level data, thereby providing a technical solution to the pervasive issue of data sharing in medicine [1].

Practical guidelines. The predictive performance obtained when training on synthetic data improved when reducing the number of classes present in

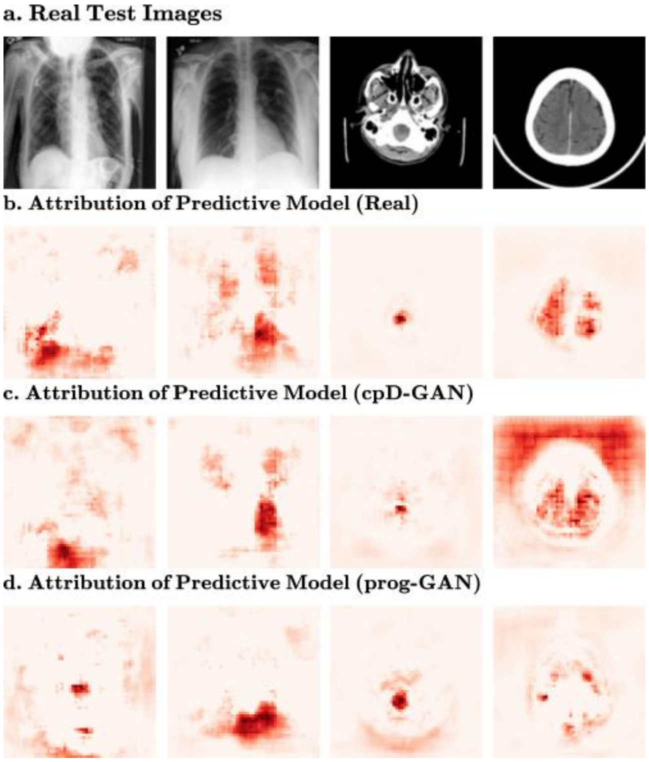
a. Real Test Images
b. Attribution of Predictive Model (Real)
c. Attribution of Predictive Model (cpD-GAN)
d. Attribution of Predictive Model (prog-GAN)

Figure 5: **Feature importance of predictive models.** Deeper red colour indicates regions that have a larger causal contribution to the label prediction. **a)** Real test images (nearest neighbours from Figure 4) at 128 × 128 resolution. From left to right: (1) Chest x-ray with support device, lung opacity, pneumonia and atelectasis. (2) Chest x-ray with cardiomegaly and edema. (3) Brain scan with subarachnoid hemorrhage. (4) Brain scan with subdural hemorrhage. **b)** Feature importance of predictive model trained on real data. **c)** Feature importance of predictive model trained on synthetic data generated by the cpD-GAN. **d)** Feature importance of predictive model trained on synthetic data generated by the prog-GAN.

the dataset. The impact of a reduced label spaces suggests that researchers should, in practice, choose datasets for GAN model development that have a manageable number of unique label combinations. Even though rare findings may be particularly interesting from a clinical perspective, they should be excluded from training when maximum performance is required, since it is currently impossible to give any guarantees for consistent quality in low sample numbers containing rare findings. Moreover, the samples per class benchmark indicates that the GAN models might overfit on rare classes by memorizing training images, resulting in potentially problematic privacy breaches. The most critical performance improvement between the prog-GAN and cpD-GAN resulted from a revised label conditioning mechanism, rooted in a probabilistic framework [40]. Therefore, the impact of the class conditioning mechanism on the predictive performance of derived classifiers suggests that research on the conditioning mechanism of GAN models may lead to further improvements in image quality. In the chest radiograph benchmark, we found that the total number of samples in the training dataset can be lowered significantly (to approximately $9,000$), at a number of samples per class of around $3,000$ without any relative performance drop. However, if low-frequency classes are included and the total number of samples is reduced too much, GAN overfitting is likely to occur, and privacy may be impacted as a result.

In terms of predictive performance in relation to different image resolutions, we found that $\overline{AUC}_{\text{syn}}$ scores for both models improved when moving to a higher resolution. However, we also observed different behaviours in the different GAN models in terms of relative performance compared to real data when adjusting the image resolution. For the prog-GAN, relative predictive performance increased at a resolution of 128×128 pixels compared to lower levels, which indicates that, once a GAN model has been fine-tuned, it can benefit from the emergence of details at a higher spatial resolution. We note that the prog-GAN hyper-parameter settings were taken from the official implementation [23], which has been well adjusted to a number of datasets. While our cpD-GAN model outperformed the prog-GAN at all evaluated resolutions, its own $\overline{AUC}_{\text{real}} - \overline{AUC}_{\text{syn}}$ scores increased when moving up from 32×32 pixels. As mentioned earlier, the increasing difference between real and synthetic data for higher resolutions is not unexpected because stable training becomes more difficult when the discriminator has access to a richer set of features to distinguish real and synthetic data.

In the presented reader study, we found that the accuracy distribution derived from the real and synthetic labels set by radiologists was not better than that of a random classifier with a mean accuracy ($\overline{acc}$) of 50%, to a statistically significant extent. The fact that trained clinicians were unable to discriminate between the real and synthetic medical imaging datasets indicates that the generated images had a realistic visual appearance and label information was included in a qualitatively reasonable manner at resolutions of 128×128 pixels. The results of the presented reader study further support the findings presented in the conducted experimental benchmark evaluation and show that the cohort level information of medical imaging data can be shared without relying on patient-level data.

We have shown that, under the right conditions, sharing synthetic medical imaging datasets may be a viable alternative to real data sharing. However, the presented results also show that there is a measurable gap in quality and predictive performance between synthetic and real medical imaging data. Across all benchmark settings, we observed that only in the extreme overfitting case $\overline{AUC}_{\text{real}} - \overline{AUC}_{\text{syn}} \leq 0$, meaning that in all other experiments, there was a reduced performance when training on the generated images. While this difference was relatively small for the cpD-GAN across the chest radiographs, it was more pronounced on the brain CT scans. From our causal contribution investigation in Figure 5, we found that while the real and cpD-GAN predictive models attributed similar regions with high feature importance, they were not identical. In the ideal case, both assigned feature importance and predictive performance would be identical when replacing real data with synthetic data. Even so, our benchmark results demonstrate that the goal of learning the real data distribution for medical images is realistic and feasible.

Limitations. In this study, we analysed synthetic medical images at an overall low resolution compared to clinical practice. A comparison on low resolution images is more acceptable for evaluating the predictive models as most deep learning systems downsample medical images to reduce the computational requirements. In clinical practice, however, radiographs are analysed at a much higher resolution, and particularly fine-scaled details are essential for the accurate diagnosis of radiology findings. The radiologist evaluation is, therefore, likely to lead to different results at higher resolutions. GAN models such as [26, 27] and other generative approaches such as [41] can generate realistic images at a much higher resolution level, and these methods may likely extend to the generation of synthetic medical images with higher resolution in the future.

While the lower resolution is more acceptable for the brain CT dataset, there is also a significant difference compared to medical practice, where computed tomography consists of 3D scans at different intensity windows. Here, we were limited by the RSNA Intracranial Hemorrhage dataset, which consists of presliced scans with only the soft-tissue window. To analyse a variety of different benchmark settings, we required a certain number of samples that rarely exist in open-source medical imaging datasets. Moreover, reliable synthetic medical data generation is currently limited to 2D settings as it becomes substantially more complex, both in terms of required computational resources and algorithmic challenges, to model 3D structures.

Finally, the presented study does not provide any mathematical guarantees for the privacy of the synthetic data. We found settings in which privacy would likely be breached in practice, which can be an important guideline, but a more formal analysis in terms of differential privacy may in the future further elucidate the degree to which generative modeling preserves individual patient-level information [42]. In Figure 4, we demonstrate that there were considerable differences between the generated images and the most closely matching nearest neighbour images from the training data, which may indicate that the GAN models learn the actual data distribution and do not merely memorize the training set. However, a retrospective analysis may not always be feasible, and more formal privacy guarantees regarding the model and training may be needed in some real-world use cases. Through the use of stochastic gradient descent, all of our GAN models have some level of intrinsic privacy [43], but it remains an area of active research to quantify how strong these privacy guarantees are. While there remain open questions for further research, our results indicate that synthetic data sharing may in the future become an attractive and privacy-preserving alternative to sharing real patient-level data in the right settings.

5 Data availability

Both datasets used in our study are publicly available and free to download for any registered user. The CheXpert chest radiograph dataset [44] can be accessed at https://stanfordmlgroup.github.io/competitions/chexpert/ and the RSNA Intracranial Hemorrhage dataset [45] is available at https://www.kaggle.com/c/rsna-intracranial-hemorrhage-detection.

6 Code availability

To reproduce our benchmark results, please see our code repository under https://github.com/AugustDS/synthetic-medical-benchmark (MIT license).

References

[1] Bernard Lo. Sharing Clinical Trial Data: Maximizing Benefits, Minimizing Risk. *JAMA*, 313(8):793–794, 02 2015.

[2] Serena Sanna et al. Causal relationships among the gut microbiome, short-chain fatty acids and metabolic diseases. *Nature Genetics*, 51:1, 2019.

[3] Hui Li et al. Quantitative MRI radiomics in the prediction of molecular classifications of breast cancer subtypes in the TCGA/TCIA data set. *npj Breast Cancer*, 2:16012, 2016.

[4] Roger Sun et al. A radiomics approach to assess tumour-infiltrating CD8 cells and response to anti-PD-1 or anti-PD-l1 immunotherapy: an imaging biomarker, retrospective multicohort study. *The Lancet Oncology*, 19(9): 1180 – 1191, 2018.

[5] Karla Miller et al. Multimodal population brain imaging in the UK Biobank prospective epidemiological study. *Nature neuroscience*, 19, 2016.

[6] Jeffrey De Fauw et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature Medicine*, 24(9):1342–1350, 2018.

[7] Miguel Monteiro et al. Multiclass semantic segmentation and quantification of traumatic brain injury lesions on head CT using deep learning: an algorithm development and multicentre validation study. *The Lancet Digital Health*, 2(6):e314–e322, 2020.

[8] Yuan Liu et al. A deep learning system for differential diagnosis of skin diseases. *Nature Medicine*, 2020.

[9] Pierre Courtiol et al. Deep learning-based classification of mesothelioma improves prediction of patient outcome. *Nature Medicine*, 25(10):1519–1525, 2019.

[10] Koji Matsuo et al. Survival outcome prediction in cervical cancer: Cox models vs deep-learning model. *The American Journal of Obstetrics & Gynecology*, 220(4):381.e1–381.e14, 2019.

[11] Hongming Chen et al. The rise of deep learning in drug discovery. *Drug Discovery Today*, 23(6):1241–1250, 2018.

[12] Bogdan Zagribelnyy et al. Deep learning enables rapid identification of potent DDR1 kinase inhibitors. *Nature Biotechnology*, 37, 09 2019.

[13] Yann Lecun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[14] Andre Esteva et al. A guide to deep learning in healthcare. *Nature Medicine*, 25(1):24–29, 2019.

[15] Sebastian Haas et al. Aspects of privacy for electronic health records. *International Journal of Medical Informatics*, 80(2):e26–e31, 2011.

[16] Clare Bycroft et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726):203–209, 2018.

[17] Kenneth Clark et al. The cancer imaging archive (TCIA): Maintaining and operating a public information repository. *Journal of Digital Imaging*, 26 (6):1045–1057, 2013.

[18] Katarzyna Tomczak, Patrycja Czerwińska, and Maciej Wiznerowicz. The Cancer Genome Atlas (TCGA): An immeasurable source of knowledge. *Wspolczesna Onkologia*, 1A:A68–A77, 2015.

[19] Willem G van Panhuis et al. A systematic review of barriers to data sharing in public health. *Bulletin of the World Health Organization*, 88(6):468–468, 2010.

[20] Mark Phillips. International data-sharing norms: from the OECD to the General Data Protection Regulation (GDPR). *Human Genetics*, 137(8): 575–582, 2018.

[21] Liangyuan Na et al. Feasibility of Reidentifying Individuals in Large National Physical Activity Data Sets From Which Protected Health Information Has Been Removed With Use of Machine Learning. *JAMA network open*, 1(8): e186040, 2018.

[22] Iheanyi Nwankwo, Stefanie Hänold, and Nikolaus Forgó. Legal and ethical issues in integrating and sharing databases for translational medical research within the EU. *IEEE 12th International Conference on BioInformatics and BioEngineering, BIBE 2012*, (November):428–433, 2012.

[23] Tero Karras et al. Progressive Growing of GANs for Improved Quality, Stability, and Variation. In *International Conference on Learning Representations*, 2018.

[24] Ian Goodfellow et al. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems 27*, pages 2672–2680. Curran Associates, Inc., 2014.

[25] Lars Mescheder, Sebastian Nowozin, and Andreas Geiger. The Numerics of GANs. In *Advances in Neural Information Processing Systems 30*, pages 1825–1835. Curran Associates, Inc., 2017.

[26] Tero Karras et al. Analyzing and Improving the Image Quality of StyleGAN. *arXiv preprint preprint arXiv:1912.04958*, 2019.

[27] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *International Conference on Learning Representations*, 2019.

[28] P. Costa et al. End-to-End Adversarial Retinal Image Synthesis. *IEEE Transactions on Medical Imaging*, 37(3):781–791, 2018.

[29] He Zhao, Huiqi Li, Sebastian Maurer-Stroh, and Li Cheng. Synthesizing retinal and neuronal images with generative adversarial nets. *Medical Image Analysis*, 49:14 – 26, 2018.

[30] Saeed Izadi et al. Generative adversarial networks to segment skin lesions. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 881–884. IEEE, 2018.

[31] Alceu Bissoto et al. Skin Lesion Synthesis with Generative Adversarial Networks. In *OR 2.0 Context-Aware Operating Theaters, Computer Assisted Robotic Endoscopy, Clinical Image-Based Procedures, and Skin Image Analysis*, pages 294–302. Springer International Publishing, 2018.

[32] Ibrahim Saad Ali, Mamdouh Farouk Mohamed, and Yousef Bassyouni Mahdy. Data Augmentation for Skin Lesion using Self-Attention based Progressive Generative Adversarial Network, 2019.

[33] Adalberto Claudio Quiros, Roderick Murray-Smith, and Ke Yuan. Pathology GAN: Learning deep representations of cancer tissue. In *Medical Imaging with Deep Learning*, 2020.

[34] Yuanpin Zhou et al. Generating high resolution digital mammogram from digitized film mammogram with conditional generative adversarial network. In *Medical Imaging 2020: Computer-Aided Diagnosis*, volume 11314, pages 508 – 513. International Society for Optics and Photonics, SPIE, 2020.

[35] Maria J.M. Chuquicusma et al. How to fool radiologists with generative adversarial networks? A visual turing test for lung cancer diagnosis. *Proceedings - International Symposium on Biomedical Imaging*, 2018-April: 240–244, 2018.

[36] Tianyu Han et al. Breaking medical data sharing boundaries by employing artificial radiographs. *bioRxiv*, 2019. doi: 10.1101/841619.

[37] Changhee Han et al. *Infinite Brain MR Images: PGGAN-Based Data Augmentation for Tumor Detection*, pages 291–303. Springer Singapore, Singapore, 2020.

[38] Karim Armanious et al. MedGAN: Medical image translation using GANs. *Computerized Medical Imaging and Graphics*, 79:101684, 2020.

[39] Patrick Schwab and Walter Karlen. CXPlain: Causal Explanations for Model Interpretation under Uncertainty. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[40] Takeru Miyato and Masanori Koyama. cGANs with projection discriminator. In *International Conference on Learning Representations*, 2018.

[41] Arash Vahdat and Jan Kautz. NVAE: A deep hierarchical variational autoencoder. *arXiv preprint arXiv:12007.03898*, 2020.

[42] Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9 (3-4):211–407, 2014.

[43] Stephanie L Hyland and Shruti Tople. On the intrinsic privacy of stochastic gradient descent. *arXiv preprint arXiv:1912.02919*, 2019.

[44] Jeremy Irvin et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 590–597, 2019.

[45] Adam E Flanders et al. Construction of a Machine Learning Dataset through Collaboration: The RSNA 2019 Brain CT Hemorrhage Challenge. *Radiology: Artificial Intelligence*, 2(3):e190211, 2020.

[46] Kaiming He et al. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.

[47] Ishaan Gulrajani et al. Improved training of Wasserstein GANs. In *Advances in neural information processing systems*, pages 5767–5777, 2017.

[48] Augustus Odena, Christopher Olah, and Jonathon Shlens. Conditional Image Synthesis with Auxiliary Classifier GANs. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 2642–2651. PMLR, 06–11 Aug 2017.

[49] T. Karras, S. Laine, and T. Aila. A Style-Based Generator Architecture for Generative Adversarial Networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4396–4405, 2019.

[50] Harm de Vries et al. Modulating early visual processing by language. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6597–6607. Curran Associates Inc., 2017.

[51] Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. Which training methods for GANs do actually converge? In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 3481–3490, 10–15 Jul 2018.

[52] Xiaolong Wang et al. Non-local neural networks. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7794–7803, 2018.

[53] Han Zhang et al. Consistency regularization for generative adversarial networks. In *International Conference on Learning Representations*, 2020.

[54] Han Zhang et al. Self-Attention Generative Adversarial Networks. In *International Conference on Machine Learning*, pages 7354–7363, 2019.

[55] Takeru Miyato et al. Spectral normalization for Generative Adversarial Networks. In *International Conference on Learning Representations*, 2018.

[56] Martin Heusel et al. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in neural information processing systems*, pages 6626–6637, 2017.

[57] Gao Huang et al. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[58] Labelbox Inc. Labelbox: The leading training data platform for data labeling. URL `https://labelbox.com`.

S Methods

S.1 Datasets and pre-processing

The CheXpert dataset consists of $224,316$ chest radiographs of $65,240$ patients, collected from radiographic examinations of the chest at the Stanford Hospital, between October 2002 and July 2017 [44]. In the dataset study, an automatic labeling tool was used to identify and classify the certainty of the presence of 14 observations from the radiology report. We turned uncertain labels into positives, to make use of all data, resulting in a binary multi-label dataset, where a large number of label combinations can co-occur.

The RSNA Intracranial Hemorrhage Dataset is composed of computed tomography studies supplied by four research institutions and labeled with the help of The American Society of Neuroradiology [45]. It consists of $752,803$ CT scan slices of the head from $18,938$ unique patients and the corresponding probabilities for the presence of 5 different hemorrhage types and the no finding label. For consistency, we turned any probability $p_{y_i} > 0$ into a positive label $y_i = 1$ and else $y_i = 0$, also resulting in a binary multi-label dataset. Since $644,874$ (85.7%) of CT scans are without any intracranial hemorrhage, we undersampled the no-finding class, resulting in a balanced dataset where at least 50% of images show some form of hemorrhage.

We randomly split the entire patient cohort into training (80%), validation (10%), and test folds (10%) within strata of radiology findings for each dataset. We excluded chest x-rays of classes with fewer then 256 samples, resulting in $117,168$ train images ($44,153$ patients), $15,318$ validation images ($5,519$ patients) and $14,687$ test images ($5,520$ patients). For the hemorrhage dataset we removed label combinations below a frequency of 100, resulting in $173,271$ train images ($15,133$ patients), $22,095$ validation images ($1,892$ patients), and $20,500$ test images ($1,892$ patients).

We developed the resolution benchmark for both datasets on the aforementioned setting. For the class benchmarking, we gradually reduced the number of clinical finding combinations present in the dataset, while keeping the total number of training images constant via over-sampling. When benchmarking the effect of samples per clinical finding, we fixed the number of classes and gradually decreased each class's frequency. Table S1 gives a complete summary of all dataset settings, the entire set of labels, the size of training, validation and test sets, and information on remaining labels and samples per class.

S.2 GAN model development

S.2.1 prog-GAN

We used the prog-GAN model as originally proposed in [23], as it is still regularly used for generating medical images [32, 36]. The input of the generator is a concatenation of the 512 dimensional random normal noise vector z and the label information y. Each resolution block is composed of two 3×3 convolutional layers followed by Leaky-ReLU activation functions and pixel-wise feature vector

Dataset Information	Benchmark	Resolution	m_{labels}	$m_{label\ comb}$	n_r	$n_{val/te}$	$n_{per\ label\ comb}$
Chest Radiographs	Classes	32 × 32	9	20	29000	3800	1450
			8	15	24000	2850	1600
Data Pool:			5	10	20000	1900	2000
n_{tr} $(n_{pat}) = 117168$ (44153)			5	6	13800	1140	2300
n_{vl} $(n_{pat}) = 15318$ (5519)			5	4	15600	760	3900
n_{te} $(n_{pat}) = 14687$ (5520)			4	2	12600	380	6300
	Samples	32 × 32	4	3	17850	2250	5950
All Labels:			4	3	13500	2250	4500
Enlarged Cardiomediastinum,			4	3	9000	2250	3000
Cardiomegaly, Lung Opacity,			4	3	4500	2250	1500
Lung Lesion, Edema,			4	3	3000	2250	1000
Consolidation, Pneumonia,			4	3	1500	2250	500
Atelectasis, Pneumothorax,			4	3	1200	2250	400
Pleural Effusion, Pleural Other,			4	3	600	2250	200
Fracture, Support Device,	Resolution	32 × 32	14	138	117168	8000	256 – 7586
No Finding	(pixels)	64 × 64	14	138	117168	8000	256 – 7586
		128 × 128	14	138	117168	8000	256 – 7586
Brain Hemorrhage CTs	Classes	32 × 32	5	10	25000	3000	2500
			5	8	24960	2400	3120
Data Pool:			5	6	25020	1800	4170
n_{tr} $(n_{pat}) = 173271$ (15133)			4	4	25000	1200	6250
n_{vl} $(n_{pat}) = 22095$ (1892)			2	2	25000	600	12500
n_{te} $(n_{pat}) = 20500$ (1892)	Samples	32 × 32	5	6	32400	3000	5400
			5	6	27000	3000	4500
All Labels:			5	6	18000	3000	3000
Epidural, Subarachnoid,			5	6	9000	3000	1500
Subdural, Intraparenchymal,			5	6	6000	3000	1000
Intraventricular, No Finding			5	6	3000	3000	500
			5	6	1800	3000	300
			5	6	600	3000	100
	Resolution	32 × 32	5	20	117168	8000	155 – 85876
	(pixels)	64 × 64	5	20	117168	8000	155 – 85876
		128 × 128	5	20	117168	8000	155 – 85876

Table S1: **All benchmark settings.** *Dataset information* summarizes the total available data for each dataset after preprocessing. m_{labels} refers to the number of labels. $m_{label\ comb}$ refers to the number of unique label combinations. n_{tr}, n_{vl}, n_{te} refers to the total number of training, validation and test samples. $n_{per\ label\ comb}$ refers to the number of training samples per unique label combination.

normalisation. For networks operating at up to 32 × 32 pixels, the generator operates at constant 512 feature channels. At higher resolution, the number of feature channels is halved with the final convolution layers of the 64×64 and 128×128 block. The discriminator consists of the same resolution blocks in the opposite order and without pixel-wise feature vector normalisation. When operating above 32 × 32 spatial resolution, the first convolutional layer in each block doubles the number of feature channels. In the final layer of the discriminator, the mini-batch standard deviation across all channels is added as an additional feature channel to increase variation. Between resolution blocks, nearest neighbour upsampling doubles the generator's resolution, and downsampling by average pooling halves it inside the discriminator. At each operating resolution, 1 × 1 convolutional layers project the number of feature channels to and from the image space, which allows to smoothly interpolate between consecutive levels of detail during progressive growth. All weights in the network are dynamically scaled with a variant of He's initialiser [46] at each optimisation step to stabilise training. The Wasserstein GAN with gradient penalty loss function is used [47].

An additional auxiliary classifier loss term is added to both the generator and discriminator [48] for conditioning. The discriminator is not only trained to classify whether input images are real or fake, but to additionally predict the label. The softmax cross-entropy loss between true and predicted labels for both real and fake images is added to the discriminator loss function, while the same loss but only for fake images is added to the generator loss. We analysed several hyper-parameter settings, mainly different batch sizes, learning rates, number of feature channels and optimiser settings, but we determined that the original parameters proposed in [23] performed best. We began training at a spatial resolution of 8×8 pixels, which we determined to be the lowest resolution at which meaningful information is still visually apparent in downsampled images. Each transition and stabilisation phase at a resolution of 32×32 pixels lasted until the discriminator had seen 1.4M real images, which corresponded to 1.4M fake images as the number of discriminator updates per generator step is $n_{critic} = 1$. At a resolution of 64×64 and 128×128 pixels, we reduced the number of real images per phase to 1M.

S.2.2 cpD-GAN

We developed the cpD-GAN based on the prog-GAN with several important improvements that we highlight below. Please see above or [23] for details on the architecture and methods if not explicitly stated. Inspired by Style-GAN [26, 49], we dropped progressive growth as we observed that it was not necessary for stable training. This allowed us to experiment with new architectures, where output skip connections within the image feature space of the generator and standard residual connections in the discriminator improved the performance the most. We achieved significantly lower $\overline{AUC}_{\text{real}} - \overline{AUC}_{\text{syn}}$ scores when replacing the auxiliary classifier conditioning with a projection based discriminator: In the last discriminator layer, the inner product between the label vector y and the feature vector is computed as the final output, resulting in a conditioning mechanism that respects the role of the conditional information in the underlining probabilistic model [40]. Inspired by conditional batch normalisation [50], we modified the pixel-wise feature vector normalisation after each generator convolution by conditioning it on a label and noise dependent scaling and bias parameter:

$$b^i_{x,y} = \frac{a^i_{x,y}}{\sqrt{1/N \sum_{j=0}^{N-1} (a^j_{x,y})^2 + \epsilon}} \cdot \gamma^i + \beta^i \tag{1}$$

where $a^i_{x,y}$ and $b^i_{x,y}$ are the original and normalised feature of channel i in pixel (x, y) and $\epsilon = 10^{-8}$. The scaling parameter is defined as $\gamma = W_1[z; y] + b_1$ and the bias parameter as $\beta = W_2[z; y] + b_2$, where W_i and b_i are trainable weight matrices and vectors, while $[z; y]$ refers to the vector concatenation of the random normal input noise z and label y. Figure S1 shows the overall model structure and a detailed description of a generator resolution block.

We evaluated various loss functions, such as the logistic GAN loss with and without R1 or R2 regularisation, the hinge loss with and without gradient penalty, or the non-saturating GAN loss [51], but the Wasserstein loss with gradient penalty worked best. Replacing a specific convolutional layer in both the generator and discriminator by a normal, sparse, or non-local self-attention layer [52] did not improve performance. Neither consistency regularisation [53], nor matching the gradients of an auxiliary classifier by minimisation of the cosine-distance when predicting the labels of fake and real images resulted in better scores. We analysed many hyper-parameters, among others the number of feature channels, batch sizes, and learning rates. The performance peaked for 512 feature channels, a batch size of 256 and learning rates of 0.005 with one discriminator update per generator update ($n_{critic} = 1$).

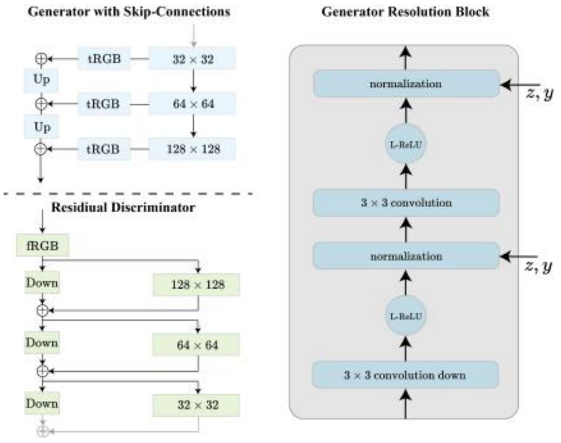

Figure S1: **Network architecture and generator block. Left:** In the generator, output skip connections in the image feature space are included after each resolution block, while the discriminator blocks have residual connections. *Up* and *Down* refer to nearest neighbour upsampling and downsampling by average pooling while *tRGB* and *fRGB* refer to the 1 × 1 convolution mappings to and from the image space. **Right:** The first convolution in each generator block doubles the spatial resolution via nearest neighbour upsampling and reduces the number of feature channels (if needed). Each pixel-wise feature vector normalisation layer is conditioned on the label information $\mathbf{y}$ and random normal noise vector $\mathbf{z}$. The Leaky-ReLU non-linearity is used as an activation function.

S.2.3 BIGGAN (discarded)

The third model that we analysed in detail is largely based on the normal BIGGAN implementation [27], with some elements of the self-attention GAN [54]. However, the implementation did not generalise across different benchmark settings which is why we excluded it from the results and discussion section. In the generator, each block has residual connections and is made up of two 3 × 3 convolutional layers (the first halves the number of feature channels), with

ReLU non-linearities followed by conditional batch normalisation and nearest neighbour upsampling layers in between. The 120-dimensional random normal noise vector z is split, concatenated with the label vector y, and fed as input to the initial fully connected generator layer and every residual block. The output layer of the generator consists of batch normalisation, a 3×3 convolutional layer and *tanh* non-linearity. In the conditional projection based discriminator, residual blocks are built in the opposite way, without batch normalisation and with average pooling for downsampling. In both the generator and discriminator a self-attention layer replaces the residual block at the second highest spatial resolution. To stabilise training spectral normalisation, along with orthogonal weight regularisation is applied to all weights [55]. Prior to the label projection embedding in the discriminator, global sum pooling is performed. We investigated a large amount of different loss functions, feature channel numbers, batch sizes, learning rates and discriminator updates per generator update. Even after performing extensive experiments we could not find a model that generalised across the different dimensions of the benchmark settings, often resulting in training collapse, high FID scores or large $\overline{AUC}_{\text{real}} - \overline{AUC}_{\text{syn}}$ scores. In our only stable training setting for the resolution at 32×32 pixels, we used a combination of the hinge loss for the discriminator and Wasserstein loss for the generator, a batch size of 256 with a maximum of 256 feature channels, learning rates for the generator and discriminator of 0.01 and 0.04 and two discriminator updates per generator update $n_{critic} = 2$.

S.3 GAN training

We used the Adam optimiser for all GAN models and the hyper-parameters as proposed in [23, 27], except for the learning rates that we fine-tuned as mentioned in S.2.2. We stopped training in all settings when the Fréchet Inception Distance (FID) between $10,000$ real and synthetic images converged. The FID score is a commonly used metric to compare the visual quality between two sets of images and allows for an unbiased GAN training evaluation [56]. We ran all models for a minimum number of steps, until the discriminator had seen as many real images as the prog-GAN discriminator after the final stabilisation phase. At a resolution of 32×32 pixels, each progressive phase lasted until the discriminator had seen 1.4M real images, resulting in a minimum number of 7M real images for the other models. For 64×64 and 128×128 pixels, we lowered the number of images per phase to 1M, resulting in a minimum number of images of 7M and 9M, respectively. At this point, we computed the FID score after every 400T real images, and if there was no improvement for two consecutive evaluations, we stopped training. We stopped all repetitions for each experiment at the same step as the first model to get comparable results.

S.4 Predictive model development and training

In all settings, we used a pre-trained densenet-121 convolutional neural network as the predictive model [57]. We added a randomly initialised fully connected layer with sigmoid activation to the pre-trained model for classification with the binary cross-entropy loss. We resized the input images to match the densenet-121 spatial input resolution of 224×224 pixels. To make training as similar as possible across different benchmark settings, we used a maximum number of 5000 images per epoch with a batch size of 48. In settings where the total number of samples is below 5000, the number of images per epoch is accordingly lower. After each epoch, we computed the area under the receiver operating characteristic curve (AUROC) for each label in all validation data samples. We reduce the initial learning rate of 0.0001 by a factor of 10 if the mean validation AUROC ($\overline{AUC}_{val}$) across all labels does not improve after two consecutive epochs (patience of 2). If the $\overline{AUC}_{val}$ does not improve for a patience of 3 epochs, we stopped training. To compute delta scores, we tested all models on the held-out, real data test set.

S.5 Statistical tests for benchmark

We repeated every experiment of our benchmark with at least four different random initialisation of the entire training and evaluation pipeline, allowing us to compute the standard deviation for each setting across repetitions. This is necessary as different parameter initialisation resulting from different random seeds can substantially impact the training of deep learning systems. For the number of classes benchmark, we repeated the cpD-GAN training and subsequent synthetic classification as well as the real data classification for 10 different random initialisation for both datasets at the extrema: For 20 and 2 classes for the chest x-rays and 10 and 2 classes for brain CT scans. Subsequently we performed the one-sided, parametric-free, Mann–Whitney U test on the $\overline{AUC}_{\text{real}} - \overline{AUC}_{\text{syn}}$ scores between the extrema to determine whether there is a statistically significant difference. We followed the same approach for the samples per class benchmark with 20 repetitions at different random initialisation: For $5,950$ and 200 samples per class for the chest x-rays and $5,400$ and 100 samples per class for brain CT scans. Here we repeated both the early-stopping cpD-GAN experiments, as well as the overfitting version. We once again performed the one-sided, parametric-free, Mann–Whitney U test on the $\overline{AUC}_{\text{real}} - \overline{AUC}_{\text{syn}}$ scores between the extreme settings to determine the statistical significance.

S.6 Nearest neighbours

To analyse differences between our generated medical images when compared to the training data, we computed the nearest neighbours for a set of randomly sampled synthetics. For both datasets, we used the synthetic images generated by the cpD or prog-GAN model at a resolution of 128×128 pixels with the lowest $\overline{AUC}_{\text{real}} - \overline{AUC}_{\text{syn}}$ scores. We used the predictive model trained on real data at the same resolution level to find the final dense layer representation for each synthetic image; a $1,024$ dimensional vector. We compute the same representation for all real training images and determine the pair of synthetic and real images for which the cosine distance between the final densenet representations is minimal. Using a measure of similarity in the predictive model's feature space results in a more reliable determination of nearest neighbours that exploits invariances to shifts and rotations within the image space of the chest radiographs or brain scans.

S.7 Feature importance

We computed the causal contribution of image neighbourhoods towards the label prediction with the method of Schwab and Karlen [39]. More precisely, we successively zero masked regions of 2×2 pixels in the input image and computed the new, increased predictive model loss. If the masking of a particular neighbourhood resulted in a significant loss increase, the region had accordingly higher importance. All regions were masked for each input image of 224×224 pixels after $12,544$ repetitions. To determine the feature importance we subtracted the original model loss and normalised the attribution map. Similar causal contribution maps indicate a similar quality and structure between real and synthetic images, leading to predictive models that attribute the same regions with high feature importance.

S.8 Reader study

We conducted the reader study by asking trained radiologists to label a set of 100 images for both data modalities as real or synthetic at a resolution of 128×128 pixels with a web-based labeling tool [58]. Each set consisted of 50 randomly sampled real and synthetic images, generated by the best performing cpD-GAN. Participants were told that each individual image was sampled at random to avoid any bias during evaluation, without knowledge about the total number of reals and synthetics. For the chest x-rays, 11 radiologists participated, while 9 radiologists labeled the brain CT sets. From each labeled set we computed the values for true reals (TR), false reals (FR), true synthetics (TS) and false synthetics (FS), to determine the classification accuracy $\text{acc} = \frac{TR+TS}{TR+TS+FR+FS}$. Next, we performed the one-sided, non-parametric Wilcoxon signed-rank test to assess whether the distribution of accuracies is equal or less than the mean accuracy of a fully random classifier with $\overline{acc} = 0.5$ (50%).

S.9 More figures

S.9.1 cpD-GAN

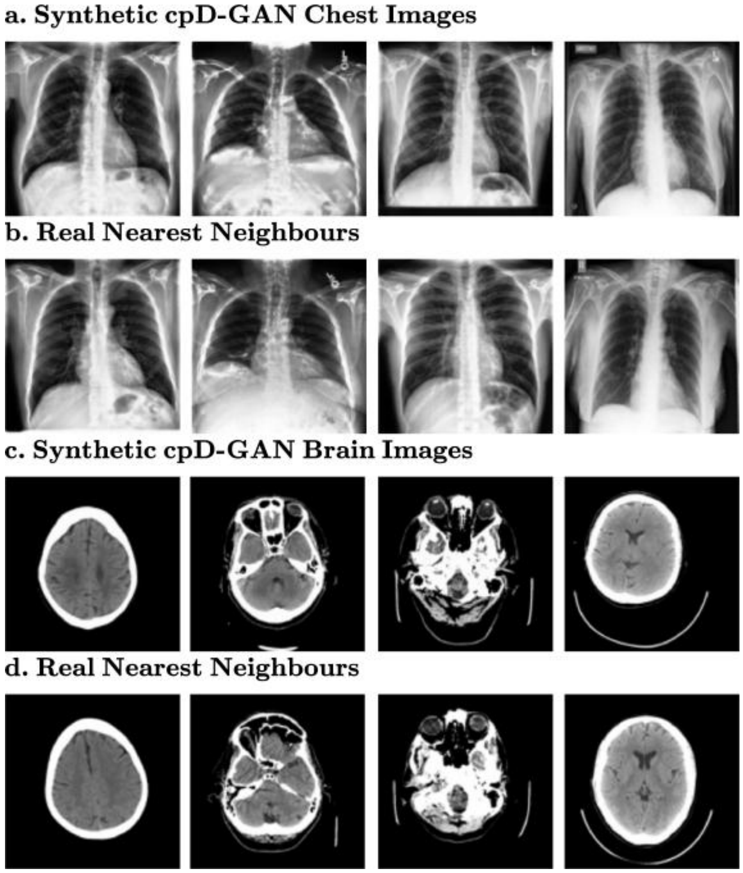
a. Synthetic cpD-GAN Chest Images
b. Real Nearest Neighbours
c. Synthetic cpD-GAN Brain Images
d. Real Nearest Neighbours

Figure S2: **More randomly sampled synthetic images from the cpD-GAN and nearest neighbours from all real training images at a resolution of** 128×128 **pixels. a)** Synthetic chest radiographs. **b)** Nearest matching real images found in the chest radiograph training set. **c)** Synthetic brain computed tomography (CT) scans. **d)** Nearest matching real images found in the brain CT training set.

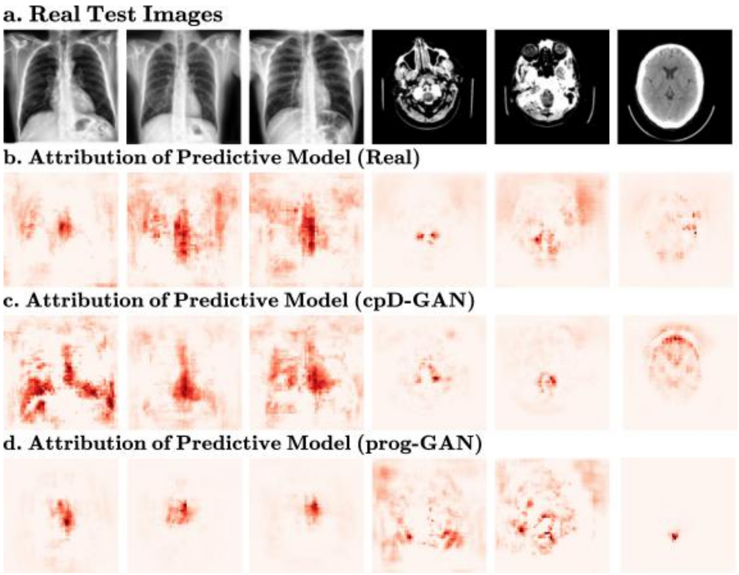
a. Real Test Images
b. Attribution of Predictive Model (Real)
c. Attribution of Predictive Model (cpD-GAN)
d. Attribution of Predictive Model (prog-GAN)

Figure S3: **More feature importance maps of predictive models.** Deeper red colour indicates regions that have a larger causal contribution to the label prediction. **a)** Real test images (nearest neighbours from Figure 4) at 128 × 128 resolution. All displayed images are without any clinical finding. **b)** Feature importance of predictive model trained on real data. **c)** Feature importance of predictive model trained on synthetic data generated by the cpD-GAN. **d)** Feature importance of predictive model trained on synthetic data generated by the prog-GAN.

S.9.2 prog-GAN

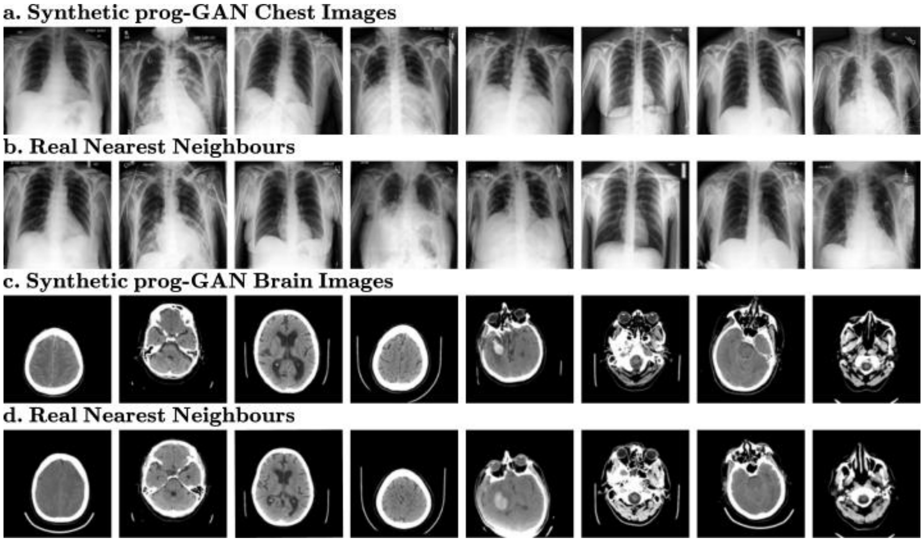
a. Synthetic prog-GAN Chest Images
b. Real Nearest Neighbours
c. Synthetic prog-GAN Brain Images
d. Real Nearest Neighbours

Figure S4: **Randomly sampled synthetic images from the prog-GAN and nearest neighbours from all real training images at resolution of** 128×128 **pixels.** **a)** Synthetic chest radiographs. **b)** Nearest matching real images found in the chest radiograph training set. **c)** Synthetic brain computed tomography (CT) scans. **d)** Nearest matching real images found in the brain CT training set.

S.10 Details on reader study

The detailed confusion matrices and accuracies from the reader study are shown in Table S2 and S3.

		Radiologist	
		Real	Synthetic
Actual	Real	$TR = 25.6$ (± 7.1)	$FS = 24.4$ (± 7.1)
	Synthetic	$FR = 31.0$ (± 8.2)	$TS = 19.0$ (± 8.2)

Accuracies of Radiologist Labels

0.45	0.52	0.45	0.52	0.50	0.25	0.49	0.39	0.40	0.55	0.39

Table S2: **Chest radiographs reader study. Top:** Means and standard deviation from 11 trained radiologists for real and synthetic images at 128×128 resolution: $TR =$ True Reals, $FR =$ False Reals, $TS =$ True Synthetics, $FS =$ False Synthetics. **Bottom:** Computed accuracies from radiologist labels.

		Radiologist	
		Real	Synthetic
Actual	Real	$TR = 25.2$ (±3.5)	$FS = 24.8$ (±3.5)
	Synthetic	$FR = 30.3$ (±5.8)	$TS = 19.7$ (±5.8)

Accuracies of Radiologist Labels

0.51	0.45	0.44	0.44	0.46	0.50	0.48	0.36	0.40

Table S3: **Brain CT scans reader study. Top:** Means and standard deviation from 9 trained radiologists for real and synthetic images at 128×128 resolution: $TR =$ True Reals, $FR =$ False Reals, $TS =$ True Synthetics, $FS =$ False Synthetics. **Bottom:** Computed accuracies from radiologist labels.