Overcoming barriers to data sharing with medical image generation: a comprehensive evaluation

npj | Digital Medicine

www.nature.com/npjdigitalmed

ARTICLE OPEN

Overcoming barriers to data sharing with medical image generation: a comprehensive evaluation

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

Privacy concerns around sharing personally identifiable information are a major barrier to data sharing in medical research. 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 utilise generative adversarial networks (GANs) to create 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 radiology findings and brain computed tomography (CT) scans with six types of intracranial haemorrhages. 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 classes. Our benchmark also indicates that at low numbers of samples per class, label overfitting effects start to dominate GAN training. We conducted a reader study in which trained radiologists discriminate between synthetic and real images. In accordance with our benchmark results, the classification accuracy of radiologists improves with an increasing resolution. Our study offers valuable guidelines and outlines practical conditions under which insights derived from synthetic images are similar to those that would have been derived from real data. Our results indicate that synthetic data sharing may be an attractive alternative to sharing real patient-level data in the right setting.

npj Digital Medicine (2021)4:141 ; https://doi.org/10.1038/s41746-021-00507-3

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[6–8], 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[16–18]. 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 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[30–32], haematoxylin and eosin (H&E) stained breast cancer tissue in digital pathology[33], X-ray mammographs[34], chest radiographs[35,36] and brain tumour magnetic resonance imaging (MRI)[37]. In[38], the authors illustrate the benefits of synthetic images as an additional form of data augmentation for tumour segmentation. They also analyse synthetic data sharing capabilities, but find that without fine-tuning a segmentation model on real data after it was trained on synthetic data, the performance gap is significant. The above-mentioned works develop and demonstrate GAN capabilities in

[1]ETH Zurich, Zurich, Switzerland. [2]Max Planck Institute for Intelligent Systems, Tübingen, Germany. [3]Department of Medical Oncology and Pneumology, University Hospital of Tübingen, Tübingen, Germany. [4]Department of Pneumology, Kantonsspital Winterthur, Winterthur, Switzerland. [5]Department of Radiology, University Hospital of Tübingen, Tübingen, Germany. [6]CIFAR Azrieli Global Scholar, Toronto, Canada. [7]GlaxoSmithKline, Artificial Intelligence & Machine Learning, Zug, Switzerland. ✉email: augustschdmnt@gmail.com

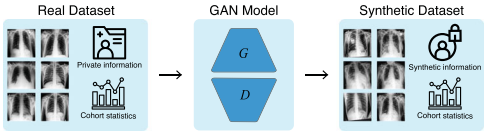

Fig. 1 Synthetic medical imaging dataset generation to overcome data sharing barriers. We train our GAN models with real medical imaging data, to generate the corresponding synthetic images. The synthetic dataset ideally no longer contains private information about individual patients while in aggregate,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

specific domains, with dataset-dependent adaptations, without providing a comprehensive evaluation of how changes within and across data modalities impact different GAN model performances for synthetic data sharing. Other works such as[39] and[40] are less related to data sharing restrictions and instead focus on utilising GANs for image-to-image translations within the medical domain. This idea has also been extended to semi-supervised settings where lower complexity images are synthesised first, before translating towards the higher complexity space[41].

Inspired by these domain-specific advancements we aim to establish a benchmark on synthetic medical imaging data generation capabilities. To the best of our knowledge, there is currently no work focused on providing a comprehensive benchmark analysis for the generation of synthetic medical images across different GAN architectures and data modalities. Our study offers guidelines for the use of GAN models to fully synthesise datasets as a potentially viable approach to privacy-preserving data sharing, as illustrated in Fig. 1. We 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.
- We additionally analyse privacy considerations, assess the feature importance of predictive models trained on the synthetic datasets, analyse visual artefacts at higher resolutions and finally conduct a large-scale reader study in which trained radiologists discriminate between real and synthetic medical images.

RESULTS

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, oedema, 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 haemorrhage 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, before filtering the available data for each benchmark setting. We developed all GAN models on the training datasets and stopped GAN training when the quality between real and synthetic images converged, as measured with the Fréchet inception distance (FID) score[42]. Next, we generated the synthetic datasets for the train, validation, and test folds by conditioning on the labels present in the respective data folds. This means that after GAN training and inference we have a real and synthetic dataset for each benchmark setting with equivalent sizes and label combinations in all folds. In theory, a trained GAN can be used to generate unlimited amounts of data, but we want the real and synthetic folds to be equivalent for a fair comparison.

Each classifier is trained on either the real or the synthetic training data fold, meaning that synthetic images are pre-computed and not generated on a batch-wise basis. 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 ($\overline{\text{AUC}}$) as the evaluation metric. For each classifier, we stopped training when the validation $\overline{\text{AUC}}$ converged. After the real predictive model is trained on the real dataset and the synthetic predictive model is trained on the synthetic dataset, we evaluated both on the separate, real data test fold to compute the difference in performance $\overline{\text{AUC}}_{\text{real}} - \overline{\text{AUC}}_{\text{syn}}$.

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 $\overline{\text{AUC}}_{\text{real}} - \overline{\text{AUC}}_{\text{syn}}$ scores differs at different benchmark settings. Additionally, we compared the predictive models' feature importance when trained on either the real or the 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.

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 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.

Up to a spatial resolution of 128×128 pixels, the prog-GAN achieved an average $\overline{\text{AUC}}_{\text{real}} - \overline{\text{AUC}}_{\text{syn}}$ 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 achieved 0.0206 (±0.0100) on the chest X-ray settings and 0.0650 (±0.0198) on the brain haemorrhage dataset experiments.

We evaluated the model performance across three benchmark dimensions, detailed in Table 1. 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. Due to the substantial computational demand at high spatial resolution, we only evaluate the cpD-GAN at 256×256 pixels for brain CT scans and 256×256 and 512×512 pixels for chest X-rays. We only performed changes across a single benchmark dimension at a time to ensure no confounding factors could 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

Table 1. All benchmark settings.

Type	Benchmark	Resolution	0/1 labels	Classes	Train set	Test/val set	Per class
Chest	Classes	32 × 32	9	20	29,000	3800	1450
		32 × 32	8	15	24,000	2850	1600
		32 × 32	5	10	20,000	1900	2000
		32 × 32	5	6	13,800	1140	2300
		32 × 32	5	4	15,600	760	3900
		32 × 32	4	2	12,600	380	6300
	Samples	32 × 32	4	3	17,850	2250	5950
		32 × 32	4	3	13,500	2250	4500
		32 × 32	4	3	9000	2250	3000
		32 × 32	4	3	4500	2250	1500
		32 × 32	4	3	3000	2250	1000
		32 × 32	4	3	1500	2250	500
		32 × 32	4	3	1200	2250	400
		32 × 32	4	3	600	2250	200
	Resolution	32 × 32	14	138	117,168	4000	256–7586
		64 × 64	14	138	117,168	4000	256–7586
		128 × 128	14	138	117,168	4000	256–7586
		256 × 256	14	138	117,168	4000	256–7586
		512 × 512	14	138	117,168	4000	256–7586
Brain	Classes	32 × 32	5	10	25,000	3000	2500
		32 × 32	5	8	24,960	2400	3120
		32 × 32	5	6	25,020	1800	4170
		32 × 32	4	4	25,000	1200	6250
		32 × 32	2	2	25,000	600	12,500
	Samples	32 × 32	5	6	32,400	3000	5400
		32 × 32	5	6	27,000	3000	4500
		32 × 32	5	6	18,000	3000	3000
		32 × 32	5	6	9000	3000	1500
		32 × 32	5	6	6000	3000	1000
		32 × 32	5	6	3000	3000	500
		32 × 32	5	6	1800	3000	300
		32 × 32	5	6	600	3000	100
	Resolution	32 × 32	6	20	117,168	4000	155–85,876
		64 × 64	6	20	117,168	4000	155–85,876
		128 × 128	6	20	117,168	4000	155–85,876
		256 × 256	6	20	117,168	4000	155–85,876
		512 × 512	6	20	117,168	4000	155–85,876

Each row defines the composition of a specific benchmark setting. After GAN training, the synthetic datasets are generated by conditioning on the real label sets, resulting in equivalent data folds. Our chest radigraph data pool consists of 117,168 (44,153) training, 15,418 (5519) validation and 14,687 (5520) test samples (patients), respectively. Our brain computed tomography scan data pool consists of 173,271 (15,133) training, 22,095 (1892) validation and 20,500 (1892) test samples (patients), respectively. The 14 binary chest X-ray labels are enlarged cardiomediastinum, cardiomegaly, lung opacity, lung lesion, oedema, consolidation, pneumonia, atelectasis, pneumothorax, pleural effusion, pleural other, fracture and support device, and no finding. The six binary brain CT scan labels are epidural, subarachnoid, subdural, intraparenchymal and intraventricular haemorrhage, and no finding. 0/1 Labels: number of binary labels. Classes: number of classes. Note: the number of classes refers to the number of unique binary label combinations. If different binary labels co-occur, we can have fewer classes than 0/1 labels. Train set: number of samples in training set. Test/val set: number of samples in each the test and validation set. Per class: number of training samples per class.

$\overline{\text{AUC}}$ scores. However, as can be seen in Fig. 2a, b, the differences in $\overline{\text{AUC}}_{\text{real}} - \overline{\text{AUC}}_{\text{syn}}$ 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

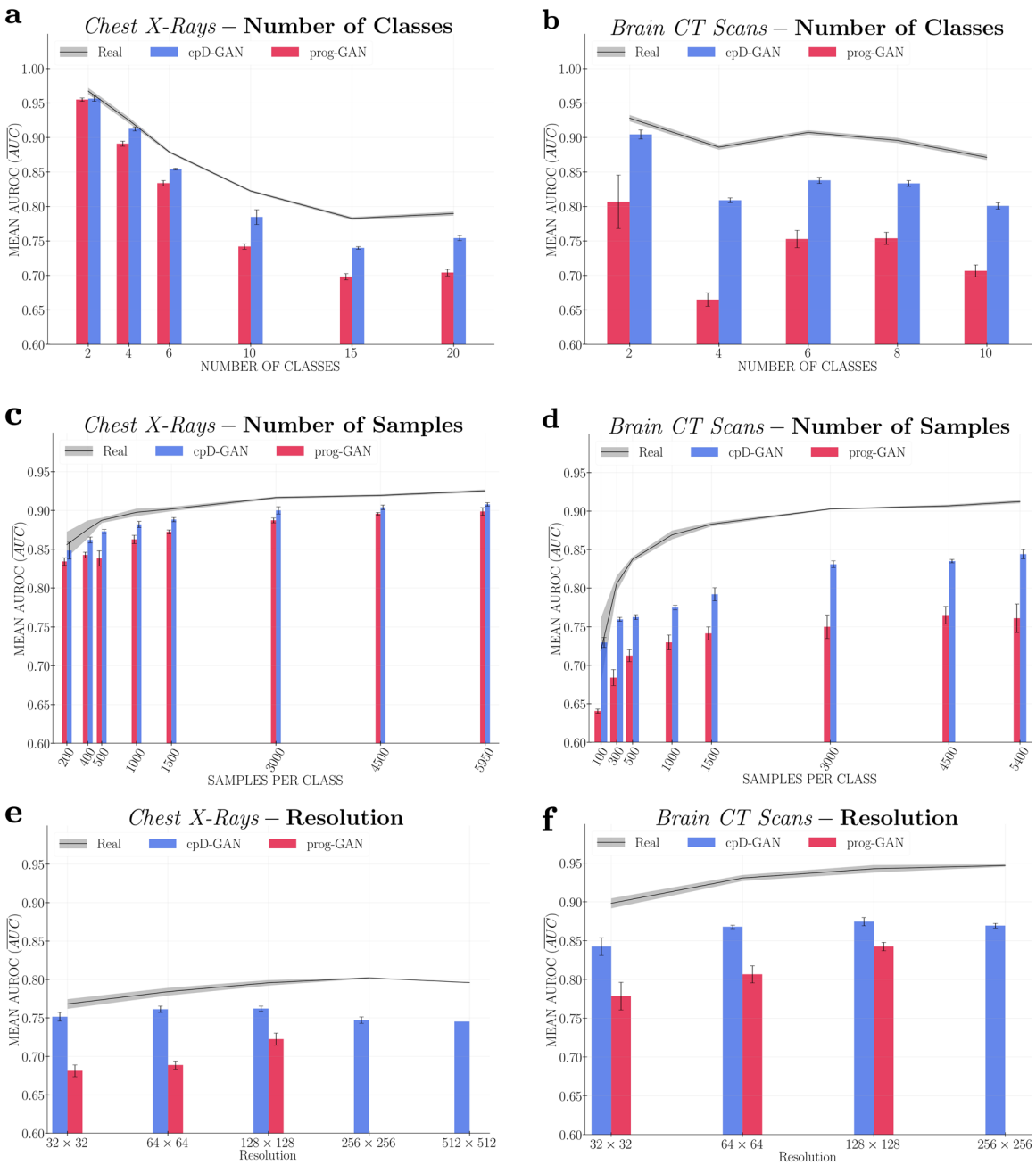

Fig. 2 Benchmark results. In each figure, we show the mean area under the receiver operator characteristic curve score ($\overline{\text{AUC}}$) on a held-out test fold of real data. The $\overline{\text{AUC}}$ scores obtained after training classifiers on real data ($\overline{\text{AUC}}_{\text{real}}$) are indicated by the black line with the shaded area representing the standard deviation across repeated experiments. The bar plots represent the $\overline{\text{AUC}}$ scores achieved after training classifiers on synthetic data ($\overline{\text{AUC}}_{\text{syn}}$) generated by the cpD-GAN (blue) and prog-GAN (red), while the error bars indicate the standard deviation. The sub-figures 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. In (**e**) and (**f**) only the cpD-GAN is evaluated at resolution levels above 128×128 pixels. In (**e**) at 512×512 pixels we perform a single training run. Please see Table 1 for more details on the dataset composition for each benchmark setting.

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

Figure 2c, d shows the benchmark findings when varying the number of samples per class included in each dataset. The predictive performances obtained when training on real and synthetic data remained similar until approximately 3000 samples per class. These results indicate that GAN model performance may be stable when the training data consists of at least 3000 samples per class. Between 1500 and 3000 samples, both the $\overline{\text{AUC}}_{\text{real}}$ and $\overline{\text{AUC}}_{\text{syn}}$ scores started to decrease substantially. However, we also observed a relative performance improvement, meaning decreasing $\overline{\text{AUC}}_{\text{real}} - \overline{\text{AUC}}_{\text{syn}}$ scores for the cpD-GAN, when moving towards low numbers of samples. This effect was particularly

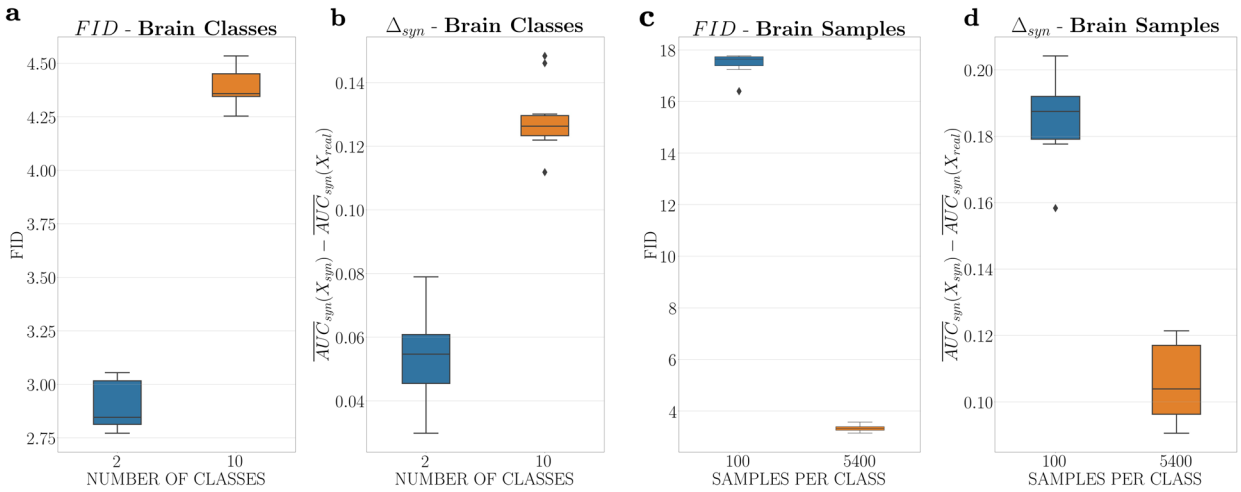

Fig. 3 Label overfitting analysis of the cpD-GAN on brain CT scans for the extrema settings from our benchmark. Box plots showing the median, the interquartile range (IQR = Q3 − Q1), the minimum (Q1 − 1.5 IQR), the maximum (Q3 + 1.5 IQR) and outliers for: **a** FID scores between 10,000 real and synthetic images after GAN convergence for 2 and 10 classes; **b** Δ_{syn} scores, the difference between testing the predictive model on synthetic and real images after training on synthetics for 2 and 10 classes; **c** FID scores between 10,000 real and synthetic images after GAN convergence for 100 and 5400 samples per class; **d** Δ_{syn} scores, the difference between testing the predictive model on synthetic and real images after training on synthetics for 100 and 5400 samples per class. Please see Table 1 for more details on the dataset composition for the extrema settings from our benchmark.

strong for the brain CT scans, where in the extreme setting of 100 samples per class $\overline{\text{AUC}}_{\text{real}} - \overline{\text{AUC}}_{\text{syn}} \leq 0$. Despite the heightened variance in predictive performance, the difference in the scores between the extrema (5950 and 200 samples per class for the chest X-rays and 5400 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 label overfitting during GAN training: given a low number of samples, the variation within real images becomes too low, and the generative model may resort to encoding the class information in unrealistic ways.

To further investigate our claim we analysed two more metrics for the extrema on the number of classes (2 vs. 10) and the number of samples per class (100 vs. 5400) benchmark for the brain CT scans. (1) We looked at the distribution of FID scores for 10,000 real and synthetic images after GAN training convergence. Low FID scores indicate synthetic images that are consistent and with an equivalent visual quality when compared to reals. The FID score is computed in a batch-wise fashion, which means that if the number of real training images is below $N = 10{,}000$ we repeatedly iterate through the real dataset. While the labels that are fed into the GAN will likewise be repetitive, the generated synthetic images will differ due to different input noises. To have comparable FID scores, we need to keep $N = 10{,}000$ constant, as the FID metric is biased with respect to the size of the sample set[43]. (2) We analysed the difference in $\overline{\text{AUC}}$ scores when testing on synthetic and real images for classifiers trained on the synthetic dataset $\Delta_{\text{syn}} = \overline{\text{AUC}}_{\text{syn}}(X_{\text{syn}}) - \overline{\text{AUC}}_{\text{syn}}(X_{\text{real}})$. In the ideal scenario where our synthetic images have captured the real data distribution, we should observe $\Delta_{\text{syn}} \to 0$. If the GAN model faces label overfitting effects by generating images that encode the label information in unrealistic ways, the Δ_{syn} scores should significantly increase.

We show the box plots for the FID and Δ_{syn} scores for the extrema on the number of classes and number of samples per class benchmark for brain CT scans in Fig. 3. In the number of classes' settings we observed the expected behaviour: when including only 2 classes, both the distribution of FID and Δ_{syn} scores was significantly lower compared to 10 classes, which is in agreement with lower $\overline{\text{AUC}}_{\text{real}} - \overline{\text{AUC}}_{\text{syn}}$ scores from our benchmark, showing the improvement in GAN training due to a reduced label space.

However, when we compared the settings for 100 and 5400 samples per class we observed that in the low-data regime we have significantly higher FID and Δ_{syn} scores. We, therefore, believe that the low $\overline{\text{AUC}}_{\text{real}} - \overline{\text{AUC}}_{\text{syn}}$ scores from our benchmark arose due to label overfitting effects. When entering the low-data regime the variation in the GAN training data became too low resulting in synthetic images with lower visual quality (high FID scores). Because the GAN model was not able to generate realistic images, it started to overfit on the class information by unrealistic label encoding. The problem that generators are encouraged to produce images that are particularly easy for auxiliary classifiers to classify has been observed in the literature for several GAN conditioning mechanisms[44,45], and was one of the motivations for the conditional projection-based discriminator of our cpD-GAN[45]. It is an important finding that these effects can also occur for projection-based discriminators, when moving towards low-data regimes. Analysing the exact way in which label information was encoded in the synthetic images and how this encoding is generalised to testing on real data in our $\overline{\text{AUC}}_{\text{real}} - \overline{\text{AUC}}_{\text{syn}}$ score computation, remains open and an important direction for future work.

Impact of resolution

When increasing the resolution from 32×32 pixels to 128×128 pixels, all $\overline{\text{AUC}}$ scores improved, as shown in Fig. 2e, f. However, in terms of relative performance we observed a different behaviour for the two GAN models. For the prog-GAN, we observed a slight increase in $\overline{\text{AUC}}_{\text{real}} - \overline{\text{AUC}}_{\text{syn}}$ scores at a resolution of 64×64 pixels, with substantially lower scores at 128×128 pixels. For the cpD-GAN, the predictive performance on real data increased disproportionately more, resulting in increased $\overline{\text{AUC}}_{\text{real}} - \overline{\text{AUC}}_{\text{syn}}$ scores on both datasets. Above 128×128 pixels, the $\overline{\text{AUC}}_{\text{real}} - \overline{\text{AUC}}_{\text{syn}}$ scores for the cpD-GAN deteriorate further. In general, GAN model training at higher resolutions is less stable and becomes more difficult due to the emergence of fine-scaled details in the images. Moreover, the significant compute demand makes it more difficult to fine-tune the model hyper-parameters at these scales. While we conducted a large-scale hyper-parameter search for the

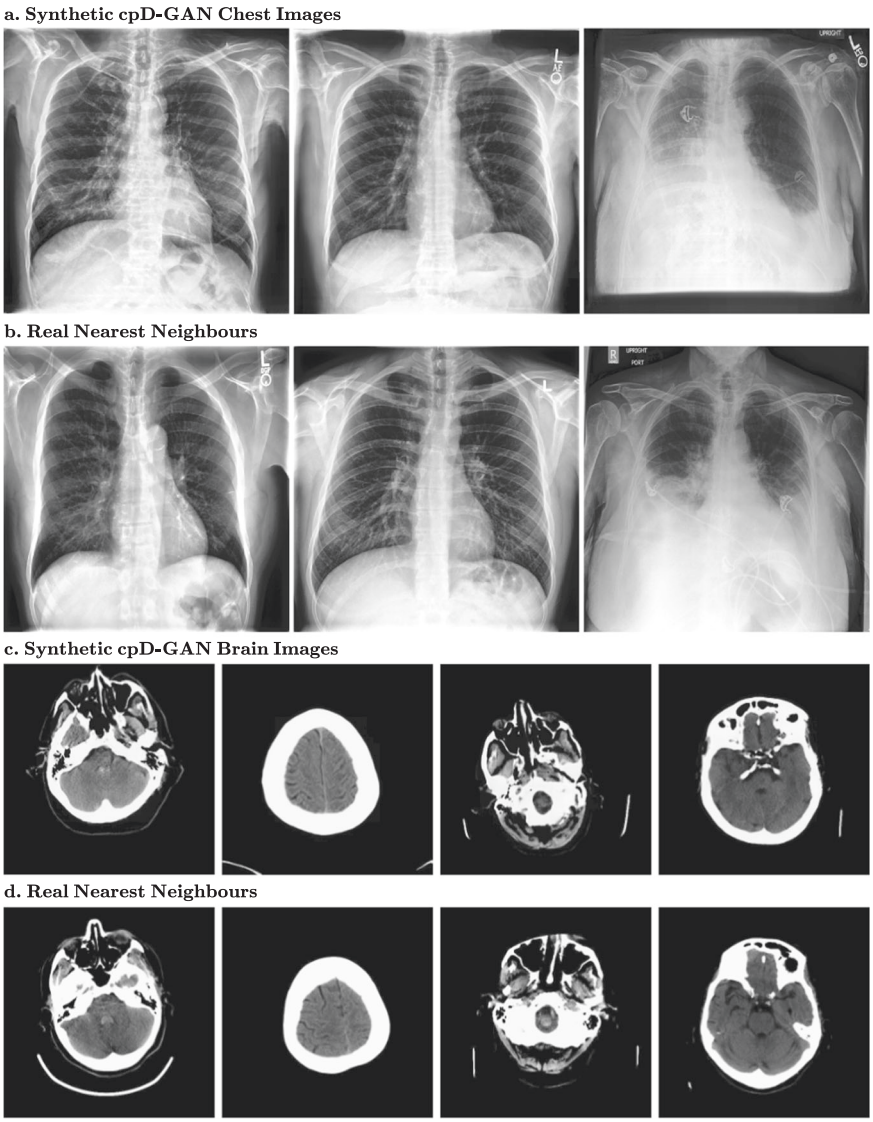

Fig. 4 Randomly sampled synthetic images generated by the cpD-GAN and real nearest-neighbour images from the training. a Synthetic chest radiographs at 512 × 512 pixels. **b** Nearest matching real images found in the chest radiograph training set. **c** Synthetic brain computed tomography (CT) scans at 256 × 256 pixels. **d** Nearest matching real images found in the brain CT training set.

high-resolution settings, which accounted for another 20,480 GPU-hours on NVIDIA's Pascal P100, we could not find training parameters that improved from the previous settings. The negative effect on performance can be clearly seen for the cpD-GAN experiments at a resolution of 256 × 256 and 512 × 512 pixels on the chest X-rays. Compared to CT scans, radiographs are analysed at a higher spatial resolution, and the resulting lower synthetic data quality due to an increased training complexity is accurately detected in the benchmark evaluation. This is in accordance with the results from our reader study and the emergence of visual artefacts for the support devices class. Up until 128 × 128 pixels, we hypothesise 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, compared to the cpD-GAN. Given an even greater amount of computational resources, it remains open whether the cpD-GAN can be fine-tuned to also benefit from resolutions above 256 × 256 pixels. 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.

Image quality and privacy

In Fig. 4, we show randomly sampled synthetic example images from the cpD-GAN at a resolution of 512 × 512 pixels for chest X-rays and 256 × 256 pixels for brain CT scans. Below each synthetic image, we show the most similar real image (nearest neighbour) out of the entire training dataset. For the brain CT scans, there appears to be little noticeable difference in visual quality between the real and synthetic images, which is in agreement with a close-to-random classification accuracy of trained radiologists in the reader study. In contrast, while the chest X-rays also have a high visual quality, there are differences between reals and synthetics, which become more apparent in the form of artefacts for one particular class. The cpD-GAN failed to realistically generate tubes and other support devices, such as pacemakers or defibrillators, as shown in Fig. 5. These devices deviate strongly in their visual appearance when compared to the physiological chest outlining and were not accurately learned by the generative model. Crucially, our benchmark successfully captured the drops in visual quality, as indicated by higher $\overline{\mathrm{AUC}}_{\mathrm{real}} - \overline{\mathrm{AUC}}_{\mathrm{syn}}$ scores. Given the frequent presence of support devices, the GAN should ideally also learn their data distribution.

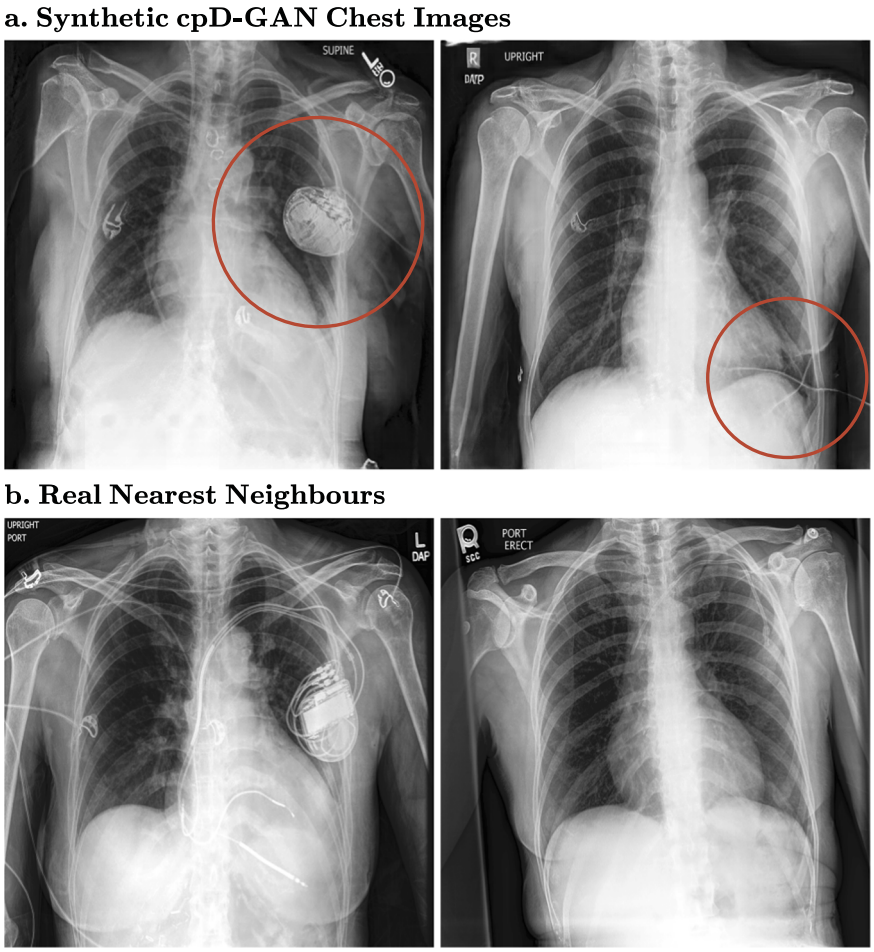

Fig. 5 Synthetic images with visual artefacts. a Synthetic images with visual artefacts, generated by the cpD-GAN, and **b** real nearest-neighbour images from the training dataset at 512×512 pixels. The red circles surround unrealistic image neighbourhoods: Artefacts resulting from the support devices class.

Nonetheless, the cpD-GAN captured the distribution of the physiological anatomic chest structure and related radiology findings, which is more important with respect to diagnosis and clinical practice than external objects. At a spatial resolution of 128×128 pixels, the visual differences are not yet apparent (see Supplementary Fig. 3), which is in accordance with a better benchmark performance. Even in settings where the GAN failed to learn the data distribution of a specific class, it did not start to copy training images: by comparing synthetics and nearest matching neighbours we demonstrate that the cpD-GAN model did not simply memorise training data, and is therefore likely to preserve private, potentially sensitive information. For more high-resolution example images from the cpD-GAN and the respective nearest neighbours, see Supplementary Fig. 2. 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 at 128×128 pixels, as can be seen in Supplementary Figs. 3 and 4.

Feature importance

To gain more interpretability, we analysed the feature importance at a resolution of 512×512 pixels for chest X-rays and 256×256 for brain CT scans from the cpD-GAN, as well as at 128×128 pixels for both datasets and models. In each setting, we estimate the feature importance by successively masking out pixel regions and computing the increased loss[46]. Similar attribution maps indicate that the pixel neighbourhoods in real and synthetic images have a similar causal loss contribution, leading to equivalent predictive models. This suggests that the local image neighbourhoods are consistent in style and texture. While the datasets only allow for classification tasks, we can extrapolate from this analysis that our synthetic images might also perform similarly on problems such as object detection or segmentation, which require global and local image consistency. Figure 6 shows the real images, 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 at high resolution. In Fig. 7, the same analysis is performed at a resolution of 128×128 pixels for both the cpD-GAN and the prog-GAN (for more examples see Supplementary Fig. 5). 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 accordance with our benchmark results, the feature-maps of the cpD-GAN appear more similar to those of real classifiers for a resolution of 128×128 pixels (and 256×256 for brain CT scans), while the differences are greater at 512×512 pixels for chest X-rays. Moreover, in line with the higher $\overline{\text{AUC}}_{\text{real}} - \overline{\text{AUC}}_{\text{syn}}$ scores for the prog-GAN synthetics, the attribution maps also appear 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

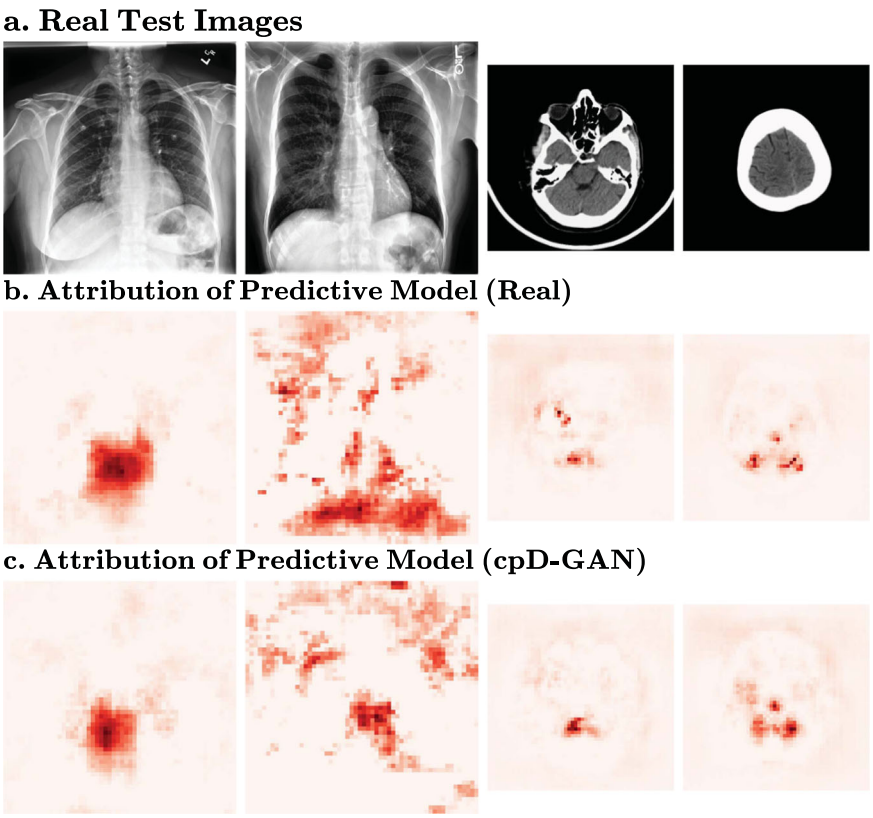

Fig. 6 Feature importance of predictive models trained on cpD-GAN synthetics Deeper red colour indicates regions that have a larger causal contribution to the label prediction. a Nearest neighbours from Fig. 4 and Supplementary Fig. 2, (512×512 pixels for chest X-rays, 256×256 pixels for brain CT scans). From left to right: (1) Chest X-ray with no finding. (2) Chest X-ray with cardiomegaly, atelectasis and support device. (3) Brain scan with no finding. (4) Brain scan with intraventricular haemorrhage. **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.

performance between the classifiers trained on real and synthetic data was greater than zero at those settings.

Reader study

We additionally conducted a reader study in which we asked trained radiologists to label a mixed set of 100 images as real or synthetic (generated by the cpD-GAN) at a resolution of 512×512 pixels for chest X-rays, 256×256 for brain CT scans and 128×128 pixels for both data modalities. 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) in both 128×128 pixel settings. This is in agreement with relatively low $\overline{\text{AUC}}_{\text{real}} - \overline{\text{AUC}}_{\text{syn}}$ scores from our benchmark evaluation. Radiologists were able to differentiate between real and synthetic brain CT scans at 256×256 pixels with an accuracy of 0.532 ± 0.126, which is still close to random, indicating a consistent visual quality and the absence of visual artefacts. In line with higher $\overline{\text{AUC}}_{\text{real}} - \overline{\text{AUC}}_{\text{syn}}$ scores and the emergence of visual artefacts, radiologists achieved an accuracy of 0.710 ± 0.148 when classifying the chest X-rays at 512×512 pixels. The presented results indicate that at lower spatial resolution, trained clinicians cannot discriminate between real and synthetic images, which further substantiates that both the general quality and label information in the synthetic images are realistic. In agreement with our other results, the classification accuracy of radiologists improves with an increasing spatial resolution due to fine-scaled image details and the emergence of visual artefacts, especially for the X-rays.

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 and only on 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 the 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. Our benchmark evaluation and detailed analysis of the synthetic images also shows limitations to medical data generation: due to an increasing training complexity at higher spatial resolution levels, the GAN models fail to accurately learn the data distribution for the support devices class on chest X-rays. This leads to a

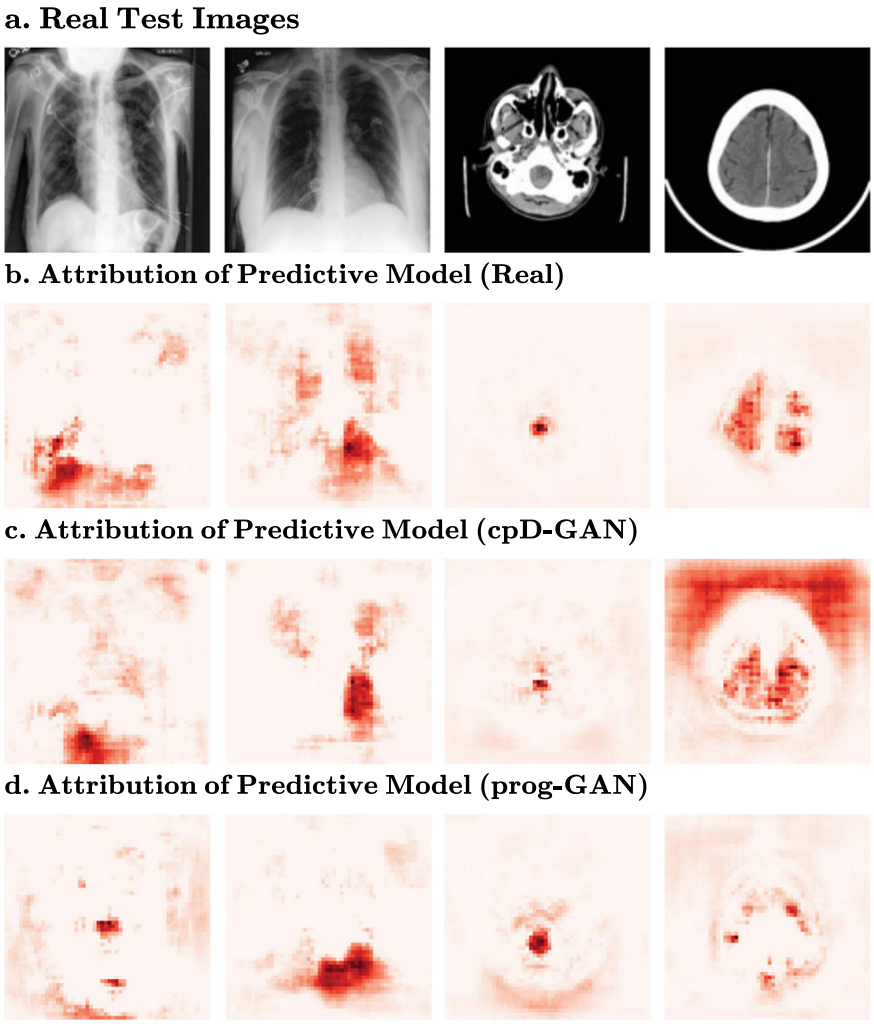

Fig. 7 Feature importance of predictive models trained on cpD-GAN and prog-GAN synthetics. Deeper red colour indicates regions that have a larger causal contribution to the label prediction. **a** Nearest neighbours from Supplementary Figs. 3 and 4 at 128×128 resolution. From left to right: (1) Chest X-ray with support device, lung opacity, pneumonia and atelectasis. (2) Chest X-ray with cardiomegaly and oedema. (3) Brain scan with subarachnoid haemorrhage. (4) Brain scan with subdural haemorrhage. **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.

decreased benchmark performance and synthetic images with a lower visual quality when compared to the real data.

We determined that both GAN models are stable across all benchmark dimensions up to 128×128 pixels, meaning that we did not observe anomalous $\overline{\text{AUC}}_{\text{real}} - \overline{\text{AUC}}_{\text{syn}}$ scores. While some GAN models that we trained were not robust across the experiments (see section 'Methods'), 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. While the $\overline{\text{AUC}}_{\text{real}} - \overline{\text{AUC}}_{\text{syn}}$ scores above 128×128 pixels further increased, there was no training collapse. The visual quality of brain CT scans remained high at 256×256 pixels. For the X-rays, we observed that at a resolution of 512×512 pixels, generated objects from the support devices class suffered from a lower visual quality, resulting in an improved radiologist classification. Crucially, the reduced performance in our benchmark findings supported our detailed analysis of the synthetic images. Also, the cpD-GAN did not fail to capture the distribution of the physiological anatomic structure of the chest, which might be more important for clinical practice. 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 two 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 in certain settings, thereby providing a technical solution to the pervasive issue of data sharing in medicine[1].

The predictive performance obtained when training on synthetic data improved when reducing the number of classes present in 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 encoding label information in unrealistic ways within the synthetic images. This can be very problematic as it can lead to predictions that are based on features not present in the real data distribution when trained on synthetics. The most critical performance improvement between the prog-GAN and cpD-GAN resulted from a revised label conditioning mechanism, rooted in a probabilistic framework[45]. 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 9000), at a number of samples per class of around 3000 without any relative performance drop. However, if low-frequency classes are included and the total number of samples is reduced too much, GAN label overfitting is likely to occur.

In terms of predictive performance in relation to different image resolutions, we found that $\overline{\mathrm{AUC}}_{\mathrm{syn}}$ scores for both models improved when moving up to 128×128 pixels. 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 $\overline{\mathrm{AUC}}_{\mathrm{real}} - \overline{\mathrm{AUC}}_{\mathrm{syn}}$ scores increased when moving up from 32×32 pixels. The dominance of training instabilities increased further when moving towards 256×256 and 512×512 pixels for chest X-rays. This behaviour is not unexpected as stable training becomes more difficult when the discriminator has access to a richer set of features to distinguish real and synthetic data. In line with our benchmark performance decrease and reader study results, we observed that at 512×512 pixels, unstable training resulted in cpD-GAN models that did not capture the real data distribution for the support devices class of chest X-rays. The cpD-GAN was not able to generate consistent and high-quality objects that are not part of the physiological chest outlining, such as tubes, pacemakers, or defibrillators. While it is more important to accurately capture the data distribution of the physiological anatomic chest structure and related radiology findings, the GAN may ideally also generate visually appealing external objects. It remains open whether more hyper-parameter fine-tuning, a refined model architecture, or different training strategies, could improve inconsistencies in the synthetic images. In practice, researcher need to carefully consider at what spatial resolution fine-scaled details emerge that differ significantly from other image parts. Importantly, our benchmark evaluation captured these inconsistencies and robustly detected visual artefacts. For the brain CT scans, we did not observe a lower image quality at a resolution of 256×256 pixels. Accordingly, trained radiologists performed only marginally better than random at differentiating between reals and synthetics. In the presented reader study, we found that at 128×128 pixels, 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 ($\overline{\mathrm{acc}}$) of 50%, to a statistically significant extent. The fact that trained clinicians were unable to discriminate between 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. 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 the quality and predictive performance between synthetic and real medical imaging data, especially when moving to high-resolution. Across all benchmark settings, we observed that only in the extreme label overfitting case $\overline{\mathrm{AUC}}_{\mathrm{real}} - \overline{\mathrm{AUC}}_{\mathrm{syn}} \leq 0$, 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 Figs. 6 and 7, we found that while the real and cpD-GAN predictive models attributed similar regions with high feature importance, they were not identical. In line with our other findings the differences appeared to increase when moving towards high resolutions. 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.

In this study, we analysed synthetic chest radiographs at an overall low resolution compared to clinical practice. Even at 512×512 pixels, the images might lack important details for an accurate clinical diagnosis. A comparison of low-resolution images is more acceptable for evaluating the predictive models as most deep learning systems downsample medical images to reduce the computational requirements. Since training generative models is even more computationally demanding, the lack of unlimited resources represents a major bottleneck for further up-scaling. However, our benchmark results indicate that state-of-the-art GAN models already have difficulties in accurately modelling the support devices class at the analysed spatial resolution. While GAN models such as[26,27] and other generative approaches such as[47,48] work at high-resolution levels, our observed drop in performance indicates that within the medical imaging domain, more model and training improvements are necessary to ensure that the full data distribution is learned in these regimes. Therefore, future work should focus on the current limitations shown in our benchmark evaluation, before scaling GAN training further.

The resolution at which we generate brain CT scans is only marginally smaller than the maximum resolution of the dataset and generally more acceptable when compared to medical practice. However, clinicians analyse 3D CT scans at different intensity windows. Here, we were limited by the RSNA Intracranial Hemorrhage dataset, which consists of pre-sliced 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.

Similarly, the lack of large-scale medical imaging corpora for object detection or segmentation, limited our ability to conduct experiments with downstream tasks, other than classification. Our results from the feature importance analysis and reader study suggest both local and global image consistency up to intermediate resolution levels. This indicates that our benchmark results may also hold for segmentation or detection performance. Nevertheless, it is an important direction for future work, once newly published

medical imaging datasets allow for a similar large-scale benchmark evaluation on the aforementioned tasks.

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 modelling preserves individual patient-level information[49]. In Fig. 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 memorise 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[50], 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.

METHODS

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[51]. In the dataset study, an automatic labelling 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 CT studies supplied by four research institutions and labelled with the help of The American Society of Neuroradiology[52]. It consists of 752,803 CT scan slices of the head from 18,938 unique patients and the corresponding probabilities for the presence of five different haemorrhage 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 haemorrhage, we undersampled the no-finding class, resulting in a balanced dataset where at least 50% of images show some form of haemorrhage.

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 than 256 samples, resulting in 117,168 train images (44,153 patients), 15,318 validation images (5519 patients) and 14,687 test images (5520 patients). For the haemorrhage dataset, we removed label combinations below a frequency of 100, resulting in 173,271 train images (15,133 patients), 22,095 validation images (1892 patients), and 20,500 test images (1892 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 1 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. Each summary refers to the real training, validation and testing dataset. The exact labels from the real settings (combined with random normal noise for variation) are used to generate equivalent synthetic training, validation and testing datasets, before developing and comparing the predictive models.

GAN model development

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 normalisation. For networks operating at up to 32×32 pixels, the generator operates at constant 512 feature channels. At a 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[53] at each optimisation step to stabilise training. The Wasserstein GAN with gradient penalty loss function is used[54]. An additional auxiliary classifier loss term is added to both the generator and discriminator[44] for conditioning. The discriminator is not only trained to classify whether input images are real or fake, but also 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 $n_{\text{critic}} = 1$. At a resolution of 64×64 and 128×128 pixels, we reduced the number of real images per phase to 1M.

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[26,55], 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 $\overline{\text{AUC}}_{\text{real}} - \overline{\text{AUC}}_{\text{syn}}$ 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[45]. Inspired by conditional batch normalisation[56], 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:

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

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. Supplementary Fig. 1 shows the overall model structure and a detailed description of a generator resolution block.

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[57]. However, the implementation did not generalise across different benchmark settings, which is why we excluded it from the sections 'Results' and 'Discussion'. 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 $\tan h$ 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[58]. 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 $\overline{\mathrm{AUC}}_{\mathrm{real}} - \overline{\mathrm{AUC}}_{\mathrm{syn}}$ 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 $n_{\mathrm{critic}} = 2$.

GAN training

We used the Adam optimiser for all GAN models and the hyperparameters as proposed in[23,27], except for the learning rates that we specifically fine-tuned. We stopped training in all settings when the FID between 10,000 real and synthetic images converged. The FID score is a commonly used metric to compare the visual quality between images synthesised by generative models and the real training data and tracking it allows for an unbiased GAN training evaluation[42]. A low FID score means that the visual quality of the synthetic images is close, compared to the set of real images. More precisely, for both sets of images the coding layer representations of a pre-trained Inception model (v3) are extracted to obtain vision-relevant features. The sets of features are summarised as a multivariate Gaussian by calculating the mean ($\mathbf{m}$) and covariance ($\mathbf{C}$), and the FID score is computed as

$$d^2((\mathbf{m}, \mathbf{C}), (\mathbf{m}_w, \mathbf{C}_w)) = ||\mathbf{m} - \mathbf{m}_w||_2^2 + \mathrm{Tr}\,(\mathbf{C} + \mathbf{C}_w - 2(\mathbf{C}\mathbf{C}_w)^{1/2}) \quad (2)$$

where ($\mathbf{m}_w$, $\mathbf{C}_w$) are the mean and covariance of real image feature representations, while ($\mathbf{m}$, $\mathbf{C}$) refers to the statistics for synthetic images and Tr refers to the trace, the sum of the diagonal elements of the matrix. 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. For the number of classes' benchmark, the convergence point for all experiments on both datasets was between 7.0M and 9.6M real images. For the number of samples per class benchmark, the FID convergence occurred between 7.2M and 10.8M real images for sample numbers above 1500. Below that it took between 14M and 16M real images, likely due to the small amount of training data and label overfitting effects. At a resolution of 64×64 and 128×128 pixels, the GAN models converged between 9M and 11M, and between 12M and 14M, respectively. At a resolution of 256×256 pixels for brain scans and X-rays and at 512×512 pixels for X-rays, the GAN models converged at around 7.5M and 8.6M, respectively.

Predictive model development and training

In all settings, we used a pre-trained densenet-121 CNN as the predictive model[59]. In dense CNNs, for each layer in every dense block, a concatenation of the feature-maps of all preceding layers are used as inputs, and its own feature-maps are used as inputs into all subsequent layers. Each block in the network consists of a number of bottleneck and composition layers, which both have batch normalisation and ReLU non-linearities. In each bottleneck layer the number of feature channels is reduced with a 1×1 convolution, before performing a 3×3 convolution in each composition layer. Between the dense blocks, the model consists of transition layers with 1×1 convolutions and average pooling operations. The dense structure allows for significantly deeper architectures without experiencing vanishing gradients, or an exploding number of weights. After the final global average pooling layer, we added a randomly initialised fully connected layer with sigmoid activation for classification with the binary cross-entropy loss. The densnet-121 architecture amounts to 117 convolution, 3 transition and 1 classification layers. 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 ($\overline{\mathrm{AUC}}_{\mathrm{val}}$) across all labels does not improve after two consecutive epochs (patience of 2). If the $\overline{\mathrm{AUC}}_{\mathrm{val}}$ 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.

Statistical tests

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 $\overline{\mathrm{AUC}}_{\mathrm{real}} - \overline{\mathrm{AUC}}_{\mathrm{syn}}$ 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 5950 and 200 samples per class for the chest X-rays, and 5400 and 100 samples per class for brain CT scans. We once again performed the one-sided, parametric-free, Mann-Whitney U test on the $\overline{\mathrm{AUC}}_{\mathrm{real}} - \overline{\mathrm{AUC}}_{\mathrm{syn}}$ scores between the extreme settings to determine the statistical significance.

Nearest neighbours

To analyse the 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 $\overline{\mathrm{AUC}}_{\mathrm{real}} - \overline{\mathrm{AUC}}_{\mathrm{syn}}$ 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 1024-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.

Feature importance

We computed the causal contribution of image neighbourhoods towards the label prediction with the method of Schwab and Karlen[46]. 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. Importantly, there is no learning involved, as we have access to the ground truth labels for all test images. After 12,544 repetitions, all regions of the 224×224 pixel input images were successfully masked. 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 of image neighbourhoods for real and synthetic images, leading to predictive models that attribute the same regions with high feature importance.

Reader study

We conducted the reader study by asking trained radiologists to label a set of 100 images for both data modalities at different resolution levels as real or synthetic with a web-based labelling tool[60]. 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 at 128×128 pixels, 11 radiologists participated, while 9 radiologists labelled the brain CT sets of 128×128 pixels. For each higher-resolution setting (512×512 pixels on chest X-rays and 256×256 pixels on brain CT scans), 5 radiologists participated in the reader study. From each labelled set we computed the values for true reals (TR), false reals (FR), true synthetics (TS) and false synthetics (FS), to determine the classification accuracy $\mathrm{acc} = \frac{\mathrm{TR}+\mathrm{TS}}{\mathrm{TR}+\mathrm{TS}+\mathrm{FR}+\mathrm{FS}}$. For the lower-resolution settings, 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 $\overline{\mathrm{acc}} = 0.5$ (50%). Due to a smaller number of participants, we do not perform statistical tests for the experiments at 256×256 and 512×512 pixels. Instead we only report the accuracy and standard deviation. Please see Supplementary Tables 1 and 2 for details on the 128×128 pixel settings and Supplementary Tables 3 and 4 for details on the 256×256 pixel and 512×512 pixel settings.

Computational cost

The computational cost of our medical imaging benchmark amounts to approximately 31,100 GPU-hours (1338 GPU-days) on NVIDIA's Pascal P100 GPU.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

DATA AVAILABILITY

Both datasets used in our study are publicly available and free to download for any registered user. The CheXpert chest radiograph dataset[51] can be accessed at https://stanfordmlgroup.github.io/competitions/chexpert/ and the RSNA Intracranial Hemorrhage dataset[52] is available at https://www.kaggle.com/c/rsna-intracranial-hemorrhage-detection.

CODE AVAILABILITY

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

Received: 3 November 2020; Accepted: 23 August 2021;
Published online: 24 September 2021

REFERENCES

1. Lo, B. Sharing clinical trial data: maximizing benefits, minimizing risk. *JAMA* **313**, 793–794 (2015).
2. Sanna, S. et al. Causal relationships among the gut microbiome, short-chain fatty acids and metabolic diseases. *Nat. Genet.* **51**, 1 (2019).
3. Li, H. 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. Sun, R. 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. *Lancet Oncol.* **19**, 1180–1191 (2018).
5. Miller, K. et al. Multimodal population brain imaging in the UK Biobank prospective epidemiological study. *Nat. Neurosci.* **19**, 1523–1536 (2016).
6. De Fauw, J. et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nat. Med.* **24**, 1342–1350 (2018).
7. Monteiro, M. 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. *Lancet Digit. Health* **2**, e314–e322 (2020).
8. Liu, Y. et al. A deep learning system for differential diagnosis of skin diseases. *Nat. Med.* **26**, 900–908 (2020).
9. Courtiol, P. et al. Deep learning-based classification of mesothelioma improves prediction of patient outcome. *Nat. Med.* **25**, 1519–1525 (2019).
10. Matsuo, K. et al. Survival outcome prediction in cervical cancer: Cox models vs deep-learning model. *Am. J. Obstet. Gynecol.* **220**, 381.e1–381.e14 (2019).
11. Chen, H., Engkvist, O., Wang, Y., Olivecrona, M. & Blaschke, T. The rise of deep learning in drug discovery. *Drug Discov. Today* **23**, 1241–1250 (2018).
12. Zagribelnyy, B. et al. Deep learning enables rapid identification of potent DDR1 kinase inhibitors. *Nat. Biotechnol.* **37**, 1038–1040 (2019).
13. Lecun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
14. Esteva, A. et al. A guide to deep learning in healthcare. *Nat. Med.* **25**, 24–29 (2019).
15. Haas, S., Wohlgemuth, S., Echizen, I., Sonehara, N. & Müller, G. Aspects of privacy for electronic health records. *Int. J. Med. Inform.* **80**, e26–e31 (2011).
16. Bycroft, C. et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
17. Clark, K. et al. The cancer imaging archive (TCIA): Maintaining and operating a public information repository. *J. Digit. Imaging* **26**, 1045–1057 (2013).
18. Tomczak, K., Czerwińska, P. & Wiznerowicz, M. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Wspolczesna Onkologia* **1A**, A68–A77 (2015).
19. van Panhuis, W. G. et al. A systematic review of barriers to data sharing in public health. *BMC Public Health* **14**, 1144 (2014).
20. Phillips, M. International data-sharing norms: from the OECD to the General Data Protection Regulation (GDPR). *Hum. Genet.* **137**, 575–582 (2018).
21. Na, L. 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 Netw. Open* **1**, e186040 (2018).
22. Nwankwo, I., Hänold, S. & Forgó, N. Legal and ethical issues in integrating and sharing databases for translational medical research within the EU. In *IEEE 12th International Conference on BioInformatics and BioEngineering, BIBE 2012,* 428–433 (IEEE, 2012).
23. Karras, T., Aila, T., Laine, S. & Lehtinen, J. Progressive growing of GANs for improved quality, stability, and variation. In *Proceedings of the 6th International Conference on Learning Representation* (ICLR, 2018).
24. Goodfellow, I. et al. Generative adversarial nets. In *Advances in Neural Information Processing Systems* **27**, 2672–2680 (Curran Associates, Inc., 2014).
25. Mescheder, L., Nowozin, S. & Geiger, A. The numerics of GANs. In *Advances in Neural Information Processing Systems* **30**, 1825–1835 (Curran Associates, Inc., 2017).
26. Karras, T. et al. Analyzing and improving the image quality of StyleGAN. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition.* 8107–8116 (IEEE, 2020).
27. Brock, A., Donahue, J. & Simonyan, K. Large scale GAN training for high fidelity natural image synthesis. In *Proceedings of the 7th International Conference on Learning Representation* (ICLR, 2019).
28. Costa, P. et al. End-to-end adversarial retinal image synthesis. *IEEE Trans. Med. Imaging* **37**, 781–791 (2018).
29. Zhao, H., Li, H., Maurer-Stroh, S. & Cheng, L. Synthesizing retinal and neuronal images with generative adversarial nets. *Med. Image Anal.* **49**, 14–26 (2018).
30. Izadi, S., Mirikharaji, Z., Kawahara, J. & Hamarneh, G. Generative adversarial networks to segment skin lesions. In 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), 881–884 (IEEE, 2018).
31. Bissoto, A., Perez, F., Valle, E. & Avila, S. 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*, 294–302 (Springer International Publishing, 2018).
32. Ali, I. S., Mohamed, M. F. & Mahdy, Y. B. Data augmentation for skin lesion using self-attention based progressive generative adversarial network. In *Expert Systems with Applications.* **165**, 113922 (Elsevier, 2021).
33. Quiros, A. C., Murray-Smith, R. & Yuan, K. Pathology GAN: learning deep representations of cancer tissue. In *Proceedings of the Third Conference on Medical Imaging with Deep Learning*, *PMLR*, Vol. **121**, 669–695 (MLResearchPress, 2020).
34. Zhou, Y. et al. Generating high resolution digital mammogram from digitized film mammogram with conditional generative adversarial network. In *Proc. SPIE Medical Imaging 2020: Computer-Aided Diagnosis* (eds. Hahn, H. K. & Mazurowski, M. A.). Vol. **11314**, 508–513 (SPIE, 2020).
35. Chuquicusma, C. J. M., Hussein, S., Burt, J. & Bagci, U. How to fool radiologists with generative adversarial networks? A visual Turing test for lung cancer diagnosis. In *Proceedings of the International Symposium on Biomedical Imaging*, 240–244 (IEEE, 2018).

36. Han, T. et al. Breaking medical data sharing boundaries by using synthesized radiographs. *Sci. Adv.* **6**, eabb7973 (2020).
37. Han, C. et al. Infinite brain MR images: PGGAN-based data augmentation for tumor detection. In *Neural Approaches to Dynamics of Signal Exchanges. Smart Innovation, Systems and Technologies*, Vol. 151 (eds. Esposito A., Faundez-Zanuy M., Morabito F. & Pasero E.) 291–303 (Springer Singapore, Singapore, 2020).
38. Shin, H.-C. et al. Medical image synthesis for data augmentation and anonymization using generative adversarial networks. In *Simulation and Synthesis in Medical Imaging* (eds. Gooya A., Goksel O., Oguz I. & Burgos N.) 1–11 (Springer International Publishing, 2018).
39. Nie, D. et al. Medical image synthesis with deep convolutional adversarial networks. *IEEE Trans. Biomed. Eng.* **65**, 2720–2730 (2018).
40. Armanious, K. et al. MedGAN: medical image translation using GANs. *Comput. Med. Imaging Graph.* **79**, 101684 (2020).
41. Yang, X., Lin, Y., Wang, Z., Li, X. & Cheng, K.-T. Bi-modality medical image synthesis using semi-supervised sequential generative adversarial networks. *IEEE J. Biomed. Health Inform.* **24**, 855–865 (2020).
42. Heusel, M. et al. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, 6629–6640 (ACM, 2017).
43. Chong, M. & Forsyth, D. Effectively unbiased fid and inception score and where to find them. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 6069–6078 (IEEE, 2020).
44. Odena, A., Olah, C. & Shlens, J. Conditional image synthesis with auxiliary classifier GANs. In *Proceedings of the 34th International Conference on Machine Learning Research*, Vol. **70**, 2642–2651 (PMLR, 2017).
45. Miyato, T. & Koyama, M. cGANs with projection discriminator. In *Proceedings of the 6th International Conference on Learning Representation* (ICLR, 2018).
46. Schwab, P. & Karlen, W. CXPlain: causal explanations for model interpretation under uncertainty. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*. (Curran Associates Inc., 2019).
47. Vahdat, A. & Kautz, J. NVAE: a deep hierarchical variational autoencoder. In *Advances in Neural Information Processing Systems* 33, 19667-19679 (Curran Associates, Inc., 2020).
48. Dhariwal, P. & Nichol, A. Diffusion models beat GANs on image synthesis. Preprint at https://www.arxiv-vanity.com/papers/2105.05233/ (2021).
49. Dwork, C. & Roth, A. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.* **9**, 211–407 (2014).
50. Hyland, S. L. & Tople, S. On the intrinsic privacy of stochastic gradient descent. Preprint at https://arxiv.org/pdf/1912.02919.pdf (2019).
51. Irvin, J. et al. CheXpert: a large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33, 590–597 (AAAI, 2019).
52. Flanders, A. E. et al. Construction of a machine learning dataset through collaboration: the RSNA 2019 brain CT hemorrhage challenge. *Radiol. Artif. Intell.* **2**, e190211 (2020).
53. He, K., Zhang, X., Ren, S. & Sun, J. Delving deep into rectifiers: surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE International Conference on Computer Vision*. 1026–1034 (IEEE, 2015).
54. Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V. & Courville, A. Improved training of Wasserstein GANs. In *Advances in Neural Information Processing Systems*. 5767–5777 (Curran Associates Inc., 2017).
55. Karras, T., Laine, S. & Aila, T. A Style-based generator architecture for generative adversarial networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 4396–4405 (IEEE, 2019).
56. de Vries, H. et al. Modulating early visual processing by language. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*. 6597–6607 (Curran Associates Inc., 2017).
57. Zhang, H., Goodfellow, I., Metaxas, D. & Odena, A. Self-attention generative adversarial networks. In *Proceedings of the 36th International Conference on Machine Learning*, Vol. 97, 7354–7363 (MLResearchPress, 2019).
58. Miyato, T., Kataoka, T., Koyama, M. & Yoshida, Y. Spectral normalization for generative adversarial networks. In *Proceedings of the 6th International Conference on Learning Representation* (ICLR, 2018).
59. Huang, G., Liu, Z. & Weinberger, K. Q. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4700–4708 (IEEE, 2017).
60. Labelbox Inc. The leading training data platform for data labeling. https://labelbox.com (Labelbox, 2020).

ACKNOWLEDGEMENTS

The authors would like to thank the Max Planck Institute for Intelligent Systems and GlaxoSmithKline plc for providing the computational resources for this study.

AUTHOR CONTRIBUTIONS

A.D.S. developed and implemented the code related to the machine learning models, data pre-processing pipeline, and post-processing analysis used in this study. A.D.S., J.H., S.G., T.H., B.D., S.B. and P.S. were involved in the conception and design of the work, and the evaluation of the results. J.H. and S.G. provided the clinical motivation. A.D.S., S.G. and T.H. supervised the reader study. All authors were involved in the reviewing, drafting, and editing process of the manuscript. A.D.S. and P.S. supervised the work.

FUNDING

Open Access funding enabled and organized by Projekt DEAL.

COMPETING INTERESTS

Dr. Patrick Schwab is an employee and shareholder of GlaxoSmithKline plc. The rest of the authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41746-021-00507-3.

Correspondence and requests for materials should be addressed to August DuMont Schütte.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.