COMPUTER SCIENCE

Breaking medical data sharing boundaries by using synthesized radiographs

Tianyu Han[1], Sven Nebelung[2], Christoph Haarburger[3], Nicolas Horst[4], Sebastian Reinartz[1,5], Dorit Merhof[4,6,7], Fabian Kiessling[6,7,8], Volkmar Schulz[1,6,7]*†, Daniel Truhn[3,5]*

Computer vision (CV) has the potential to change medicine fundamentally. Expert knowledge provided by CV can enhance diagnosis. Unfortunately, existing algorithms often remain below expectations, as databases used for training are usually too small, incomplete, and heterogeneous in quality. Moreover, data protection is a serious obstacle to the exchange of data. To overcome this limitation, we propose to use generative models (GMs) to produce high-resolution synthetic radiographs that do not contain any personal identification information. Blinded analyses by CV and radiology experts confirmed the high similarity of synthesized and real radiographs. The combination of pooled GM improves the performance of CV algorithms trained on smaller datasets, and the integration of synthesized data into patient data repositories can compensate for underrepresented disease entities. By integrating federated learning strategies, even hospitals with few datasets can contribute to and benefit from GM training.

INTRODUCTION

The application of computer vision (CV) in medicine promises to personalize diagnosis, decision management, and therapy based on the combination of patient information with knowledge of thousands of experts and the outcomes of billions of patients. In recent years, scientific effort has focused on applications of CV in medicine, in particular in radiology (*1*). Where there has been progress toward this vision of an omniscient radiological CV, this has mostly been anticipated by corresponding technical advances in the field of CV on natural images. A prominent example is convolutional neural networks (CNNs), which had their breakthrough when the performance of AlexNet surpassed more conventional CV algorithms in 2012 (*2*). Since then, CNNs have matched and even surpassed human performance on natural image recognition tasks (*3*). Similar developments took place in medicine, where CNNs performed comparably to the performance of experts in computed tomography (CT) screening for lung cancer (*4*) and retinal disease detection (*5*). However, human performance in CV on medical images has so far only been achieved but not surpassed. Whenever human performance in CV on medical images was achieved, large datasets were used, often pooled from many sites, containing thousands of images. Going a step further and surpassing human performance in CV on natural images, however, always required even larger databases containing up to billions of natural images (*6*).

Unfortunately, collecting and sharing such large quantities of medical images seem inconceivable, caused, in part, by their insufficient public availability. Even if the combined data worldwide reach billions of images, like in the case of thoracic radiographs, patient privacy issues prohibit combining data from multiple sites. This is even more conspicuous given that the majority of patients are willing to share their data for research purposes if adequate measures have been taken to protect their privacy (*7*). Secure ways to share and merge medical images are essential for the development of future CV algorithms (*8*).

Federated learning has gathered attention and is suitable where data sharing is hindered by privacy considerations. In this paradigm, a central model is updated by exchanging encrypted gradients or weights between global and selected models (*9*). To further improve privacy in medical applications, a fraction of weights or gradients within local models can be blurred by injecting random noise, i.e., differential privacy. Such a random module has been successfully integrated into a federated brain segmentor (*10*). However, in the conventional federated learning settings, the central instance cannot inspect the raw training data due to privacy concerns, and hence, modeling tasks become challenging.

Another promising solution to overcome data sharing limitations is the use of generative adversarial networks (GANs), which enable the generation of an anonymous and potentially infinite dataset of images based on a limited database of radiographs. GANs are a special class of neural networks that were first introduced by Goodfellow *et al.* (*11*) in 2014 and have since then been advanced to generate high-resolution, photorealistic synthetic images (*12*). While the first implementations of GANs made it possible to synthesize unconditioned images, the development and usage of informative priors to drive generators that output conditional samples are desired in medical applications. A common choice for such a conditional prior is an existing image as used in pix2pix (*13*) and Cycle-GAN (*14*). Recently, Cycle-GAN–based networks have gained attention in the medical imaging community due to their capabilities of achieving intermodality image transitions. On the basis of Cycle-GAN frameworks, researchers such as Wolterink *et al.* (*15*) and Chartsias *et al.* (*16*) successfully demonstrated bidirectional CT–magnetic resonance imaging (MRI) transitions in both brain and heart imaging. Furthermore, Zhang *et al.* (*17*) introduced a segmentor-based shape consistency term to the Cycle-GAN loss and achieved realistically looking volumetric CT-MRI data transitions. The performance of

[1]Physics of Molecular Imaging Systems, Experimental Molecular Imaging, RWTH Aachen University, Aachen, Germany. [2]Department of Diagnostic and Interventional Radiology, University Hospital Düsseldorf, Düsseldorf, Germany. [3]Aristra GmbH, Berlin, Germany. [4]Institute of Imaging and Computer Vision, RWTH Aachen University, Aachen, Germany. [5]Department of Diagnostic and Interventional Radiology, University Hospital Aachen, Aachen, Germany. [6]Fraunhofer Institute for Digital Medicine MEVIS, Bremen, Germany. [7]Comprehensive Diagnostic Center Aachen (CDCA), University Hospital RWTH Aachen, Aachen, Germany. [8]Institute for Experimental Molecular Imaging, RWTH Aachen University, Aachen, Germany.
*These authors contributed equally to this work.
†Corresponding author. Email: schulz@pmi.rwth-aachen.de

segmentors and GANs was boosted by shape consistency and online augmentation, respectively. Nevertheless, image-based conditioning always carries the risk of leaking patient-sensitive data to the generator during the training process.

Here, we propose to use generative models (GMs) on the basis of convolutional GANs (*18*) to break the boundary of sharing medical images and to enable merging of disparate databases without the limitations that are now restricting the collection of radiographs in a public database (see Fig. 1). To demonstrate the performance of our concept, we show that fully synthetic and thus anonymous images can be generated, which look deceivingly real—even to the expert's eye—and that these images can be used in the medical data sharing process. Our concept proposes how medical images or data can be shared in the future.

RESULTS

Generation of synthesized radiographs

Generating synthesized two-dimensional images in high resolution is a nontrivial task and has just recently been made feasible by using progressive growing during training (*12*) or by using large-scale networks that demand massive amounts of computing power. As the computing power required for the latter approach is, in general, not accessible to most hospitals, we used progressive spatial resolution growing during training of our networks. Thus, the GAN was trained by starting with a spatial resolution of 4 × 4 and stepping up in powers of 2 (8 × 8, 16 × 16, 32 × 32, 64 × 64, 128 × 128, 256 × 256, 512 × 512) to a spatial resolution of 1024 × 1024.

We measured the time needed to train a GAN on a dataset size of 112,120 radiographs with a hardware setup that is accessible to any small hospital: We used a desktop computer with an Intel Xeon(R) E5-2650 v4 processor (Intel, Santa Clara, CA) and an Nvidia Tesla P100 16 GB GPU (Nvidia, Santa Clara, CA). To completely train the GAN with this setup to generate synthesized x-rays with a spatial resolution of 256 × 256 took 60 hours. Continuing the training to generate synthesized x-rays of spatial resolutions 512 × 512 and 1024 × 1024 took 114 and 272 hours of computational time, respectively. Once the training had finished, inference, i.e., generation of synthesized radiographs, was much faster with a rate of 67,925, 41,379, and 4511 generated radiographs per hour at the three spatial resolution stages. Sample images are shown in fig. S4A for a spatial resolution of 256 × 256. Further images for spatial resolutions of 512 × 512 and 1024 × 1024 are given in the Supplementary Materials.

We have chosen the multiscale structural similarity (MS-SSIM) as a metric (*19*) to detect a possible mode collapse of our GAN (i.e., missing diversity in the images). The MS-SSIM has been successfully used in predicting perceptual similarity judgments of humans. A lower MS-SSIM reflects perceptually distinct samples and proves the high diversity of a dataset. In fig. S2, we depict the MS-SSIM of 1000 randomly selected pairs of samples within a given pathology class. As can be seen, the overall MS-SSIM among synthesized pairs is comparable to that of real sample pairs.

Ability of human readers to distinguish synthesized radiographs from real x-ray images

To test the quality of the synthesized radiographs (i.e., radiographs synthesized by the generator), six readers were presented 50 synthesized radiographs each and 50 radiographs of real patients in randomized order, and the readers were separately tasked with deciding whether the presented radiograph was real or synthesized. The tests were repeated with spatial resolutions of 256 × 256, 512 × 512, and 1024 × 1024, resulting in a total of 18 tests with 100 radiographs each.

To assess whether experience with machine learning or radiological expertise was necessary to identify synthesized radiographs, the readers were grouped and chosen as follows: group 1 consisted of three readers that had a background in CV (readers 1, 2, and 3, who had 4, 2, and 5 years of experience in CV, respectively), while group 2 consisted of experienced radiologists (readers 4, 5, and 6, who had 4, 19, and 6 years of experience in general radiology, no dedicated specialization to thoracic radiology).

Accuracies in differentiating the synthesized images from the real images at spatial resolutions of 256 × 256 were 60 ± 5% for group 1 and 51 ± 5% for group 2. Generating convincing radiographs at higher spatial resolutions proved more difficult, and experts were able to distinguish real from synthesized radiographs

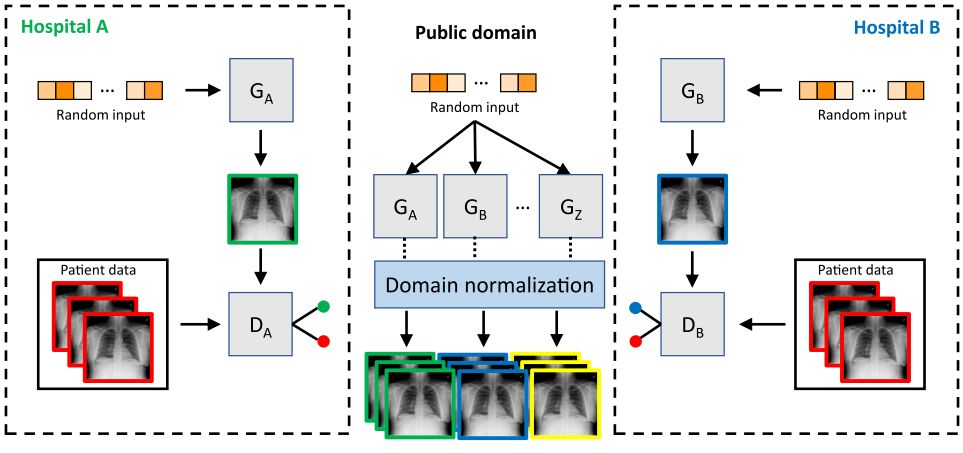

Fig. 1. Concept of constructing a public database without disclosing patient-sensitive data. The GAN in each hospital consists of a generator G and a discriminator D. During training, patient-sensitive data (shown in red) are never exhibited to the generators G directly. Patient-sensitive data is only exhibited to discriminator D while it is trying to differentiate between real and synthesized radiographs. After training is completed, only the generators G are transferred to a public database and can be used to generate synthesized radiographs.

more easily at spatial resolutions of (512 × 512)/(1024 × 1024) with accuracies of 67 ± 17%/82 ± 4% for group 1 and 65 ± 5%/77 ± 13% for group 2. Thus, radiologists and CV experts performed similarly when identifying synthesized radiographs at high resolutions when judged by their accuracy. As shown in table S2, the sensitivity, i.e., the correct identification of a synthesized radiograph, was higher than the specificity, i.e., the correct identification of a real radiograph. This is probably attributable to the fact that some synthesized radiographs show telltale signs of synthesization (see fig. S4E) and thus allow for a more reliable identification. While radiologists predominantly detected errors in anatomical details such as bone shape or rib cage morphology, CV experts tended to focus more on tiny details such as wave-like patterns (see fig. S4E). There was no interreader agreement between the readers for spatial resolutions of 256, underlining the fact that identification of synthesized radiographs at this spatial resolution stage is hardly possible (see Table 1). At higher spatial resolutions, the interreader agreement was consistently higher following the found higher accuracy in identifying synthesized radiographs. These results were observed under restrictions: The radiographs were assessed on conventional 24-inch computer monitors without zooming into the images. The radiographs were presented in a given order: first, the low-resolution radiographs, followed by the mid- and then the high-resolution radiographs. The readers were not allowed to go back and change previous decisions. When these restrictions were lifted, accuracy in determining whether a radiograph is real or synthesized was significantly increased. A radiologist with 9 years of experience who was given unlimited amounts of time and who first examined the high-resolution radiographs on specialized radiological monitors to identify typical GAN-related artifacts before going back to the 256 × 256 radiographs was able to identify synthesized radiographs in 86% of cases.

The difficulties to generate convincing radiographs at high resolutions were understandable, as the task becomes more difficult with the growing number of pixels: even for low-resolution grayscale images of 100 × 100 pixels and 8-bit grayscale depth, the number of possible different images amounts to $256^{(100 \times 100)}$. The GAN was tasked with identifying the subset of real-looking images out of this set that grows exponentially in size with increasing spatial resolution. Not unexpectedly, this process was not perfect, and although the GAN managed to capture the general appearance of a real radiograph at high resolutions, small details revealed the synthesized origin. After having performed the tests and with the knowledge of the ground truth, the readers conferred to identify these typical patterns that allowed for the differentiation of real from synthesized images at high resolution. Among these were unphysiological configurations of the pulmonary vessels, aberrant bone structures, and subtle periodic, wave-like patterns superimposed on the lung parenchyma, which reflect the network's difficulty to generate fine details (see fig. S4E).

Ensuring non-transference of private information

To exclude the possibility that the GAN memorizes and subsequently merely reproduces the given training examples, 1000 randomly synthesized radiographs were generated, and their nearest neighbors in the database of real radiographs were sought according to the structured similarity index. All 1000 radiographs along with their respective three nearest neighbors were then plotted, and a board-certified radiologist assessed whether an entity from the database of real radiographs had been duplicated.

In this set of 1000 randomly drawn GAN images, we did not find a single instance in which the synthesized radiograph looked identical to its closest neighbor in the real dataset (fig. S4B). When assessing similarity in terms of the SSIM, we did not find a single case in a set of 10^5 randomly drawn synthesized radiographs, in which a digital twin was found in any of the real radiographs. In addition, the reader was asked to examine the synthesized radiograph for local information that might lead to the identification of a specific patient, e.g., an anatomic variant unique to a patient or a necklace with a name on it. No such information was found in this set of 1000 images.

We reason that the duplication of images from the database of real radiographs is unlikely. The GAN consists of a generator and a discriminator network. Only the discriminator network will be in direct contact with patient images. The generator is never directly presented a patient image in the training process. Thus, only the part of the architecture (the generator) that has never been presented with real patient images is transferred to the central database.

Performance of classifiers trained on synthesized radiographs

To demonstrate the feasibility of our approach in a clinical setting, as shown in Fig. 1, we have decided to apply our concept to the detection of pneumonia. In the United States alone, pneumonia accounted for over 500,000 visits to emergency departments and over 50,000 deaths in 2015 (*20*). The Radiological Society of North America (RSNA) has recently hosted a challenge to automatically detect pneumonia in x-rays using machine learning algorithms. Often, local hospitals can only gather medical datasets with limited diversity due to a specific patient population with associated pathologies. However, diversity of the datasets is crucial to the performance of deep learning algorithms due to the complex features of a specific pathology. By using our approach of pooled GANs, different patients from different locations can be jointly considered and thus boost the diversity of the local dataset without violating any privacy protection

Table 1. Real/synthesized radiographs test. Accuracy and interreader agreement for the group of three CV experts, three radiologists, and all readers when differentiating whether the presented radiograph is real or synthesized.

	256 × 256		512 × 512		1024 × 1024	
	Accuracy, %	Fleiss' kappa	Accuracy, %	Fleiss' kappa	Accuracy, %	Fleiss' kappa
CV experts	60 ± 5	−0.03	67 ± 17	0.07	82 ± 4	0.46
Radiologists	51 ± 5	0.10	65 ± 5	0.18	77 ± 13	0.39
All readers	55 ± 7	0.00	67 ± 14	0.07	80 ± 10	0.37

laws. We simulated a local dataset with a limited diversity by using a subset of the RSNA dataset with 1000 real x-rays for training, of which only 5% exhibited signs of pneumonia. The resulting classifier achieved an area under the curve (AUC) of 0.74 on a test set of 6000 previously unseen x-rays from the RSNA dataset (see Fig. 2B). To alleviate the limiting scarcity of pathological images and improve the classifier performance, we used our public database of generated images [trained on the National Institutes of Health (NIH) and the Stanford dataset]. From this database, we randomly sampled a total of 500 synthesized x-rays: 100 that exhibited signs of pneumonia and 400 that exhibited no signs of pneumonia. These were then joined with 500 real x-rays from the RSNA subset (450 healthy and 50 pneumonia), which resulted in a set of 1000 x-rays for training of the classifier (healthy: 450 real and 400 synthesized; pneumonia: 50 real and 100 synthesized). When trained on this artificially enriched set of x-rays, the performance of the classifier increased with an overall AUC of 0.81. We hypothesize that the reason for the improvement in performance was probably due to the greater diversity of pathological cases as produced by the generator: As reflected by the lower MS-SSIM in Fig. 2A, the GAN-augmented dataset (MS-SSIM, 0.18 ± 0.09) achieved a higher level of diversity in comparison with the local RSNA subset (MS-SSIM, 0.24 ± 0.12). Note that for both cases, we have chosen to train the classifiers on the same number of x-rays to exclude any potential influence that the size of the training set might have had on the performance of the classifiers. Similarly, improved performance measures were found for sensitivity, specificity, accuracy, positive predictive value (PPV), negative predictive value (NPV), and F1 score (see Fig. 2C). This experiment thus demonstrated that our pooled dataset approach is capable of improving deep learning classifiers by enriching scarce datasets.

To simulate the data merging scenario as outlined in Fig. 1, we compared the results of a comprehensive pathology classification, i.e., cardiomegaly, effusion, pneumothorax, atelectasis, edema, consolidation, and pneumonia, with a classifier solely trained on the NIH-GAN versus a classifier that was trained on merged synthesized images from different sources. Generated samples of our Stanford-GAN can be found in fig. S5. The average values of the AUC, accuracy, sensitivity, and specificity all increased significantly after integration of the synthesized external dataset (see Fig. 3). This demonstrated that the merging of multiple databases of generated radiographs can boost the performance of classifier networks and can alleviate the performance bottleneck due to insufficient amounts of training data.

The performance improvements have been achieved without any techniques of domain adaption, i.e., without any efforts to homogenize the appearance of the radiographs from different databases. Adopting these techniques not only offers an opportunity for

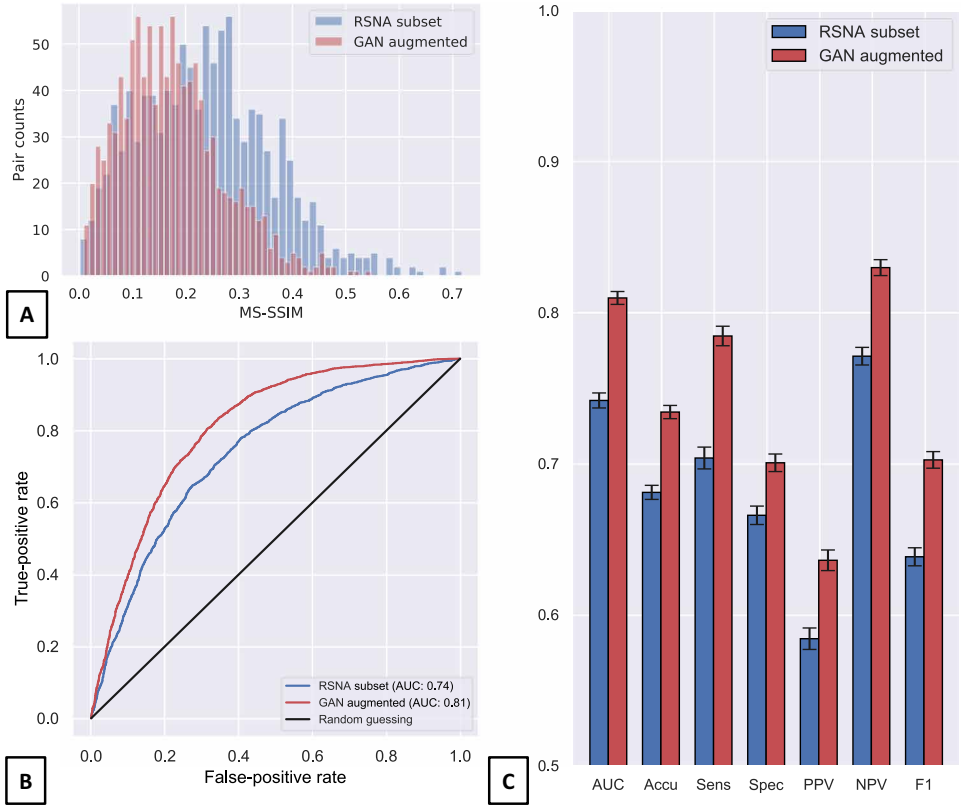

Fig. 2. Pooled GAN training can improve pneumonia detection by enriching the diversity of the dataset. (**A**) Distributions of MS-SSIM of randomly selected 2450 pneumonia-positive pairs. Higher diversity of pneumonia cases in the GAN-augmented dataset is confirmed by a lower MS-SSIM (GAN-augmented MS-SSIM: 0.18 ± 0.09 versus RSNA subset MS-SSIM: 0.24 ± 0.12). (**B**)The performance of the classifier when trained on 1000 x-rays from the GAN-enriched dataset (healthy: 450 real and 400 synthesized; pneumonia: 50 real and 100 synthesized) reaches an AUC of 0.81 in pneumonia detection, outperforming that of a classifier trained on 1000 real x-rays (healthy, 950; pneumonia, 50) that reaches an AUC of 0.74. (**C**) Similarly, improved performance measures were found for sensitivity (Sens), specificity (Spec), accuracy (Accu), PPV, NPV, and F1 score. We used a test set of 6000 x-rays randomly sampled from the RSNA dataset to calculate those scores. The GANs used to generate the synthesized x-rays were trained based on the NIH and Stanford datasets.

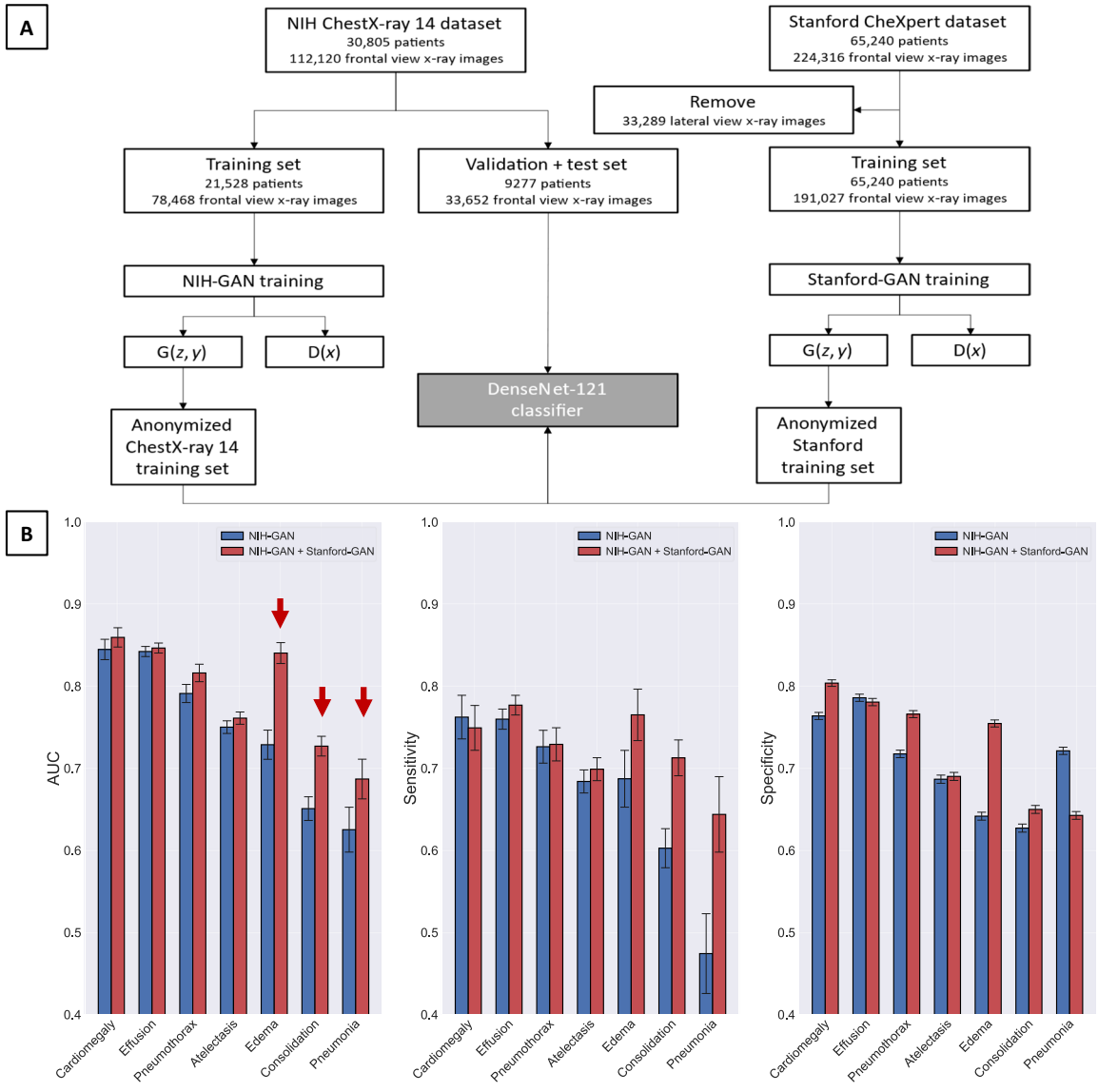

Fig. 3. Using pooled synthesized data from different sites, classification performance can be increased. To simulate the scenario in Fig. 1, two classifiers were trained and compared: a classifier solely trained on anonymous radiographs generated with the NIH-GAN (blue) and a classifier trained on the pooled anonymized dataset generated with the NIH-GAN and the Stanford-GAN (red). The schematic of the data selection process is shown in (**A**). AUC, sensitivity, and specificity for the seven diseases are given in (**B**). In particular, the classification performances of formerly problematic cases such as edema, consolidation, and pneumonia were boosted by merging data from multiple sites (red arrows).

further performance improvements through domain adaption—now, an active area of research (*21*)—but would also most likely make the classifier network more robust to deployments in different environments. This is an important aspect in the translation of CV algorithms from the workbench to clinics.

Federated averaging facilitates the training of local GANs

Large amounts of data are required to obtain robust results from GANs that are trained locally. This potentially limits the sites at which a GAN can be trained to large hospitals. Federated learning algorithms (*22*) offer a remedy to this limitation as the GAN can be trained without the original images, leaving the protected space of the hospital. One possible reservation is that, because of the uncontrollable gradient/model updates, it is difficult to detect adversarial attacks and protect against them (*23*). However, the security of GAN-based federated learning has the advantage of offering an additional degree of freedom for screening of databases by using confidence calibrated checking (*24*) or manual inspection (*25*). We therefore investigated the use of federated learning in training one central GAN as an alternative to the pooled GAN approach.

To simulate hospitals with limited amount of training data, we randomly sampled 20,000 patient radiographs from the Stanford CheXpert dataset and then partitioned them into 20 local clients each receiving 1000 patient radiographs. We trained and compared the following models: a centralized "20k model," which was trained on 20,000 patient radiographs, a centralized "1k model," which was

solely trained on 1000 local radiographs, and a federated "20 × 1k model," which was trained federally (*22*) on 20 distributed datasets consisting of 1000 radiographs each (see Fig. 4A).

An important property of Wasserstein GANs is that their discriminator loss directly reflects the quality of generated samples (*26*). We therefore visualized the negative discriminator loss in Fig. 4B. As can be seen in Fig. 4B, because of insufficient training images, the centralized 1k GAN overfitted quickly and led to an unstable training. However, as indicated by a lower loss and Fréchet inception distance (FID) in Fig. 4 (B and C), the federated trained GAN (federated 20 × 1k) overcame this local overfitting issue and significantly outperformed the locally trained GAN. The loss curve of the federated GAN was smoother because it represented an average over local iteration losses.

Generated images as a visualization of what neural networks see

The images generated by the generator could be specifically controlled: By changing the part of the input vector signifying the disease, radiographs with specific pathologies could be generated. We used two techniques of visualizing the disease-specific hallmark changes. First, the disease-specific entry in the input vector was gradually changed from 0 to 1, while all other entries were kept at 0. The generated images then showed the transition from healthy to diseased states and were stitched together to form an animation. Exemplary frames visualizing the transition are given in fig. S4C for cardiomegaly and effusion. With cardiomegaly, we observed an enlargement of the projected heart shape, reflecting the expected radiological change. Similarly, effusion showed the typical opacification of the lower lung field mirroring the collection of fluid there. Animations for all of the 14 disease states are given in the Supplementary Materials.

Second, the pixel-wise difference image between the fully diseased and the healthy radiograph was calculated and superimposed on the healthy radiograph as a colormap (see Fig. 5A for a schematic of the process). Examples of such found visualizations are given in Fig. 5B for all 14 pathologies.

One advantage of having full control over the disease state of the GAN radiographs is that any combination of diseases in a single radiograph can be generated by changing the corresponding entries in the input vector simultaneously. We found that the disease state as represented by the GAN transition reflected the underlying disease and was in good agreement with radiological expertise if many marked examples of this disease were present in the original dataset and if the disease-related changes occurred on a large scale of the radiographs (e.g., cardiomegaly or effusion) rather than on small patches at different sites (nodules).

To uncover correlations between disparate pathologies, we let the classifier rate the score of a specific pathology when the GAN was tasked with generating a synthesized radiograph of another disease

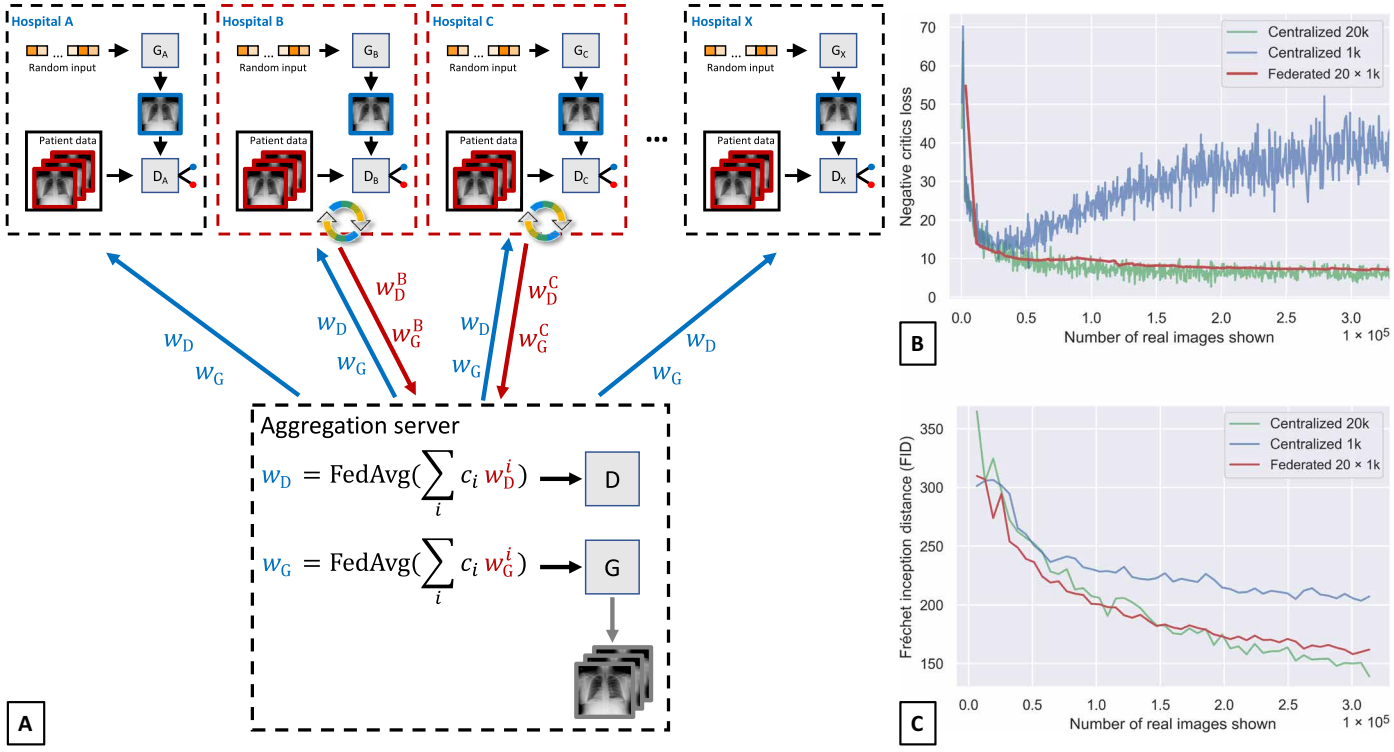

Fig. 4. Federated learning facilitates GAN training when facing insufficient amounts of local data. Hospitals can use federated learning algorithms to train a global GAN, and the central GAN deposit can serve as a hub. (**A**) Illustration of the GAN-related federated learning system. After local model initialization, local hospitals B and C (in red frames) were selected to update their local models. The global generator and discriminator were updated by the weights (*w*) transferred to the aggregation server (red arrows). All local models were subsequently redefined by the updated global GAN (blue arrows). The exchange of local and global weights continued until the global GAN converged. (**B**) Discriminator loss curves for three trained Wasserstein GANs. The Wasserstein GAN trained by federated averaging algorithm (federated 20 × 1k) outperformed the centralized GAN trained on only 1000 x-rays (centralized 1k) and performed comparably to the centralized 20k GAN. (**C**) FID evaluations of the GAN training process.

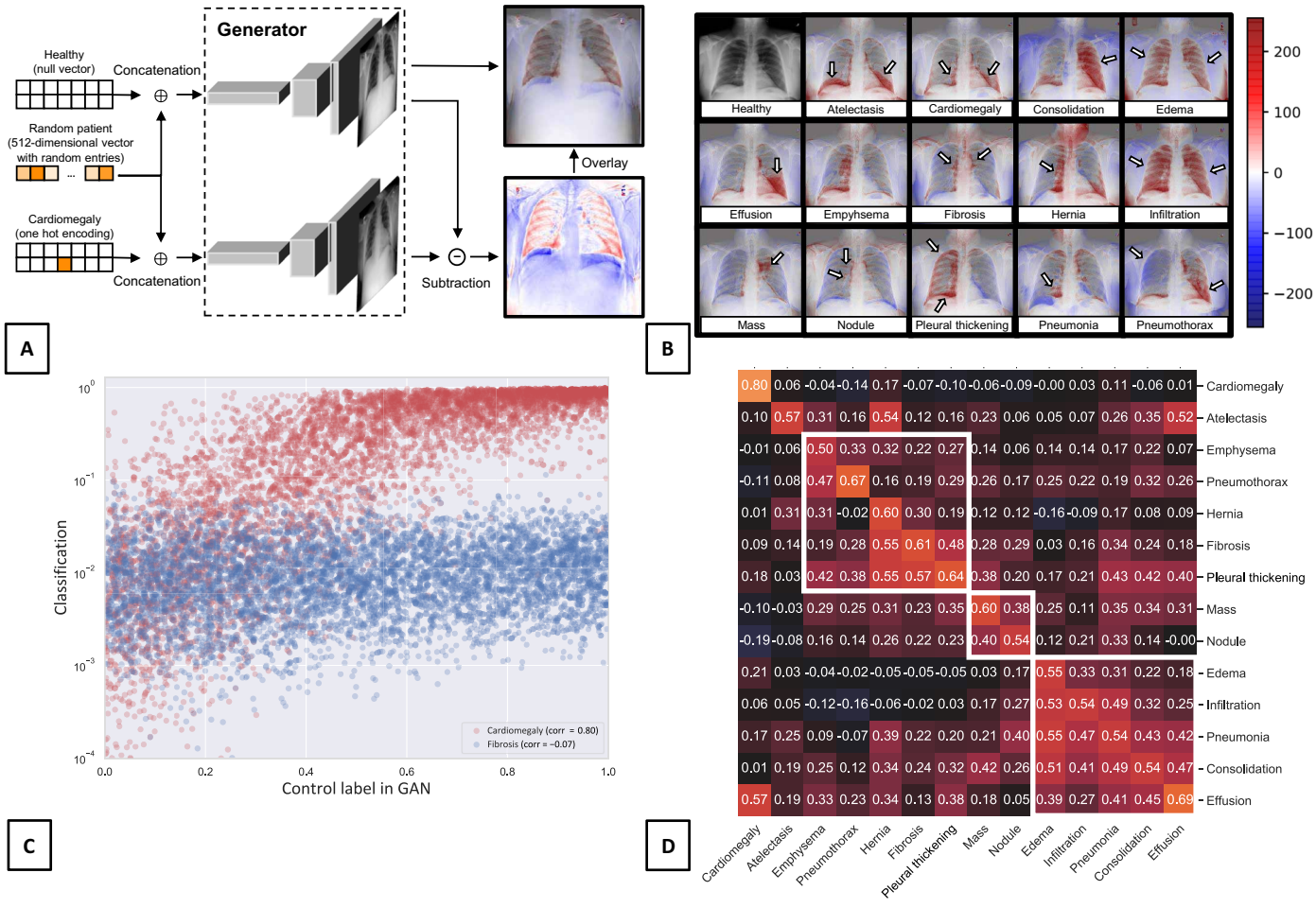

Fig. 5. Learned pathological features. (**A**) Generation of the disease-specific pixel map. A randomly chosen vector with 512 Gaussian distributed entries characterizes one specific patient. The GAN was tasked with generating a healthy and a diseased radiograph of that patient (cardiomegaly in this example). A subtraction map was generated to denote the changes brought about by the disease and was superimposed as a colormap over the generated healthy radiograph. (**B**) Disease-specific patterns generated by the generator for an exemplary randomly drawn pseudopatient. Red denotes higher signal intensity in the pathological radiograph, while blue denotes lower signal intensity. Note that for some diseases such as cardiomegaly and edema, the pattern looks realistic, while the GAN struggled with diseases that have a variable appearance and where ground truth data are limited, e.g., pneumonia. (**C**) Revealing correlations within generated pathological radiographs by the classifier trained on the real dataset. For each pathology, 5000 random synthesized radiographs with a pathology label drawn from a uniform distribution between 0.0 and 1.0 were generated. The images were then rated by the classifier network, and Pearson's correlation coefficient was calculated for each pairing of pathologies [shown in (C) with the GAN cardiomegaly label on the *x* axis and the cardiomegaly and fibrosis classifier output on the *y* axis in red and blue, respectively]. (**D**) Resulting correlation coefficients for all 14 × 14 pairings are displayed and color coded in (B). Clustering on the *x* axis (i.e., the GAN label axis) was performed to group related diseases. The obtained clustered blocks are marked with white-bordered boxes.

and calculated the Pearson's correlation coefficient (Fig. 5C). We found that the clustering of related pathologies based on the correlation coefficient agreed well with clinical intuition: Infiltration, pneumonia, consolidation, effusion, and edema—all pathologies where lung opacity increases—were related to each other while being distant from, e.g., emphysema or pneumothorax—diseases that are associated with increased radiolucency. The magnitude of diagonal elements in the 14 × 14 matrix in Fig. 5D directly reflects the quality of the generation of pathological images with our method. Diseases, such as cardiomegaly (corr = 0.8) and effusion (corr = 0.7) could be reliably generated due to their localized and predictable pathological features. However, the GAN trained on these particular datasets performed less reliable in generating infiltration (corr = 0.5), emphysema (corr = 0.5), and pneumonia (corr = 0.5). This might be due to the limited number of those cases in the datasets. For example, for pneumonia-labeled cases in the NIH dataset, only 1431 cases are positive and 110,689 are negative. As can be seen in Fig. 5D, the NIH generator cannot reliably generate pneumonia-related features. The nature of pneumonia was better captured by the Stanford-GAN in Fig. 5D, which shows the challenging cases from this particular dataset. This GAN was trained with a much higher number of 20,656 pneumonia cases in the Stanford CheXpert dataset.

DISCUSSION

In this study, we demonstrated that GANs can be used to generate deceivingly real-looking radiographs and to merge databases of

radiological images without disclosing patient-sensitive information. This helps to build large radiological image databases for the training of CV algorithms. While radiographs may, in principle, be abundantly available, universal access is, in general, severely restricted due to data protection laws: privacy concerns restrict the export of sensitive patient information to extramural institutions, and often, only a small fraction of the available data can be used (e.g., a patient consent form may not be universally available). In these cases, GANs that have been trained in-house may serve as a mean to distribute the information contained within the database without actually providing a real snapshot of patient-sensitive data: only the weight distribution of the GAN needs to be transferred, and a representative synthesized dataset of millions of radiographs may be generated in reasonable computational time at a peripheral site. This is in contrast to previous works of Shin and co-workers, in which lower–spatial resolution synthesized images could be produced but always required recourse to the original patient images as inputs to an image to image translational network (*13*). Another group has previously demonstrated that synthesized chest radiographs can be used to augment training, see, e.g., Salehinejad *et al.* (*27*). However, they used a less advanced GM and were only capable of synthesizing radiographs with limited quality, which are of less clinical value. We demonstrated the feasibility of using GANs as a tool of effective oversampling when the pathology distribution within a medical dataset was highly imbalanced. In particular, we demonstrated how a deep learning classifier can benefit from using synthesized x-rays from a publicly accessible database in cases where only few instances of a particular disease are present. Moreover, our developed GANs could be used to visualize what the generator neural network sees and to reveal correlations between diseases. The synthesis of pathological radiographs and the subsequent analysis by classification networks reveal correlations between diseases. For example, there seemed to be a block of diseases (lower right corner of Fig. 5D) that was characterized by lung opacification, namely, effusion, consolidation, pneumonia, infiltration, and edema. This makes sense from a clinical viewpoint as these diseases are clinically correlated, and similar observations hold for the remaining clusters in this figure. Cardiomegaly was an outlier in the sense that it was associated with seemingly only one disease from that block (effusion), but not the others. This again makes sense, considering that effusion can be a direct consequence of congestive heart failure. The correlation to edema was quite low for this case, which may indicate that the edema was not described consistently in the radiological reports and was therefore difficult for the neural network to synthesize and classify.

A potential problem in training a GAN on-site exists when the set of training images is limited—as is the case in small community hospitals. In those cases, the GAN might not converge, and realistically looking radiological images might not be produced. To overcome this problem, we proposed to use federated learning for training of the GAN: radiological images remained on-site and never left the premises, while only the weight updates got transferred to the central repository. We simulated this approach by splitting a set of 20,000 patients from the CheXpert dataset into 20 smaller datasets, and we found that the federally trained model significantly outperformed the locally constrained model.

One caveat when dealing with many smaller distributed databases is the potential low quality of locally trained GANs. To prevent the central repository from being contaminated by inferior synthetic x-rays, we propose two possible remedies: One way of pursuing the approach of pooling locally trained GANs is to apply a quality criterion such as the FID or the inception score (*28*) assessment. Locally trained generators can be rejected to be included in the central GAN repository. Second, federated learning allows for training of a single global GAN with several smaller distributed databases as demonstrated by Fig. 4A. In this way, several smaller databases can be combined and act as one large database without actually sharing the underlying patient information.

Attention needs to be paid to adversarial attacks on distributed learning systems. Models might be affected by poisoning attacks (*29*). Local gradients can be easily manipulated and distorted before being transmitted to central servers, and adversarial attacks might not be detected in the federated learning approach. Our GAN-based distributed learning approach offers passive and active robustness against adversarial attacks. The posterior distribution could be estimated (*30*), and the confidence threshold (*24*) of any given example in the local training set could be deduced. Such confidence thresholds could be used to detect and filter the suspicious training examples to secure GAN training from dataset poisoning (*29*). In addition, adversarial training (*31*) is an efficient method to increase model robustness against adversarial perturbations. We demonstrated (in fig. S3) that the robustness of our radiograph classifier was significantly improved by adversarial training.

The concepts demonstrated in this work rely on two-dimensional images, but there is no principal restriction on the number of dimensions that the real and synthesized images are allowed to have or even that the data have to consist of images only. Thus, the same concept could be translated to volumetric CT or magnetic resonance images, to fluoroscopy, to time series of volumetric data (e.g., contrast-enhanced CT), or even to imaging data in conjunction with clinical data (e.g., an MRI with associated expression profiles of laboratory tumor markers). However, because of the exponentially increasing size of the data, we expect that the problem of generating synthesized data of very high dimensionality is much more difficult and that a far greater number of real cases would be needed for the GAN to converge.

Diagnosis in the clinical setting usually relies on more than just imaging and comprises the patients' demographics, their medical history, and previous and ongoing treatments. Future work will investigate how to include these important parameters into our approach by letting the GANs generate not only radiographs but also accompanying clinical data such as laboratory values. However, to realize this, more training data are probably needed, as the data to be synthesized will have higher dimensions/degrees of freedom. Federated learning as presented here can help overcome those difficulties by providing the means to combine several distinct databases.

MATERIALS AND METHODS

Dataset and preprocessing

Three datasets were used in this study: first, the ChestX-ray dataset released by the NIH in 2017, containing 112,120 frontal radiographs of 30,805 unique patients. At the time of its publication, this dataset comprised 8 disease entities and was later updated to contain 14 pathologies (*32*). To ensure that no information leaked into the test set used for the evaluation of the algorithms, patient-wise stratification into training (21,528 patients, 78,468 radiographs, 70%), validation (3,090 patients, 11,219 radiographs, 10%), and test set (6187 patients, 22,433 radiographs, 20%) was performed. The test

set was kept separately until the final testing of the algorithms. Detailed label statistics for the ChestX-ray 14 dataset can be found in the "Preprocessing steps in CheXpert dataset" section in the Supplementary Materials and in table S3.

The second dataset used in this study is the CheXpert dataset, which has been released by Irvin *et al.* (*33*) in January 2019. It contains 224,316 chest radiographs of 65,240 patients. This dataset was used to train a second GAN to demonstrate the feasibility of the proposed data sharing approach (see Fig. 1). A detailed explanation of the label preparation and statistics for the CheXpert dataset is given in table S3. Algorithms of classification were tested on the NIH test set. Therefore, no subdivision of the CheXpert dataset into test, training, and validation sets was needed, and all available frontal radiographs of the CheXpert dataset (n = 191,027) were used for training of the GAN.

The third dataset used in this study is a dataset of x-rays released by the RSNA to host a challenge about pneumonia detection. We used this dataset to train a classifier for pneumonia detection and to test whether the inclusion of synthesized x-rays could improve the performance of said classifier.

Before training, radiograph datasets such as NIH and CheXperts were downsampled to dedicated spatial resolutions, i.e., ranging from 4×4, 8×8, …, $2^{10} \times 2^{10}$, and converted into separate files. Thus, each of those files contained all training radiographs with a fixed spatial resolution. The radiographs' intensity values were normalized to the range of [−1, 1] (*12*).

Model architecture and implementation

Two neural network architectures were used here. First, GANs as introduced by Goodfellow *et al.* (*11*) were adapted to incorporate an input condition (*19*) to selectively generate synthesized radiographs with a certain pathology. We used two different inputs to the networks: the conditional vector, which controls the type of disease present in the synthetic image, and the random noise vector, which determines which item from the set of possible x-rays is generated. Both vectors are concatenated and fed to the network as an input (*19*). As depicted in fig. S1C, such concatenation-based conditioning is equivalent to adding bias to the hidden activations based on the conditional input (*34*). In addition, we also added an auxiliary classifier at the end of the discriminator and additional classification loss terms in the objective

$$L_G^C = \mathbb{E}_{\tilde{x}\sim\mathbb{P}_g}[-\log P(C=c\,|\,\tilde{x})]$$
$$L_G^C = \mathbb{E}_{x\sim\mathbb{P}_r}[-\log P(C=c\,|\,x)] \quad (1)$$

where c is the pathological class label.

To generate high–spatial resolution images, we used progressive growing, a technique in which the GAN is trained in progressively higher–spatial resolution stages (*12*). The network architecture resulting in a final spatial resolution of 1024×1024 is shown in table S5. We picked leaky rectified linear unit (ReLU) (α = 0.2) and pixel norm (*12*) as the major activation function and normalization layer. Note that, instead of using a common tanh activation function, Karras *et al.* (*12*) suggested to use linear activation at the end of the generator. During training, we used a mini-batch size of 128 for spatial resolutions $4^2 - 32^2$ and then decreased the batch size by a factor of 2 when spatial resolution doubled to account for the limited memory size: $64 \times 64 \rightarrow 64$, $128 \times 128 \rightarrow 32$, $256 \times 256 \rightarrow 16$, $512 \times 512 \rightarrow 8$, and $1024 \times 1024 \rightarrow 4$. Dedicated explanations about techniques used in our GANs can be found at the "Network training details" section in the Supplementary Materials.

Second, a densely connected CNN with 121 layers (DenseNet-121) was used as a classifier. It was pretrained on 14 million natural images [ImageNet database (*2*)] and subsequently trained on the radiographs in this study. The architecture has been shown to achieve state-of-the-art performance on the ChestX-ray dataset (*35*) before. Implementations were done using TensorFlow 1.9.0 and PyTorch 1.1.0.

Training of the GANs

We trained two GANs on the basis of two separate datasets in a progressive growing strategy: on the NIH ChestX-ray14 dataset and the Stanford CheXpert dataset. Note that weights were initialized randomly. Training proceeded in repetitive stages: once training of one spatial resolution stage stabilized after being presented a total of 600,000 real radiographs (with repetitions), the layers responsible for the next spatial resolution stage were gradually faded in and training continued with another 600,000 radiographs during this fade-in stage (again with repetitions). In total, discriminators of GANs were each presented 12 million radiographs. The training scheme was chosen so that the GANs learned to first explore the large-scale pattern and overall contrast before focusing their attention on finer details.

To measure whether the images generated by the generator converged to real-looking images, we used the FID between a set of 10,000 real x-rays and 10,000 generated x-rays at each training epoch (*28*). We ensured an equal contribution from each pathological class by using a uniform distribution among the 14 classes, i.e., roughly 700 radiographs per class. To compute the FID, we extracted features of radiographs from the third pooling layer of the inception network (*28*). The FID score among real and synthesized radiographs was then computed according to

$$\mathrm{FID}(x,g) = \left\|\mu_x - \mu_g\right\|_2^2 + \mathrm{Tr}\left(\Sigma_x + \Sigma_g - 2\left(\Sigma_x\Sigma_g\right)^{\frac{1}{2}}\right) \quad (2)$$

where μ and Σ are mean and covariances of a multivariate Gaussian distribution that models the feature distributions. We found that the FID decreased nearly monotonically, indicating that the general appearance of the generated images approaches that of the real x-rays. The corresponding figures depicting the evolution of the FID are given in fig. S1D.

Training of the classifier with real and synthesized data

All classifier models used validation-based early stopping with sigmoid binary cross-entropy loss as the criterion. No oversampling of underrepresented classes was used except for the experiment, in which we specifically tested for the effect of oversampling. Training of the classifier network was done for a variety of different settings.

In the experiment depicted in Fig. 2, we first trained a classifier on a set of 1000 real x-rays (950 healthy, 50 exhibited signs of pneumonia) provided by the RSNA. Subsequently, we trained a classifier on a set of 500 real and 500 synthesized x-rays (450 healthy real, 50 pneumonia real, 400 healthy synthesized, and 100 pneumonia synthesized), whereby the synthesized radiographs were generated by generators that had previously been trained on the NIH and Stanford datasets. As a test set, we used a random subset of the dataset published by the RSNA, comprising 6000 real x-rays with the relation healthy:pneumonia as 2:3. In addition, we tested whether this concept could also be used in a more challenging task of differentiating

between a variety of diseases. As not all of the 14 pathologies labeled in the NIH dataset had been labeled in the Stanford dataset, we only classified those pathologies that were present in both datasets' labels, namely, cardiomegaly, effusion, pneumothorax, atelectasis, edema, consolidation, and pneumonia. We trained a classifier to differentiate between these classes with three different training sets: (i) synthesized x-rays generated by the NIH-GAN, (ii) synthesized x-rays generated by both the NIH-GAN and the Stanford-GAN, and (iii) real x-rays from the NIH dataset.

In addition, an experiment was carried out, in which the generated images were evaluated by the trained DenseNet to discover correlations between different pathologies. For each pairing of pathology as generated by the generator and pathology as classified by the classifier, we calculated Pearson's correlation coefficient and performed clustering on the resulting correlation matrix.

Federated averaging GAN

The pseudocode of our federated averaging GAN is given in algorithm S1. Specifically, we controlled our federated learning experiment by setting 10% of local clients that ran local GAN updates ($C = 10\%$), 10 local generator iterations on each round ($E = 10$), and a local batch size of 32 ($b = 32$). Following Gulrajani *et al.* (*26*), parameters of local Wasserstein GAN training was set to $\lambda = 10$ and $n_{\text{discriminator}} = 5$. All local models were initialized identically. During one global update round, as shown in Fig. 4A, a subset of clients (10% here) was picked to run local GAN updates on isolated datasets. Local clients were asked to transmit updated weights (red arrows in Fig. 4A) to the aggregation server once local updates were finished. The global model was updated by the weighted average over collected weights (*22*). To finish the global round, all local models were updated by the weights from the global model (blue arrows in Fig. 4A).

Reader study

Six readers were tasked with identifying whether a radiograph was real or synthesized. The tests were performed as follows. Each reader was given 30 s within which she or he had to decide whether the presented radiograph was real or synthesized. To prevent readers from identifying GAN-related features on the high–spatial resolution radiographs first—which are harder to produce and thus presumably more prone to artifacts—and transferring that knowledge to the low–spatial resolution images, the radiographs were presented in the following order: 100 radiographs of 256 × 256, 100 radiographs of 512 × 512, and, lastly, 100 radiographs of 1024 × 1024. All presented radiographs were different, i.e., the 256 × 256 were different from the 512 × 512 and 1024×1024 radiographs. Reading tests were done on a 24-inch computer monitor. To exclude the possibility that readers investigated the metal markers or pixel-hardcoded letters (e.g., denoting patient side—L or R) as potential artifacts to differentiate between real and synthesized images, these were covered by an independent investigator before handing out the x-ray to the testers.

Statistical analysis

For each of the experiments, we calculated the following parameters on the test set: AUC, accuracy, sensitivity, and specificity. To assess the errors due to sampling of the specific test set, we used bootstrapping with 10,000 redraws. The SE of the accuracy in the real versus synthesized tests for each human reader was calculated among the reader performances, and Fleiss' kappa was used to assess interreader agreement between readers.

To determine the number of needed samples for the performed experiments, we used power analyses according to (*36*). In general, all of our performed experiments followed a binomial distribution, because each decision for a radiograph was binary: either yes (e.g., was real for the case of deciding between real and synthesized radiograph or disease was present for the case of the classifiers) or no (was not real or disease not present). We could thus use the binomial formula for the SD of absolute numbers: $\text{SD}_{\text{absolute numbers}} = \sqrt{n \times p \times q}$, or equally well the SD of percentages: $\text{SD}_{\text{percentages}} = \sqrt{\frac{p \times q}{n}}$.

The difference of metrics, such as AUC, sensitivity, and specificity, was defined as a Δmetric (see table S4). For the total number of $n = 1000$ bootstrapping, models were built after randomly permuting predictions of two classifiers, and metric difference Δmetric_i were computed from their respective scores. We obtained the P value of individual metrics by counting all Δmetric_i above the threshold Δmetric. Statistical significance was defined as $P < 0.001$.

SUPPLEMENTARY MATERIALS

Supplementary material for this article is available at http://advances.sciencemag.org/cgi/content/full/6/49/eabb7973/DC1

View/request a protocol for this paper from *Bio-protocol*.

REFERENCES AND NOTES

1. S. Bickelhaupt, P. F. Jaeger, F. B. Laun, W. Lederer, H. Daniel, T. A. Kuder, L. Wuesthof, D. Paech, D. Bonekamp, A. Radbruch, S. Delorme, H.-P. Schlemmer, F. H. Steudle, K. H. Maier-Hein, Radiomics based on adapted diffusion kurtosis imaging helps to clarify most mammographic findings suspicious for cancer. *Radiology* **287**, 761–770 (2018).
2. A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks. *Commun. ACM* **60**, 84–90 (2017).
3. K. He, X. Zhang, S. Ren, J. Sun, Delving deep into rectifiers: surpassing human-level performance on ImageNet classification, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Santiago, Chile, 7 to 13 December 2015.
4. A. Hosny, C. Parmar, T. P. Coroller, P. Grossmann, R. Zeleznik, A. Kumar, J. Bussink, R. J. Gillies, R. H. Mak, H. J. W. L. Aerts, Deep learning for lung cancer prognostication: A retrospective multi-cohort radiomics study. *PLOS Med.* **15**, e1002711 (2018).
5. V. Gulshan, L. Peng, M. Coram, M. C. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, K. Widner, T. Madams, J. Cuadros, Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA* **316**, 2402–2410 (2016).
6. D. Mahajan, R. Girshick, V. Ramanathan, K. He, M. Paluri, Y. Li, A. Bharambe, L. van der Maaten, Exploring the limits of weakly supervised pretraining, in *Proceedings of the European Conference on Computer Vision (ECCV)*, Munich, Germany, 8 to 14 September 2018.
7. M. M. Mello, V. Lieou, S. N. Goodman, Clinical trial participants' views of the risks and benefits of data sharing. *N. Engl. J. Med.* **378**, 2202–2211 (2018).
8. L. M. Prevedello, S. S. Halabi, G. Shih, C. C. Wu, M. D. Kohli, F. H. Chokshi, B. J. Erickson, J. Kalpathy-Cramer, K. P. Andriole, A. E. Flanders, Challenges related to artificial intelligence research in medical imaging and the importance of image analysis competitions. *Radiol. Artif. Intell.* **1**, e180031 (2019).
9. J. Xu, F. Wang, Federated learning for healthcare informatics (2019); arXiv:1911.06270.
10. W. Li, F. Milletarì, D. Xu, N. Rieke, J. Hancox, W. Zhu, M. Baust, Y. Cheng, S. Ourselin, M. J. Cardoso, A. Feng, Privacy-preserving Federated Brain Tumour Segmentation, in *International Workshop on Machine Learning in Medical Imaging* (Springer, 2019), pp. 133–141.
11. I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in *Advances in Neural Information Processing Systems* (Neural Information Processing Systems Foundation, Inc., 2014), pp. 2672–2680.
12. T. Karras, T. Aila, S. Laine, J. Lehtinen, Progressive growing of GANs for improved quality, stability, and variation (2017); arXiv:1710.10196.
13. H.-C. Shin, N. A. Tenenholtz, J. K. Rogers, C. G. Schwarz, M. L. Senjem, J. L. Gunter, K. P. Andriole, M. Michalski, Medical image synthesis for data augmentation and anonymization using generative adversarial networks, in *International Workshop on Simulation and Synthesis in Medical Imaging* (Springer, 2018), pp. 1–11.
14. J.-Y. Zhu, T. Park, P. Isola, A. A. Efros, Unpaired image-to-image translation using cycle-consistent adversarial networks, *Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy, 22 to 29 October 2017.

15. J. M. Wolterink, A. M. Dinkla, M. H. F. Savenije, P. R. Seevinck, C. A. T. van den Berg, I. Išgum, Deep MR to CT synthesis using unpaired data, in *Simulation and Synthesis in Medical Imaging* (Springer, 2017), pp. 14–23.
16. A. Chartsias, T. Joyce, R. Dharmakumar, S. A. Tsaftaris, Adversarial image synthesis for unpaired multi-modal cardiac data, in *Simulation and Synthesis in Medical Imaging* (Springer, 2017), pp. 3–13.
17. Z. Zhang, L. Yang, Y. Zheng, Translating and segmenting multimodal medical volumes with cycle- and shape-consistency generative adversarial network, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, Utah, USA, 18 to 22 June 2018.
18. A. Radford, L. Metz, S. Chintala, Unsupervised representation learning with deep convolutional generative adversarial networks (2015); arXiv:1511.06434.
19. A. Odena, C. Olah, J. Shlens, Conditional image synthesis with auxiliary classifier GANs, *Proceedings of the 34th International Conference on Machine Learning-Volume 70* (JMLR.org, 2017), pp. 2642–2651.
20. P. Rui, K. Kang, National Hospital Ambulatory Medical Care Survey: 2015 Emergency Department Summary Tables (2015). https://www.cdc.gov/nchs/data/nhamcs/web_tables/2015_ed_web_tables.pdf [accessed 16 January 2020].
21. S. Conjeti, A. Katouzian, A. G. Roy, L. Peter, D. Sheet, S. Carlier, A. Laine, N. Navab, Supervised domain adaptation of decision forests: Transfer of models trained in vitro for in vivo intravascular ultrasound tissue characterization. *Med. Image Anal.* **32**, 1–17 (2016).
22. B. McMahan, E. Moore, D. Ramage, S. Hampson, B. A. y Arcas, Communication-Efficient Learning of Deep Networks from Decentralized Data, in *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, Fort Lauderdale, FL, USA, 20 to 22 April 2017.
23. A. N. Bhagoji, S. Chakraborty, P. Mittal, S. Calo, Analyzing federated learning through an adversarial lens. arXiv:1811.12470 (2018).
24. D. Stutz, M. Hein, B. Schiele, Confidence-calibrated adversarial training: Generalizing to unseen attacks. arXiv:1910.06259 (2019).
25. S. Augenstein, H. B. McMahan, D. Ramage, S. Ramaswamy, P. Kairouz, M. Chen, R. Mathews, B. A. y Arcas, Generative models for effective ML on private, decentralized datasets. arXiv:1911.06679 (2019).
26. I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, A. C. Courville, Improved training of Wasserstein GANs, in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett, Eds. (Curran Associates, Inc., 2017), pp. 5767–5777.
27. H. Salehinejad, S. Valaee, T. Dowdell, E. Colak, J. Barfett, Generalization of deep neural networks for chest pathology classification in X-rays using generative adversarial networks, in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2018), pp. 990–994.
28. M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, S. Hochreiter, GANs trained by a two time-scale update rule converge to a local nash equilibrium, in *Advances in Neural Information Processing Systems* (Curran Associates, Inc., 2017), pp. 6626–6637.
29. S. G. Finlayson, J. D. Bowers, J. Ito, J. L. Zittrain, A. L. Beam, I. S. Kohane, Adversarial attacks on medical machine learning. *Science* **363**, 1287–1289 (2019).
30. Y. Li, Y. Gal, Dropout inference in Bayesian neural networks with alpha-divergences, in *Proceedings of the 34th International Conference on Machine Learning-Volume 70* (JMLR.org, 2017), pp. 2052–2061.
31. A. Madry, A. Makelov, L. Schmidt, D. Tsipras, A. Vladu, Towards deep learning models resistant to adversarial attacks. arXiv:1706.06083 (2017).
32. X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, R. M. Summers, ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, 2017), pp. 2097–2106.
33. J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya, J. Seekins, D. A. Mong, S. S. Halabi, J. K. Sandberg, R. Jones, D. B. Larson, C. P. Langlotz, B. N. Patel, M. P. Lungren, A. Y. Ng, CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. arXiv:1901.07031 (2019).
34. E. Perez, F. Strub, H. de Vries, V. Dumoulin, A. Courville, FiLM: Visual reasoning with a general conditioning layer. arXiv:1709.07871 (2017).
35. P. Rajpurkar, J. Irvin, R. L. Ball, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. P. Langlotz, B. N. Patel, K. W. Yeom, K. Shpanskaya, F. G. Blankenberg, J. Seekins, T. J. Amrhein, D. A. Mong, S. S. Halabi, E. J. Zucker, A. Y. Ng, M. P. Lungren, Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists. *PLOS Med.* **15**, e1002686 (2018).
36. B. Rosner, *Fundamentals of Biostatistics* (Nelson Education, 2015).
37. M. Arjovsky, S. Chintala, L. Bottou, Wasserstein GAN. arXiv:1701.07875 (2017).
38. D. P. Kingma, P. Dhariwal, Glow: Generative flow with invertible 1x1 convolutions, in *Advances in Neural Information Processing Systems* (2018), pp. 10215–10224.
39. T. Salimans, A. Karpathy, X. Chen, D. P. Kingma, PixelCNN++: Improving the PixelCNN with discretized logistic mixture likelihood and other modifications. arXiv:1701.05517 (2017).
40. J. Hoffman, E. Tzeng, T. Park, J.-Y. Zhu, P. Isola, K. Saenko, A. A. Efros, T. Darrell, CyCADA: Cycle-consistent adversarial domain adaptation. arXiv:1711.03213 (2017).

Acknowledgments

Funding: This research project was supported by the START program of the Faculty of Medicine, RWTH Aachen, Germany, through the START rotation program granted to D.T. and by the DFG, Germany, through a grant given to S.N. **Author contributions:** T.H., D.T., V.S., and F.K. conceived the idea and approach. F.K., V.S., S.R., S.N., C.H., N.H., D.M., and D.T. contributed to the experiments. T.H., D.T., C.H., and N.H. developed the code infrastructure and GAN training setup. T.H., D.T., F.K., and V.S. wrote the manuscript. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** This study used three publicly available datasets: NIH ChestX-ray14 dataset (https://nihcc.app.box.com/v/ChestXray-NIHCC), Stanford CheXpert dataset (https://stanfordmlgroup.github.io/competitions/chexpert), and RSNA pneumonia dataset (https://kaggle.com/c/rsna-pneumonia-detection-challenge). The full images used in our real/synthesized radiograph test are available at https://drive.google.com/open?id=1_snb7hQ47WYxJEYK95G3cYlWSqckvRDW. Details of the implementation as well as the weights of the neural networks after training and the full code producing the results of this paper are made publicly available at https://github.com/peterhan91/Thorax_GAN.git. All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. Additional data related to this paper may be requested from the authors.

Submitted 19 March 2020
Accepted 14 October 2020
Published 2 December 2020
10.1126/sciadv.abb7973

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

Science Advances

Breaking medical data sharing boundaries by using synthesized radiographs

Tianyu Han, Sven Nebelung, Christoph Haarburger, Nicolas Horst, Sebastian Reinartz, Dorit Merhof, Fabian Kiessling, Volkmar Schulz, and Daniel Truhn

Sci. Adv. **6** (49), eabb7973. DOI: 10.1126/sciadv.abb7973