AMERICAN ACADEMY
OF OPHTHALMOLOGY®

Synthetic Medical Images for Robust, Privacy-Preserving Training of Artificial Intelligence

Application to Retinopathy of Prematurity Diagnosis

Aaron S. Coyner, PhD,[1] Jimmy S. Chen, MD,[1,2] Ken Chang, PhD,[3,4] Praveer Singh, PhD,[3,4] Susan Ostmo, MS,[1] R. V. Paul Chan, MD, MSc,[5] Michael F. Chiang, MD, MA,[6] Jayashree Kalpathy-Cramer, PhD,[3,4,] J. Peter Campbell, MD, MPH,[1,*] on behalf of the Imaging and Informatics in Retinopathy of Prematurity Consortium[†]*

Purpose: Developing robust artificial intelligence (AI) models for medical image analysis requires large quantities of diverse, well-chosen data that can prove challenging to collect because of privacy concerns, disease rarity, or diagnostic label quality. Collecting image-based datasets for retinopathy of prematurity (ROP), a potentially blinding disease, suffers from these challenges. Progressively growing generative adversarial networks (PGANs) may help, because they can synthesize highly realistic images that may increase both the size and diversity of medical datasets.

Design: Diagnostic validation study of convolutional neural networks (CNNs) for plus disease detection, a component of severe ROP, using synthetic data.

Participants: Five thousand eight hundred forty-two retinal fundus images (RFIs) collected from 963 preterm infants.

Methods: Retinal vessel maps (RVMs) were segmented from RFIs. PGANs were trained to synthesize RVMs with normal, pre-plus, or plus disease vasculature. Convolutional neural networks were trained, using real or synthetic RVMs, to detect plus disease from 2 real RVM test datasets.

Main Outcome Measures: Features of real and synthetic RVMs were evaluated using uniform manifold approximation and projection (UMAP). Similarities were evaluated at the dataset and feature level using Fréchet inception distance and Euclidean distance, respectively. CNN performance was assessed via area under the receiver operating characteristic curve (AUC); AUCs were compared via bootstrapping and Delong's test for correlated receiver operating characteristic curves. Confusion matrices were compared using McNemar's chi-square test and Cohen's κ value.

Results: The CNN trained on synthetic RVMs showed a significantly higher AUC (0.971; $P = 0.006$ and $P = 0.004$) and classified plus disease more similarly to a set of 8 international experts ($\kappa = 0.922$) than the CNN trained on real RVMs (AUC = 0.934; $\kappa = 0.701$). Real and synthetic RVMs overlapped, by plus disease diagnosis, on the UMAP manifold, showing that synthetic images spanned the disease severity spectrum. Fréchet inception distance and Euclidean distances suggested that real and synthetic RVMs were more dissimilar to one another than real RVMs were to one another, further suggesting that synthetic RVMs were distinct from the training data with respect to privacy considerations.

Conclusions: Synthetic datasets may be useful for training robust medical AI models. Furthermore, PGANs may be able to synthesize realistic data for use without protected health information concerns. *Ophthalmology Science 2022;2:100126 © 2022 by the American Academy of Ophthalmology. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).*

Supplemental material available at www.ophthalmologyscience.org.

Artificial intelligence (AI), specifically deep learning (DL), for automated image-based disease detection and segmentation has gained significant traction over the last decade.[1–8] Convolutional neural networks (CNNs), a subset of DL techniques, have demonstrated clinician-level efficacy in pattern recognition in many medical research applications; however, a number of barriers exist to development and implementation.[1–6] Institutional regulations can limit access to patients' medical imaging data for privacy reasons.[9–14] Rare diseases can be challenging to acquire enough data

https://doi.org/10.1016/j.xops.2022.100126
ISSN 2666-9145/22

even without institutional barriers.[9–12] Discrepancies may exist in the label quality or consistency of annotators, leading to an increased training data requirement.[15] Finally, models trained in one population may not generalize well to external populations that differ from the training population. One disease that suffers from all of these challenges is retinopathy of prematurity (ROP), a potentially blinding disease that affects prematurely born infants.[12,16]

Previously, using thousands of retinal fundus images (RFIs) collected by the Imaging and Informatics in ROP multicenter study over the course of many years, we trained and evaluated the ability of CNNs to diagnose plus disease—a strong indicator of treatment-requiring ROP, defined as venous dilation and arterial tortuosity—and found that they could achieve clinician-level performance.[1] However, it is likely that the development process could have been expedited had other sites been willing to share data, because CNN training efficiency and generalizability are not only related to the ease of a pattern recognition task, but also the size and diversity of the training data.[9,11,12] For optimal generalizability, the bigger and more diverse a dataset, the better. One way to address this issue is to simulate new data. Generative adversarial networks (GANs), another subset of DL techniques, could be used for this task.[17,18] As is implied in the name, GANs are generative models which aim to synthesize new observations of data based off of supplied labels, as opposed to classification models which aim to label observations based off of already existing data. Generative adversarial networks have an uncanny ability to synthesize highly realistic images, evidenced by our prior study demonstrating that medical experts have a difficult time discerning between real and synthetic RFIs.[19]

We hypothesize that progressively growing GANs (PGANs) can be trained to generate an infinitely large synthetic dataset of images representing a spectrum of vascular severities, which can then be used to train CNNs for plus disease detection.[18] If CNNs trained on synthetic images are able to detect plus disease from real images as well as CNNs trained on real images, it stands to reason that synthetic medical datasets may be a viable option for artificially enhancing small datasets. Furthermore, if it can be shown that synthesized images are not similar to those on which GANs were trained, then the risk of inadvertently revealing protected health information in synthetic image datasets is low, supporting their dissemination as a privacy-preserving alternative to data sharing. Ultimately, GANs may allow researchers to increase the size and diversity of medical datasets and to share those datasets with other researchers without privacy concerns.

Methods

Imaging and Informatics in Retinopathy of Prematurity Dataset

The Imaging and Informatics in Retinopathy of Prematurity study was approved by the institutional review boards at the coordinating center (Oregon Health & Science University) and at each of 7 study centers (Columbia University, University of Illinois at Chicago, William Beaumont Hospital, Children's Hospital Los Angeles, Cedars-Sinai Medical Center, University of Miami, and Weill Cornell Medical Center) and was conducted in accordance with the tenets of the Declaration of Helsinki.[20] Written informed consent was obtained from parents of all enrolled infants.

Between July 2011 and December 2016, a RetCam (Natus) was used during routine ROP screenings to collect posterior pole-centered RFIs from premature infants born weighing < 1501 g or born before 31 weeks of gestation. A reference standard diagnosis was assigned to each RFI using previously published methods based on independent image-based diagnoses by at least 3 trained graders (2 ROP specialists and 1 ROP study coordinator; S.O., R.V.P.C., M.F.C., J.P.C.) and the clinical diagnoses (dilated, in-person, ophthalmoscopic evaluation) by expert ophthalmologists.[21] The RFIs were classified as having normal, preplus, or plus disease vasculature. Those not centered on the posterior pole, labeled as "not of acceptable quality for diagnosis" by most image graders, or of stage 4 or 5 ROP were excluded (in stages 4 and 5 ROP, the retinal blood vessels prove difficult to visualize because of partial or total retinal detachments, respectively, and the diagnosis of plus disease is not relevant).[16]All but 100 images were split at the patient level into training, validation, and test datasets at a ratio of 3:1:1, while maintaining the natural disease prevalence. The remaining 100 images were placed into an expert test dataset that was specifically designed to have a higher prevalence of plus disease images to better capture the incidence of ROP worldwide and was designated for use as the final dataset for evaluating model performance.[1,22,23] Images in this dataset were graded for plus disease by a set of 8 international ROP experts (M.F.C., R.V.P.C.), each with more than 10 years of clinical ROP experience and more than 5 peer-reviewed articles on ROP published. Five of the 8 experts served as principal investigators for the multicenter Early Treatment for ROP study.[24]

Our previous study used retinal vessel maps (RVMs)—grayscale segmentations of the major retinal arteries and veins in RFIs—rather than RFIs to train CNNs to detect plus disease because they strictly focus DL techniques on evaluating the retinal vasculature, rather than other clinical abnormalities that may be present in RFIs.[1] To produce RVMs, a previously trained u-net was used, which takes 640 × 480 color RFIs as input and outputs single channel 640 × 480 RVMs, where each pixel intensity value represents the overall probability of that pixel belonging to a major artery or vein.[1,25] All pixel intensity values that represented a < 20% probability of said pixels indicating the presence of major retinal blood vessels (pixel intensity values, < 51) were set equal to 0, because most of these pixels encode information related to choroidal blood vessel patterns, which are not relevant for plus disease diagnosis.[16,19] Finally, RVMs were resized to 256 × 256, because that was the size of the images used to pretrain the ResNet-18 CNN (discussed below).

Generative Adversarial Network Development

Progressively growing GAN code was acquired from Facebook Research's publicly available GitHub repository.[18,26] This GAN was used, as opposed to a more state-of-the-art GAN, because the desired outputs were black-and-white 256 × 256 images as opposed to larger, more complex color images. Three PGANs were trained to synthesize 256 × 256 de novo RVMs by learning the data distribution of real RVMs in the training dataset and, specifically, learning how to generate RVMs representing normal, preplus, or plus disease vasculatures. Models were trained at 7 progressively grown image sizes (4 × 4, 8 × 8, 16 × 16, 32 × 32, 64 × 64, 128 × 128, and 256 × 256). All parameters were set at the default values reported by Karras et al.[18]

Table 1. Partitions of the Informatics in Retinopathy of Prematurity Dataset

Dataset	No. of Patients	No. of Images	Normal (%)	Preplus (%)	Plus (%)
Training	534	3477	2908 (83.6)	464 (13.4)	105 (3.0)
Validation	178	1167	903 (77.4)	222 (19.0)	42 (3.6)
Test	179	1098	927 (84.4)	136 (12.4)	35 (3.2)
Expert test	72	100	54 (54.0)	31 (31.0)	15 (15.0)
Total	963	5842	4792 (82.0)	853 (14.6)	197 (3.4)

Unlike discriminative CNNs, monitoring loss statistics of GANs during training does not provide information regarding the ability of the GANs to synthesize images, and it is best to continue training even if loss metrics have begun to diverge.[17,18] After training, a synthetic dataset consisting of the same number of images and disease distribution as the training dataset was generated using the PGANs, and a subset of images were evaluated visually and subjectively to ensure that images represented grayscale RVMs with medically plausible vascular trees that represented the desired vascular severities (normal, preplus, or plus).

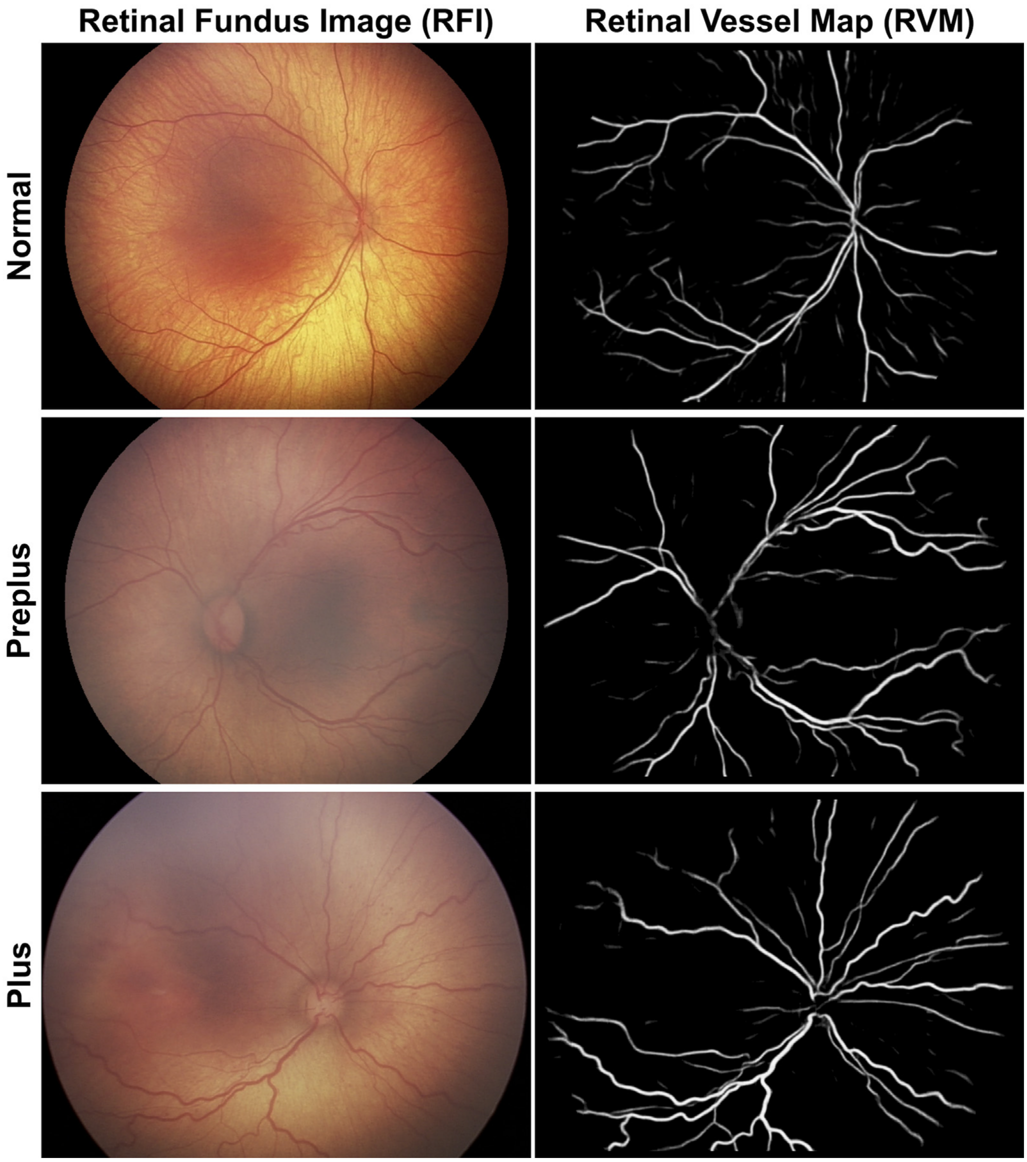

Figure 1. The main arteries and veins present in color retinal fundus images (RFIs; left column) can be automatically segmented into grayscale retinal vessel maps (RVMs; right column) using a previously trained u-net.

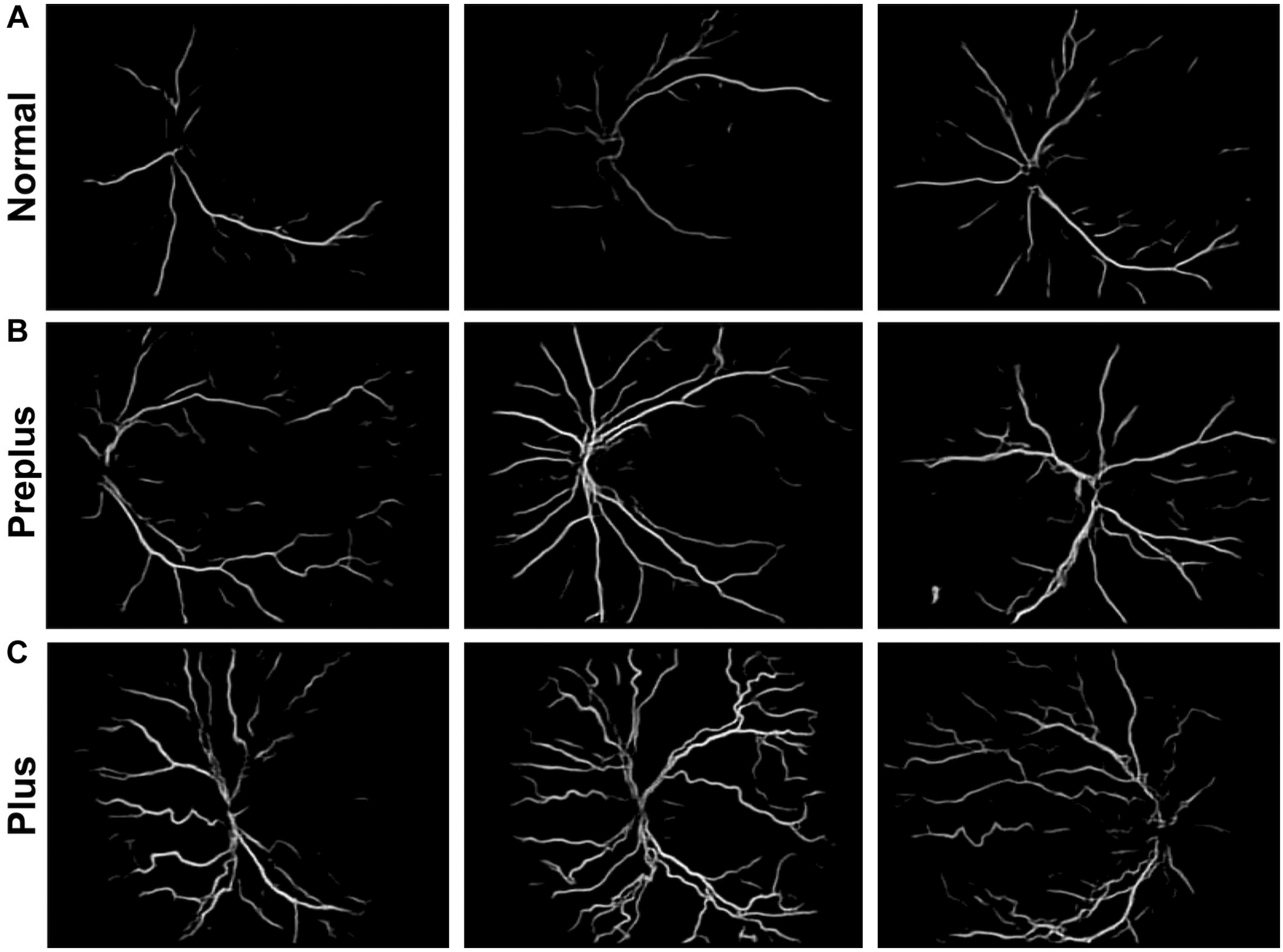

Figure 2. Synthetic retinal vessel maps (RVMs) of eyes with (**A**) normal retinal vasculature, (**B**) preplus disease, or (**C**) plus disease can be generated by progressively growing generative adversarial networks trained on a limited number of real RVMs.

Feature Space Analysis

To confirm, beyond visual inspection, that synthetic RVMs represented real RVMs, features of real RVMs from the training dataset were extracted from the penultimate layer of a CNN trained to diagnose plus disease (as discussed below) from those same RVMs. This layer output 512 features to the last layer of the CNN, which was responsible for predicting normal, preplus, or plus disease using said features. Uniform manifold approximation and projection (UMAP) was used to reduce these features into a 2-dimensional space (manifold). The Euclidean distance between the 25 nearest neighbors for each point, with a minimum distance of 0.99, was used to develop the manifold. All other parameters remained at the default values reported by McInnes et al.[27] The features of the synthetic RVMs were then extracted from the same CNN, and their place on the manifold, given these features, was predicted.

Similarity Analysis

Although demonstrating that real and synthetic RVMs of matching disease severity overlap in a dimensionally reduced feature space suggests that synthetic RVMs are similar to real RVMs and that the desired disease severities are being produced, it does not inform whether the RVMs produced by the PGANs are unique or simply reproductions or slightly modified versions of the RVMs from which they were trained. Therefore, to evaluate image similarity, the Fréchet inception distance (FID)—a measure of similarity between two datasets of images, where the minimum score of 0 indicates that 2 image datasets are identical—was computed between real and synthetic RVMs.[28] This metric compares datasets by fitting 2 Gaussians to the extracted features of RVMs from each dataset output by a ImageNet-pretrained InceptionV3 CNN, and the FID between the two is calculated.[29,30] To get an expected FID value, the FID between real RVMs in the training dataset and real RVMs in the combined validation and test datasets were compared. This essentially tested how similar RVMs are between different people. The FID was then used to compare real RVMs in the training dataset with synthetic RVMs.

However, FID is a dataset-level measure of similarity. To investigate at the image level, the Euclidean distances of images to one another, determined using the features output by the same InceptionV3 network used to compute FID, were determined. Specifically, the closest pair of RVMs in the (1) training dataset alone, (2) training dataset versus combined validation and test dataset, or (3) synthetic dataset versus training dataset were evaluated. These comparisons allowed us to test the following, respectively: (1) the shortest Euclidean distance between RVMs

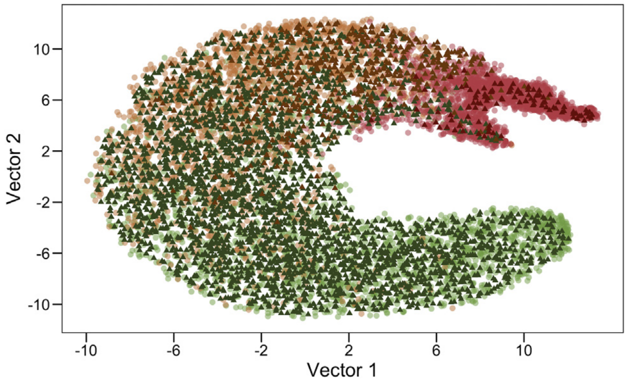

Figure 3. Two-dimensional manifold generated by uniform manifold approximation and projection (UMAP) using the extracted image features of real retinal vessel maps from a convolutional neural network trained to diagnose plus disease using said retinal vessel maps (opaque triangles). Features of synthetic retinal vessel maps (transparent circles) were extracted from the same model and their locations on the UMAP manifold were predicted. Real and synthetic normal (green), preplus (orange), and plus disease (red) retinal vessel maps, respectively, overlapped with one another on the UMAP manifold.

from the same participant, (2) the shortest Euclidean distance between RVMs from different participants, and (3) the shortest Euclidean distance between synthetic RVMs and the real RVMs on which the PGANs were trained.

Convolutional Neural Network Development

A SoftMax layer, which converts raw CNN predictions into a probability distribution for all classes, was appended to 2 ResNet-18 CNNs, pretrained on the ImageNet database.[29,31,32] These CNNs were fine tuned on real RVMs or synthetic RVMs produced by PGANs. During training, the CNN trained on real RVMs was validated using real RVMs, and the CNN trained on synthetic RVMs was validated using a synthetic dataset with the same size and same disease distribution as the validation dataset. Both models were tested on real RVMs from 2 test datasets.

During training, RVMs were input to CNNs at a resolution of 224 × 224. Random rotations (−45° to 45°) and horizontal or vertical flips, or both, were applied to the crops, each with a probability of occurring equal to 0.5. Pixel values were standardized with mean of 0 and standard deviation of 1. For validation and testing, images were input at a resolution of 224 × 224 and only image standardization was applied. To train the real RVM CNN using the natural disease prevalence, a weighted random sampler was used. The Adam optimizer was used at a learning rate of 0.001 with a batch size of 8.[33] Models were trained for up to 50 epochs; however, early stopping was implemented (with patience of 10 epochs) and only epochs that resulted in progressively better accuracy on the validation dataset were saved.

Convolutional neural network performances were evaluated using area under the receiver operating characteristic curve (AUC).[32] Specifically, discrimination between the binary outcome of plus disease versus normal or preplus disease, as compared with the reference standard diagnosis, was evaluated. To evaluate differences between respective AUCs, CNN predictions were bootstrapped 1000 times to compute sample means and 95% confidence intervals.[34] Significance was determined if the bootstrapped mean of the AUC of the CNN trained on synthetic RVMs fell outside of the 95% confidence interval of the bootstrapped mean of the AUC of the CNN trained on real RVMs. In addition, Delong's test for correlated receiver operating characteristic curves was used to verify these findings (significance determined if $P < 0.05$). Finally, model performances were evaluated on the expert test dataset. Confusion matrices were statistically evaluated for similarity using McNemar's chi-square test and Cohen's κ value (interpreted using a commonly accepted scale: 0−0.20, slight agreement; 0.21−0.40, fair agreement; 0.41−0.60, moderate agreement; 0.61−0.80, substantial agreement; and 0.81−1.0, near-perfect agreement).[35]

Results

Imaging and Informatics in Retinopathy of Prematurity Dataset

Retinal fundus images were partitioned into training, validation, test, and expert test datasets (Table 1), which contained similar distributions of normal, preplus, and plus disease RFIs. The expert test dataset was additionally enriched with plus disease images. Color RFIs in each dataset were segmented into grayscale RVMs (Fig 1).

Generative Adversarial Network Development

Throughout the GAN training process, synthetic RVMs were evaluated visually. After subjective review, we found that generated images: (1) appeared similarly to RVMs, (2) showed medically plausible vascular trees, and (3) showed dilation and tortuosity consistent with normal, preplus, and plus disease (Fig 2).

Feature Space Analysis

Uniform manifold approximation and projection analysis suggested that real and synthetic RVMs had similar features because they overlapped with one another on the dimensionally reduced UMAP manifold (Fig 3). Furthermore, synthetic normal, preplus, and plus disease RVMs overlapped with real RVMs of the same diagnoses. Taken together, these results suggest that real and synthetic RVMs not only contain the same general features, but also disease-specific features. Furthermore, this result shows that synthetic RVMs span the entire disease spectrum represented by real RVMs, because real RVM features were encapsulated entirely within the space of synthetic RVM features.

Similarity Analysis

The FID for real RVMs from the training dataset compared with real RVMs from the combined validation and test datasets was 4.64, whereas the FID for synthetic RVMs compared with the real RVMs from the training dataset was 6.53. This finding suggests that more similarity exists between real RVMs from different groups of participants than exists between synthetic RVMs and the real RVMs on which they were trained. Furthermore, the pair of RVMs closest in Euclidean distance (5.54) were from the same participant's eye, and the closest pair of RVMs from different participants showed a Euclidean distance of 5.71 (Fig 4). However, the closest pair of a synthetic RVM and a

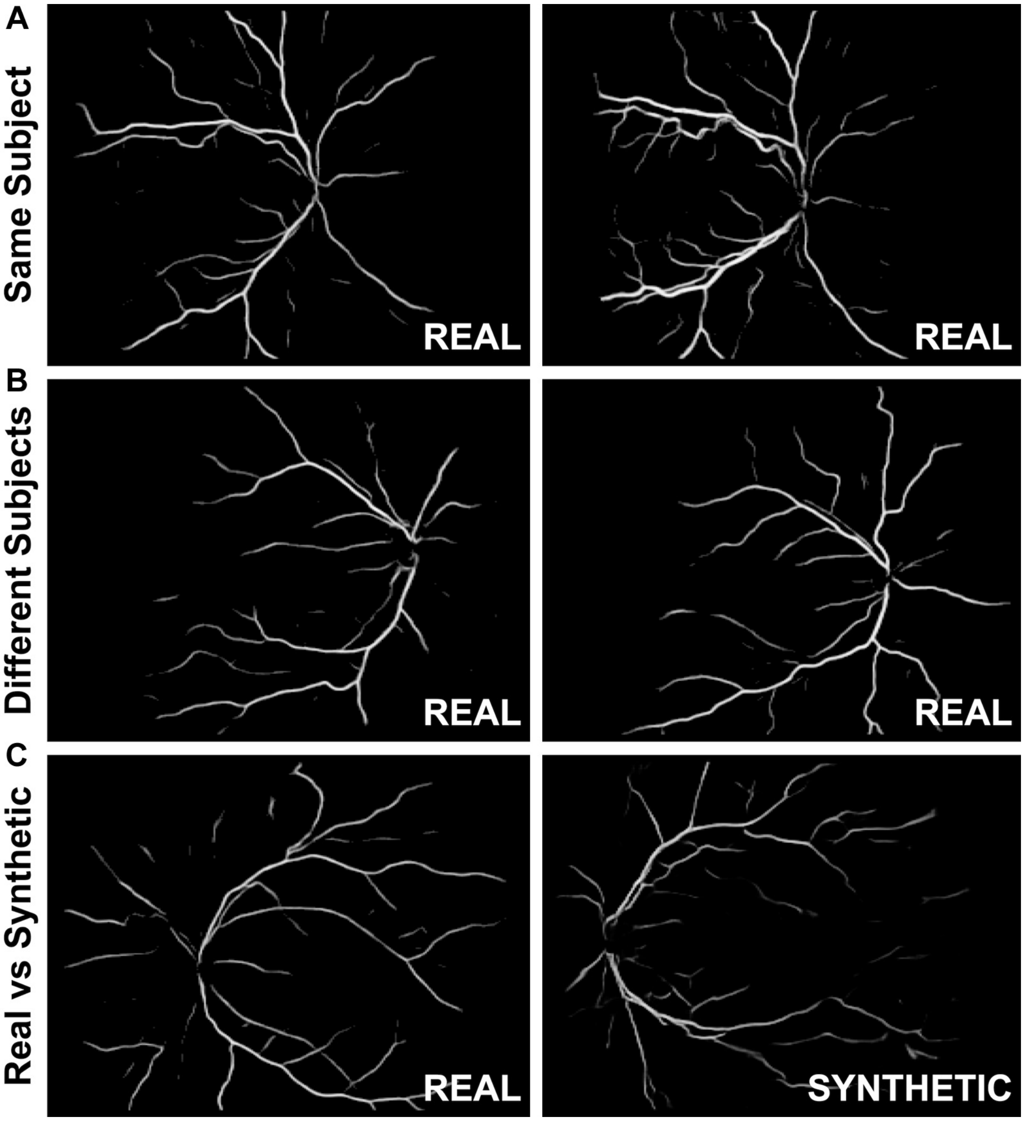

Figure 4. Pairs of real retinal vessel maps (RVMs) are closer in feature space than pairs of real and synthetic RVMs. Using the same features output by InceptionV3 to compute Fréchet inception distance, the Euclidean distances between RVMs from the same participants, different participants, and synthetic RVMs and the real RVMs they were trained on were calculated. **A**, The closest Euclidean distance was 5.54 and occurred between RVMs from the same participant's eye. **B**, The closest RVMs from different participants showed a distance of 5.71. **C**, The closest synthetic RVM to a real RVM from the training dataset showed a Euclidean distance of 6.15.

real RVM from the training dataset showed a Euclidean distance of 6.15. These results suggest that real RVMs from different participants are more similar to one another than any synthetic RVMs are to real RVMs from the training dataset. That is, the PGAN is not simply reproducing real images that could be identifiable.

Convolutional Neural Network Evaluation

The ability of CNNs trained on real or synthetic RVMs to discriminate between plus disease and normal or preplus disease was evaluated via receiver operating characteristic curves on the test dataset (Fig 5). The bootstrapped mean AUC (0.934) for the CNN trained on real RVMs was significantly different from the bootstrapped mean AUC (0.971) of the CNN trained on synthetic RVMs ($P = 0.004$). Delong's test for correlated receiver operating characteristic curves confirmed significance ($P = 0.006$). Taken together, these findings suggest that, for detection of plus disease, models trained on synthetic RVMs are at least as performant as, if not superior to, those trained on real RVMs.

Finally, both models' predictions of 100 RVMs were compared with the consensus diagnoses of 8 international experts (Table 2). McNemar's chi-square test suggested that predictions made by the CNN trained on real RVMs ($P = 0.450$) and those made by the CNN trained on synthetic RVMs ($P > 1.000$) were not significantly different from the diagnoses of 8 international experts. However, although Cohen's κ value

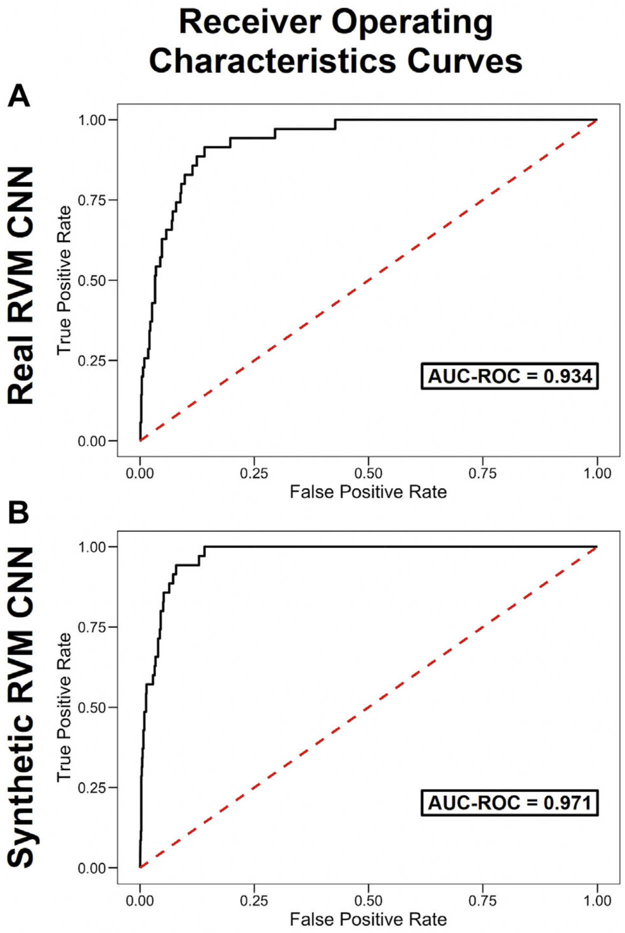

Figure 5. Models trained on synthetic retinal vessel maps (RVMs) detect plus disease better than those trained on real RVMs. Receiver operating characteristic curves (ROCs) of convolutional neural networks (CNNs) trained on (**A**) real or (**B**) synthetic RVMs for detection of plus disease versus normal or preplus disease from real RVMs in the test dataset are depicted. Areas under the ROC curves (AUCs) were significantly different as determined by 2 different tests ($P = 0.004$, $P = 0.006$).

suggested moderate agreement (κ = 0.701) between expert diagnoses and predictions of the model trained on synthetic RVMs, the model trained on synthetic RVMs showed near-perfect agreement (κ = 0.922) with experts' diagnoses.

Discussion

In this work, we trained multiple PGANs to generate RVMs depicting normal, preplus, or plus disease vasculature. We then trained 2 ResNet-18 CNNs, one on real RVMs and the other on synthetic RVMs, to diagnose plus disease. The 3 key findings are: (1) PGANs can synthesize realistic RVMs that accurately represent the desired disease state, (2) synthesized RVMs are less similar to the RVMs they were trained on than real RVMs are to one another, and (3) CNNs trained on synthetic RVMs can detect plus disease at least as well as, if not better than, those trained on real RVMs.

Plus disease, by definition, is a disease of the main retinal blood vessels.[16] For this reason, the main arteries and veins present in color RFIs were segmented into grayscale RVMs using a previously trained u-net (Fig 1).[1] These RVMs were then used to train PGANs to generate RVMs with normal, preplus, or plus disease vasculature. In general, RVMs produced by these GANS were highly realistic, especially when compared visually with real RVMs (Figs 1 and 2). Uniform manifold approximation and projection visualization confirmed that real and synthetic RVMs were contained within the same feature space, meaning that they shared similar features (Fig 3). Furthermore, synthetic and real RVMs overlapped by disease severity (i.e., normal, preplus, and plus), suggesting that not only were PGANs generating images that had similar features to real RVMs, but also that each PGAN was generating features unique to normal, preplus, or plus disease vasculatures and that synthetic RVMs spanned the entire spectrum of disease represented within real data. Taken together, these results suggested that synthetic RVMs accurately depicted their real RVM counterparts and could be useful for training CNNs to detect plus disease.

Table 2. A Model Trained on Synthetic Retinal Vessel Maps Predicts Plus Disease Diagnoses More Similarly to International Experts Than a Model Trained on Real Retinal Vessel Maps

	Ground Truth	
	Not Plus	*Plus*
Real RVMs		
Not Plus	83	5
Plus	2	10
Synthetic RVMs		
Not Plus	84	1
Plus	1	14

RVM = retinal vessel map.

However, the high degree of overlap between real and synthetic normal, preplus, and plus disease RVMs in the visualized UMAP embedding provoked the question of whether the PGANs were simply replicating the RVMs they had been trained on. To investigate this possibility, the FID between real and synthetic RVMs was assessed. First, to obtain an expected or baseline FID value, participants in the training dataset were compared with participants in the combined validation and test datasets (FID = 4.64). Retinal vessel maps from participants in the training dataset were then compared with synthetic RVMs (FID = 6.53). Synthetic RVMs were more dissimilar (higher FID scores) from the training RVMs than the training RVMs were dissimilar to real RVMs from other participants. That is, synthetic RVMs are not just copies of the training data. In fact, their diversity seems to supersede the natural diversity found between humans at the population level in our dataset.

We performed further analysis on the Euclidean distances of CNN output features between all RVMs in the training dataset compared with one another, RVMs in the training dataset compared with RVMs in the combined validation and test datasets, and synthetic RVMs compared with training dataset RVMs. The pair of RVMs with the closest Euclidean distance belonged to the same participant (Fig 4).

However, the closest real and synthetic RVMs were further apart from one another than the closest real RVMs that belonged to different participants. This finding, once again, suggests that more dissimilarity exists between real and synthetic RVMs than exists between real RVMs from different participants. Together, UMAP, FID, and Euclidean distance analyses suggest that: (1) the PGANs are not simply replicating the RVMs on which they were trained, meaning that (2) the synthesized RVMs likely contain less protected health information than real RVMs, but (3) they still have features relevant not only to RVMs, but specifically to normal, preplus, and plus diseases with (4) a diversity that supersedes that of our dataset of real RVMs.

Given these findings, real or synthetic RVMs were used to train CNNs to detect real normal, preplus, or plus disease RVMs. The CNN trained on synthetic RVMs was significantly better able to detect plus disease than the CNN trained on real RVMs, as evidenced by their respective AUCs (Fig 5). Both models were ultimately evaluated on the expert test set, which was graded by a set of 8 international experts. The model trained on synthetic RVMs outperformed the model trained on real RVMs, as evidenced by the Cohen's κ value. Given that the real and synthetic RVM datasets contained the same number of normal, preplus, and plus disease images, it is reasonable to assume that the performance increase is the result of there being more diversity among synthetic RVMs as compared with real RVMs (supported by the higher FID score between synthetic and real RVM datasets than between real RVM datasets). That is, although the GANs learned to generate RVMs from the same dataset that a CNN was trained on to diagnose plus disease, the diversity of the synthetic data they were able to generate may have increased the performance of the CNN trained on synthetic RVMs as compared with the CNN trained on real RVMs.

As shown above, using synthetic data has at least 2 potential benefits. First, PGAN synthesizes de novo RVMs. In theory, this means that GANs, which capture underlying disease characteristics and data distributions, could be used to generate synthetic medical images that can be shared publicly with a lower degree of risk of identifying unique patients. Although RFIs and RVMs currently are difficult to link to individual patients without a reference scan to which they can be compared, it is still theoretically possible.[36] Using GANs could allow for entirely synthetic datasets to be created and shared, so that broader teams of researchers can investigate and train models on existing data without privacy concerns. Second, synthetic images may contain more variations in their patterns and textures than real images, which may serve as a viable data augmentation technique for increasing robustness of CNNs for disease classification. In addition to increasing the prevalence of rare diseases, this may be especially useful in cases where one wishes to increase the proportion of demographic minorities within a dataset. For example, GANs have been used to alter the retinal pigmentation present in RFIs to better ameliorate racial imbalances in datasets.[37]

Still, because we did not use any privacy-preserving techniques during GAN training, theoretical concerns remain about whether synthetic images are truly de-identifed.[38] Although FID suggests that real and synthetic RVMs are not similar at the dataset level, and Euclidean distances between RVMs suggest that real RVM features were more similar to one another than to synthetic RVM features at the image level; further image-level analyses should be performed because the consequences of being wrong could be detrimental. However, this proves more difficult than initially thought. For example, one could compute the structural similarity index measure or peak signal-to-noise ratio between image pairs to test for similarity.[39] However, if a GAN were to output training dataset images that were slightly warped or rotated, then structural similarity index measure and peak signal-to-noise ratio would suggest that the images were different, even though they were actually the same image. This is why FID and Euclidean distance computed over Inceptionv3 features were used, because they are more invariant to flips, rotations, and so forth (because of data augmentation techniques used during training) and have been shown to correlate well with human judgement of visual quality and diversity.[28,40] Future work will investigate ways to further validate these findings, for example, by building siamese networks for identifying whether 2 images are from the same patient.[41]

In conclusion, these findings suggest that GAN-generated medical images may be just as powerful for training CNNs for disease diagnosis and may even increase their robustness. Ultimately, this may be broadly relevant to other medical domains, where datasets are often limited and disseminating datasets containing patient-identifiable images or proprietary information is of concern. Future work will center around exploration of similarity metrics, as well as applications to other imaging methods and diseases within and outside of ophthalmology. Adopting this technology could not only improve medical CNN diagnoses, but could also allow for legal, ethical sharing of medical datasets.

Footnotes and Disclosures

Originally received: November 24, 2021.
Final revision: February 1, 2022.
Accepted: February 7, 2022.
Available online: February 11, 2022. Manuscript no. XOPS-D-21-00221

[1] Department of Ophthalmology, Casey Eye Institute, Oregon Health & Science University, Portland, Oregon.

[2] Department of Ophthalmology, Shiley Eye Institute, University of California, San Diego, San Diego, California.

[3] Department of Radiology, Athinoula A. Martinos Center for Biomedical Imaging, Massachusetts General Hospital, Charlestown, Massachusetts.

[4] Center for Clinical Data Science, Massachusetts General Hospital and Boston Women's Hospital, Boston, Massachusetts.

[5] Department of Ophthalmology and Visual Sciences, Eye and Ear Infirmary, University of Illinois, Chicago, Illinois.

[6] National Eye Institute, National Institutes of Health, Bethesda, Maryland.

*Both authors contributed equally as senior authors.

[†]Members of the i-ROP research consortium appear in the Appendix (www.ophthalmologyscience.org).

Disclosure(s):

All authors have completed and submitted the ICMJE disclosures form.

The author(s) have made the following disclosure(s): R.V.P.C.: Scientific advisory board – Phoenix Technology Group; Consultant – Alcon; Financial support – Genentech; Equity owner – Siloam Vision

M.F.C.: Previous Consultant – Novartis; Previous Financial support – Genentech; Previous Equity owner – InTeleretina, LLC

J.K.-C.: Financial support – Genentech

J.P.C.: Consultant – Boston AI Lab; Financial support – Genentech; Equity owner – Siloam Vision

The i-ROP DL system has been licensed to Boston AI Lab (Boston, MA) by Oregon Health & Science University, Massachusetts General Hospital, Northeastern University, and the University of Illinois, Chicago, which may result in royalties to Drs Chan, Campbell, Brown, and Kalpathy-Cramer in the future.

Supported by the National Institutes of Health, Bethesda, Maryland (grant nos.: R01 EY19474, R01 EY031331, R21 EY031883, and P30 EY10572); and Research to Prevent Blindness, Inc., New York, New York (unrestricted departmental funding and a Career Development Award [J.P.C.]). These funding organizations had no role in the design or conduct of this research.

Because of institutional review board and data privacy considerations, real retinal fundus images, retinal vessel maps, and associated clinical variables will not be made publicly available. However, the synthetic retinal vessel maps used in these experiments may be found at https://doi.org/10.17632/fscyyhg6vt.1, and all code may be found at: https://github.com/aaroncoyner/synthetic-plus-classifier.

HUMAN SUBJECTS: Human subjects were included in this study. The Institutional Review Boards at the coordinating center (Oregon Health & Science University) and at each of seven study centers (Columbia University, University of Illinois at Chicago, William Beaumont Hospital, Children's Hospital Los Angeles, Cedars-Sinai Medical Center, University of Miami, Weill Cornell Medical Center) and was conducted in accordance with the Declaration of Helsinki. Written, informed consent was obtained from parents of all enrolled infants.

No animal subjects were included in this study.

Author Contributions:

Conception and design: Coyner, Chen, Chiang, Kalpathy-Cramer, Campbell

Analysis and interpretation: Coyner, Chen, Chang, Singh, Ostmo, Chan, Chiang, Kalpathy-Cramer, Campbell

Data collection: Coyner, Chen, Ostmo, Chan, Chiang, Kalpathy-Cramer, Campbell

Obtained funding: Chan, Chiang, Kalpathy-Cramer, Campbell

Overall responsibility: Coyner, Chen, Chang, Singh, Ostmo, Chan, Chiang, Kalpathy-Cramer, Campbell

Abbreviations and Acronyms:

AI = artificial intelligence; **CNN** = convolutional neural network; **DL** = deep learning; **FID** = Fréchet inception distance; **GAN** = generative adversarial network; **PGAN** = progressively growing generative adversarial network; **RFI** = retinal fundus image; **ROP** = retinopathy of prematurity; **RVM** = retinal vessel map; **UMAP** = uniform manifold approximation and projection.

Keywords:

Artificial intelligence, Deep learning, Generative adversarial network, Retinopathy of prematurity.

Correspondence:

J. Peter Campbell, MD, MPH, Department of Ophthalmology, Casey Eye Institute, Oregon Health & Science University, 545 SW Campus Drive, Portland, OR 97239. E-mail: campbelp@ohsu.edu.

References

1. Brown JM, Campbell JP, Beers A, et al. Automated diagnosis of plus disease in retinopathy of prematurity using deep convolutional neural networks. *JAMA Ophthalmol*. 2018;136(7):803–810.
2. Rajpurkar P, Irvin J, Zhu K, et al. CheXNet: radiologist-level pneumonia detection on chest x-rays with deep learning. Published online November 14, 2017. Available at: http://arxiv.org/abs/1711.05225. Accessed 15.06.21.
3. Chen JS, Coyner AS, Ostmo S, et al. Deep learning for the diagnosis of stage in retinopathy of prematurity: accuracy and generalizability across populations and cameras. *Ophthalmol Retina*. 2021;5(10):1027–1035.
4. Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*. 2016;316(22):2402–2410.
5. Shen L, Margolies LR, Rothstein JH, et al. Deep learning to improve breast cancer detection on screening mammography. *Sci Rep*. 2019;9(1):1–12.
6. De Fauw J, Ledsam JR, Romera-Paredes B, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nat Med*. 2018;24(9):1342–1350.
7. Alom MZ, Yakopcic C, Hasan M, et al. Recurrent residual U-Net for medical image segmentation. *Journal of Medical Imaging*. 2019;6(1):014006.
8. Zhou Z, Siddiquee MMR, Tajbakhsh N, Liang J. UNet++: a nested u-net architecture for medical image segmentation. Deep learning in medical image analysis and multimodal learning for clinical decision support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, held in conjunction with MICCAI 2018, Granada, Spain. 2018;11045. Available at: https://arxiv.org/abs/1807.10165
9. Singh RP, Hom GL, Abramoff MD, et al. Current challenges and barriers to real-world artificial intelligence adoption for the healthcare system, provider, and the patient. *Transl Vis Sci Technol*. 2020;9(2):45.
10. Watson J, Hutyra CA, Clancy SM, et al. Overcoming barriers to the adoption and implementation of predictive modeling and machine learning in clinical care: what can we learn from US academic medical centers? *JAMA Open*. 2020;3(2):167–172.
11. Kohli MD, Summers RM, Raymond Geis J. Medical image data and datasets in the era of machine learning—whitepaper from the 2016 C-MIMI meeting dataset session. *J Digit Imaging*. 2017;30(4):392–399.
12. Gensure RH, Chiang MF, Campbell JP. Artificial intelligence for retinopathy of prematurity. *Curr Opin Ophthalmol*. 2020;31(5):312–317.
13. Larson DB, Magnus DC, Lungren MP, et al. Ethics of using and sharing clinical imaging data for artificial intelligence: a proposed framework. *Radiology*. 2020;295(3):675–682.
14. Office for Civil Rights (OCR), United States Department of Health and Human Services. Guidance regarding methods for DE-identification of protected health information in

accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule. Published September 7, 2012. Available at: https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html. Accessed 27.01.22.

15. Kalpathy-Cramer J, Campbell JP, Erdogmus D, et al. Plus disease in retinopathy of prematurity: improving diagnosis by ranking disease severity and using quantitative image analysis. *Ophthalmology*. 2016;123(11):2345–2351.
16. Chiang MF, Quinn GE, Fielder AR, et al. International Classification of Retinopathy of Prematurity, Third Edition. *Ophthalmology*. 2021;128(10):e51–e68.
17. Goodfellow IJ, Pouget-Abadie J, Mirza M, et al. Generative adversarial networks. Published online June 10, 2014. Available at: http://arxiv.org/abs/1406.2661. Accessed 15.06.21.
18. Karras T, Aila T, Laine S, Lehtinen J. Progressive growing of GANs for improved quality, stability, and variation. Published online October 27, 2017. Available at: http://arxiv.org/abs/1710.10196. Accessed 15.06.21.
19. Coyner AS, Chen J, Campbell JP, et al. Diagnosability of synthetic retinal fundus images for plus disease detection in retinopathy of prematurity. *AMIA Annu Symp Proc*. 2021;2020:329–337.
20. World Medical Association. World Medical Association Declaration of Helsinki: ethical principles for medical research involving human subjects. *JAMA*. 2013;310(20):2191–2194.
21. Ryan MC, Ostmo S, Jonas K, et al. Development and evaluation of reference standards for image-based telemedicine diagnosis and clinical research studies in ophthalmology. *AMIA Annu Symp Proc*. 2014;2014:1902–1910.
22. Gilbert C, Fielder A, Gordillo L, et al. Characteristics of infants with severe retinopathy of prematurity in countries with low, moderate, and high levels of development: implications for screening programs. *Pediatrics*. 2005;115(5):e518–e525.
23. Quinn GE. Retinopathy of prematurity blindness worldwide: phenotypes in the third epidemic. *Eye Brain*. 2016;8:31–36.
24. Good WV, Hardy RJ, Dobson V, et al. The incidence and course of retinopathy of prematurity: findings from the early treatment for retinopathy of prematurity study. *Pediatrics*. 2005;116(1):15–23.
25. Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation. Published online May 18, 2015. Available at: http://arxiv.org/abs/1505.04597. Accessed 15.06.21.
26. Facebook Research. facebookresearch/pytorch_GAN_zoo. Available at: https://github.com/facebookresearch/pytorch_GAN_zoo. Accessed 15.06.21.
27. McInnes L, Healy J, Melville J. UMAP: uniform manifold approximation and projection for dimension reduction. Published online February 9, 2018. Available at: http://arxiv.org/abs/1802.03426. Accessed 15.06.21.
28. Brock A, Donahue J, Simonyan K. Large scale GAN training for high fidelity natural image synthesis. Published online September 28, 2018. Available at: http://arxiv.org/abs/1809.11096. Accessed 15.06.21.
29. Russakovsky O, Deng J, Su H, et al. ImageNet large scale visual recognition challenge. Published online September 1, 2014. Available at: http://arxiv.org/abs/1409.0575. Accessed 15.06.21.
30. Szegedy C, Vanhoucke V, Ioffe S, et al. Rethinking the inception architecture for computer vision. Published online December 2, 2015. Available at: http://arxiv.org/abs/1512.00567. Accessed 08.11.21.
31. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. Published online September 4, 2014. Available at: http://arxiv.org/abs/1409.1556. Accessed 15.06.21.
32. Choi RY, Coyner AS, Kalpathy-Cramer J, et al. Introduction to machine learning, neural networks, and deep learning. *Transl Vis Sci Technol*. 2020;9(2):14.
33. Kingma DP, Ba J. Adam: a method for stochastic optimization. Published online December 22, 2014. Available at: http://arxiv.org/abs/1412.6980. Accessed 19.10.21.
34. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning*. 2nd ed. New York, NY: Springer; 2017.
35. Hripcsak G, Heitjan DF. Measuring agreement in medical informatics reliability studies. *J Biomed Inform*. 2002;35(2):99–110.
36. Office for Civil Rights (OCR), United States Department of Health and Human Services. Methods for de-identification of PHI. Published September 7, 2012. Available at: https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html. Accessed 14.10.21.
37. Burlina P, Joshi N, Paul W, et al. Addressing artificial intelligence bias in retinal diagnostics. *Transl Vis Sci Technol*. 2021;10(2):13.
38. Beaulieu-Jones BK, Wu ZS, Williams C, et al. Privacy-preserving generative deep neural networks support clinical data sharing. *Circ Cardiovasc Qual Outcomes*. 2019;12(7):e005122.
39. Wang Z, Bovik AC, Sheikh HR, Simoncelli EP. Image quality assessment: from error visibility to structural similarity. *IEEE Trans Image Process*. 2004;13(4):600–612.
40. Barratt S, Sharma R. A note on the inception score. Published online January 6, 2018. Available at: http://arxiv.org/abs/1801.01973. Accessed 19.10.21.
41. Chicco D. Siamese neural networks: an overview. In: Cartwright H, ed. *Artificial Neural Networks*. New York: Humana; 2021:73–94. https://doi.org/10.1007/978-1-0716-0826-5_3.