Journal of Imaging Informatics in Medicine (2024) 37:1228–1238
https://doi.org/10.1007/s10278-024-00976-4

Generative Adversarial Networks for Brain MRI Synthesis: Impact of Training Set Size on Clinical Application

MM Zoghby[1] · BJ Erickson[1] · GM Conte[1]

Received: 12 September 2023 / Revised: 27 November 2023 / Accepted: 27 November 2023 / Published online: 16 February 2024

Abstract

We evaluated the impact of training set size on generative adversarial networks (GANs) to synthesize brain MRI sequences. We compared three sets of GANs trained to generate pre-contrast T1 (gT1) from post-contrast T1 and FLAIR (gFLAIR) from T2. The baseline models were trained on 135 cases; for this study, we used the same model architecture but a larger cohort of 1251 cases and two stopping rules, an early checkpoint (early models) and one after 50 epochs (late models). We tested all models on an independent dataset of 485 newly diagnosed gliomas. We compared the generated MRIs with the original ones using the structural similarity index (SSI) and mean squared error (MSE). We simulated scenarios where either the original T1, FLAIR, or both were missing and used their synthesized version as inputs for a segmentation model with the original post-contrast T1 and T2. We compared the segmentations using the dice similarity coefficient (DSC) for the contrast-enhancing area, non-enhancing area, and the whole lesion. For the baseline, early, and late models on the test set, for the gT1, median SSI was .957, .918, and .947; median MSE was .006, .014, and .008. For the gFLAIR, median SSI was .924, .908, and .915; median MSE was .016, .016, and .019. The range DSC was .625–.955, .420–.952, and .610–.954. Overall, GANs trained on a relatively small cohort performed similarly to those trained on a cohort ten times larger, making them a viable option for rare diseases or institutions with limited resources.

Keywords GAN · Synthetic MRI · Image-to-image translation · Glioma · MRI

Abbreviations

GAN	Generative adversarial network
cGAN	Conditional generative adversarial network
T1	Pre-contrast T1-weighted MRI scan
T1Gd	Post-contrast T1-weighted MRI scan
T2	T2-weighted MRI scan
FLAIR	Fluid attenuated inversion recovery MRI scan

Introduction

Generative adversarial networks (GANs) have gained prominence due to their capacity to generate synthetic images. GANs consist of two competing neural networks: the generator and the discriminator. The generator creates the images, while the discriminator attempts to distinguish real from generated images, forcing the generator to refine its output to deceive the discriminator [1]. A specialized form of GAN, known as conditional generative adversarial network (cGAN), has proven especially valuable in medical applications. Unlike standard GANs, cGANs generate images guided by target images instead of noise [2]. For example, they can transform a sketch into a painting or convert a post-contrast T1-weighted MRI scan into its pre-contrast counterpart. GANs have demonstrated substantial success in medical contexts, spanning tasks from CT scan denoising and image super-resolution to generating synthetic images of cell embryos and liver lesions [3].

A fundamental challenge in neural network training is the availability of high-quality data. The creation of sizable, well-curated datasets has proven pivotal in advancing state-of-the-art algorithms. In this regard, one of the most influential datasets for neuroimaging and neuro-oncology applications is the one collected for the Multimodal Brain Tumor Image Segmentation Benchmark (BraTS) challenge [4]. The significance of synthetic data in medicine is underscored by the inclusion of two BraTS 2023 challenges to synthetic MRI image generation [5, 6]. However, a knowledge gap persists regarding the influence of dataset

✉ GM Conte
conte.gianmarco@mayo.edu

[1] Department of Radiology, Mayo Clinic, Rochester, MN, USA

size on training cGANs. Determining the impact of varied training set sizes is crucial for assessing the viability of model development under data-scarce conditions, such as institutions with constrained access to data and computational resources, or in cases of rare diseases.

Motivated by the common issue of missing data, in a previous study, we trained two cGANs to generate pre-contrast T1-weighted (T1) sequences from post-contrast T1-weighted (T1Gd) sequences and fluid attenuated inversion recovery (FLAIR) sequences from T2-weighted (T2) sequences. We found that cGANs can effectively synthesize missing brain MRI data and that the segmentations of brain lesions obtained using GAN-generated images were comparable to those obtained with the original MRI scans [7]. In this study, we extend our previous work by assessing how the performance of cGANs for the same two medical image tasks is influenced by training set size. Additionally, we conduct a supplementary analysis by including an early stopping paradigm to evaluate the feasibility of decreasing training time on larger datasets.

Materials and Method

Data

For this study, we utilized a publicly available dataset (BraTS challenge) [4, 8–10] as well as data originating from our institution. We were granted an exemption from the requirement for IRB approval (45 CFR 46.104d, Category 4).

The BraTS dataset includes clinically acquired T1, T1Gd, T2, and FLAIR multiparametric MRI (mpMRI) scans of gliomas from various institutions. In this paper, we distinguish between two versions of this dataset: the BraTS 2017 and BraTS 2021 datasets. The BraTS 2017 dataset constitutes a subset of the BraTS 2021 dataset. We label the BraTS 2017 dataset as the *small dataset* ($n = 168$) and the BraTS 2021 dataset as the *large dataset* ($n = 1{,}470$). The training and validation sets for the *small dataset* included 135 and 33 subjects, comprising a total of 20,925 and 6,510 slices for each image type, respectively [4, 8–10]. Likewise, the training and validation sets for the large dataset encompassed 1,251 and 219 BraTS volumes, containing a total of 193,905 and 33,945 slices for each image type, respectively [4, 8–10].

The local dataset consisted of 485 newly diagnosed IDH-wt gliomas, which we collected at our institution between 2017 and 2022. We obtained a single preoperative MRI examination, including axial T1, T1Gd, T2, and FLAIR sequences, for each subject. This cohort served as the test set for all our experiments.

Data Preprocessing

The BraTS and local datasets stored each MRI volume in the NIfTI file format. We used the NiBabel and NumPy Python libraries to convert the NIfTI files into the PNG file format [11, 12]. We resized each MRI slice to 256×256 and then rescaled them between -1 and 1. The same preprocessing pipeline was applied during training and inference.

Study Design

Previously, we trained two cGANs on the *small dataset* (BraTS 2017, $n = 135$) to generate pre-contrast T1 sequences from T1Gd (T1 Model) sequences and FLAIR from T2 sequences (FLAIR Model) [7]. We refer to these models as the *baseline models*. For this study, we trained two additional sets of models, utilizing an identical network architecture on the *large dataset* (BraTS 2021, $n = 1{,}251$): one that stopped training at an early checkpoint (*early models*) and one trained for 50 epochs (*late models*). The study design is depicted in Fig. 1.

cGANs Training

We used the pix2pix framework, a specialized type of cGAN for image-to-image translation, to train all models [13]. We trained the *early* and *late c*GAN models on the *large dataset* for two distinct medical image translation tasks. Throughout the training process, the model received a pair of images: one image served as the input sequence, while the other image represented the target sequence. In the T1 Model, we designated the post-contrast T1-weighted sequence as the input image, while the pre-contrast T1-weighted sequence was assigned to the target image. Similarly, in the FLAIR Model, we designated the T2-weighted sequence as the input image, and the FLAIR sequence as the target image.

We trained the models for 300 epochs on the *small dataset* and 50 epochs on the *large dataset*. During the training process, we alternately trained the discriminator and generator for one gradient descent step. The final loss function combined entropy and least absolute deviation (L1 loss). We updated the model weights using mini-batch stochastic gradient descent, employing a batch size of one, an Adam optimizer with a learning rate of 0.0002, and momentum parameters of $\beta 1 = 0.5$ and $\beta 2 = 0.999$. To augment the data, we applied techniques such as random cropping and horizontal (left/right) flipping.

We implemented the models in Python (version 3.7) and TensorFlow (version 2.11.0) and trained them on an

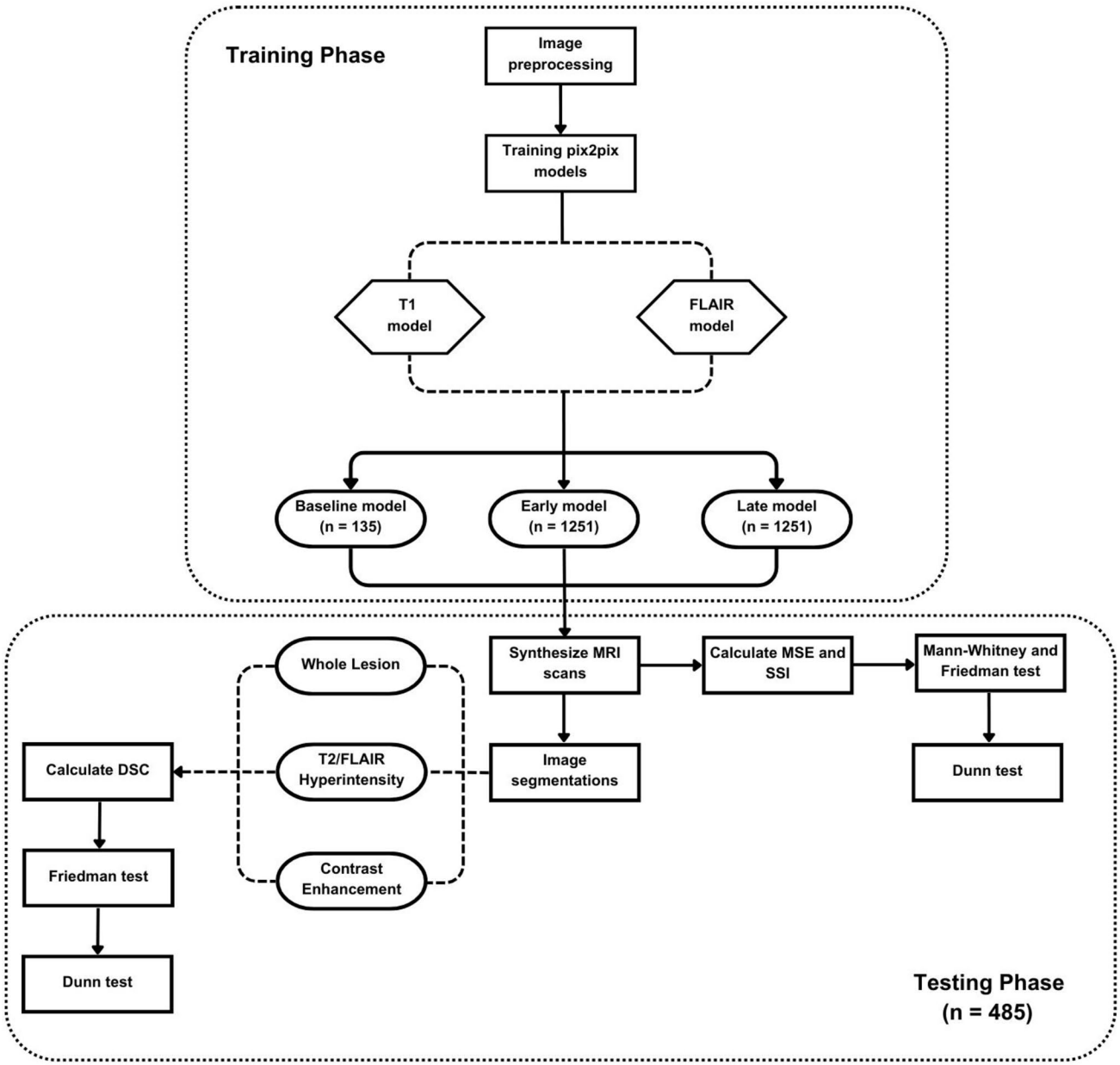

Fig. 1 Workflow showing the study of the design. T1 model: post-contrast T1-weighted to non-contrast T1-weighted model; FLAIR model: T2-weighted to fluid attenuated inversion recovery model; MSE: mean squared error; SSI: structural similarity index; DSC: Dice similarity coefficient

NVIDIA V100 card with 32 GB of memory. We utilized the Weights & Biases software for experiment tracking and loss visualization [14].

Determining the Early Stopping Point for the Early Model

We determined the early stopping point for the *early models* by identifying the smallest difference between the generator and discriminator loss. This point was reached at the 2nd epoch for the T1 model (Online Reference 1, E1-E2) and the 4th epoch for the FLAIR model (Online Reference 1, E3-E4).

Evaluation of Generated Images

We assessed differences in the MRI signal intensities between the GAN-generated images and the original images using the mean squared error (MSE) [15] and the structural similarity index (SSI) [16]. A value closer to 0

in the MSE indicated a higher degree of similarity, while a value closer to 1 in the SSI also indicated greater similarity. The SSI is theoretically more informative about human perceived similarity since it considered structural, contrast, and luminance information between the two images compared to the MSE [17, 18].

Segmentation Model

To quantify the ability of cGAN-generated images to function as surrogates for missing MRI sequences (pre-contrast T1 and FLAIR sequence), we used a publicly available pre-trained brain tumor segmentation model developed outside our institution (HD-GLIO) [19, 20]. We selected this model for its comprehensive training set, its ease of implementation, and its good performance [21]. In brief, the model was trained by other researchers on three datasets (3220 MRI scans in 1450 patients with brain tumors) collected across various institutions. It requires four MRI sequences (T1-weighted, postcontrast T1-weighted, T2-weighted, and FLAIR images) as input and generates two tumor segmentation masks—one for the contrast-enhancing regions and one for the areas of T2/FLAIR hyperintensities. This model is a modified U-Net and can be accessed at https://github.com/NeuroAI-HD/HD-GLIO [22].

Evaluation of Segmentation

We compared the generated data produced by the models through four scenarios: (a) no missing data, (b) missing T1 and FLAIR, (c) missing FLAIR, and (d) missing T1. We used the first scenario (no missing data) as a reference when comparing the segmentations produced by HD-GLIO (Fig. 2a). For the remaining scenarios, we utilized the GAN-generated MRI sequences (gT1 and gFLAIR) produced by the 3 models (*baseline, early, and late*) as replacement inputs for HD-GLIO (Fig. 2b). We compared the segmentations produced using the synthetic data to the reference scenario by computing the Dice similarity coefficient (DSC) for the whole lesion, T2/FLAIR hyperintensity, and contrast-enhanced component. We calculated the DSC score using the following formula:

$$DiceScore = \frac{2\left|Y_{gt} \cap Y_{pred}\right|}{\left|Y_{gt}\right| + \left|Y_{pred}\right|}$$

Here, Y_{gt} represents the segmentations acquired from the four original MRI scans, which served as the ground truth. Y_{pred} represents the segmentations acquired in the three scenarios where we simulated the absence of MRI sequences. The DSC score ranges from 0 (no overlap) to 1 (perfect overlap).

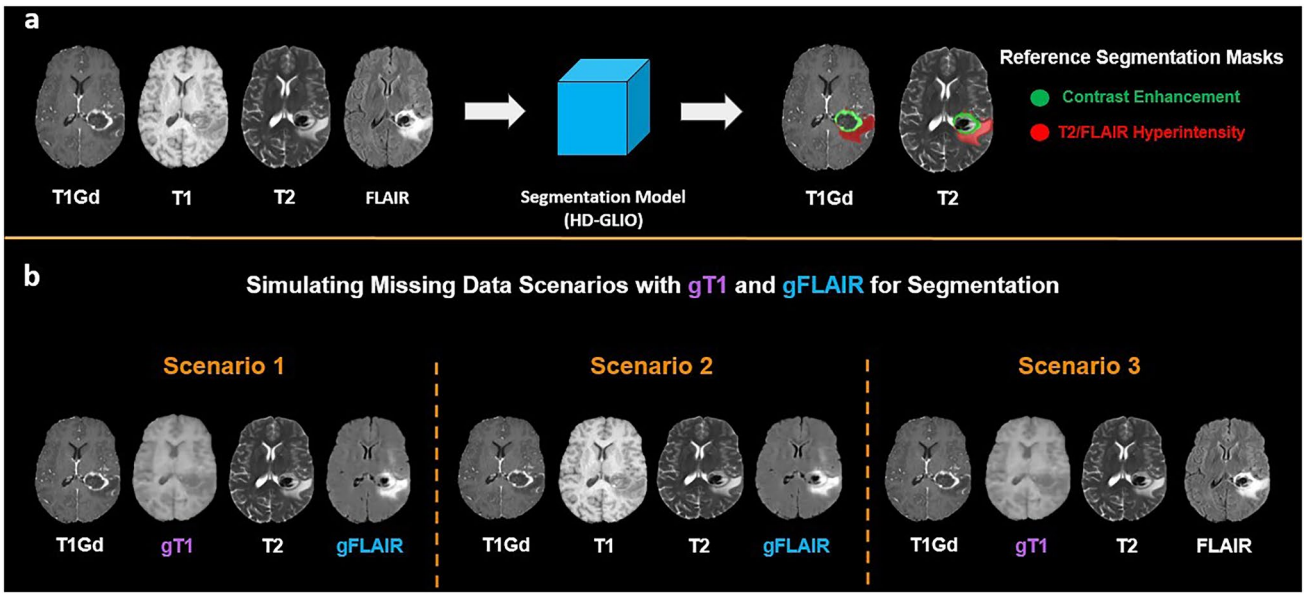

Fig. 2 Lesion segmentation. T1Gd: post-contrast T1-weighted MRI sequence; FLAIR: fluid attenuated inversion recovery; T2: T2-weighted MRI sequence; gT1: cGAN-generated T1-weighted sequence; gFLAIR: cGAN-generated FLAIR sequence. **a** The segmentation model (HD-GLIO) requires four MRI sequences (T1 Gd, T1, T2, FLAIR) as inputs to obtain masks of contrast-enhanced areas (green) and areas of T2 and FLAIR hyperintensities (red). The masks obtained with original MRI scans serve as reference in comparison with those obtained in different scenarios. **b** We simulated three different clinical scenarios utilizing the cGAN-generated sequences as inputs (gT1 and gFLAIR) for the segmentation model: missing T1 and FLAIR, missing FLAIR, and missing T1

Statistical Analysis

We conducted all statistical computations using Prism 10 software (GraphPad 10.0.0). We employed nonparametric tests for all analyses. To compare the MSE and SSI between the T1 models and the FLAIR models, we utilized the Mann–Whitney test. Additionally, we employed the Friedman test to compare the MSE, SSI, and DSC values derived from the three models (baseline, early, and late). To correct multiple comparisons, we implemented the Dunn test. A statistically significant difference was denoted by $P < 0.01$.

Results

Evaluation of Generated Images

Table 1 displays the MSE and SSI (median, 25th, and 75th percentiles) for the three versions of the T1 and FLAIR models obtained on the institutional test set ($n = 485$). For both the T1 and FLAIR models, the baseline models yielded the lowest MSE and the highest SSI, except for one instance (where the median MSE for the baseline FLAIR and early FLAIR model were both 0.016). For the T1 models, median MSE values ranged from 0.006 to 0.014, and the median SSI ranged from 0.947 to 0.957. For the FLAIR models, the median MSE ranged from 0.016 to 0.019, while the median SSI ranged from 0.908 to 0.924. Overall, the T1 models consistently exhibited a trend of lower MSE and higher SSI compared to the FLAIR models.

Evaluation of Segmentations

Summary results for all scenarios and segmentations are reported as violin plots in Figs. 3, 4, and 5. Examples of the best and worst DSC scores for the segmentations when missing both T1 and FLAIR scans (Scenario 1) are displayed in Figs. 6 and 7. Table 2 displays the DSC values (median, 25th, and 75th percentiles) obtained from the internal test set ($n = 485$). Median DSC values ranged between 0.625–0.955, 0.420–0.952, and 0.610–0.954 for the *baseline*, *early*, and *late* models, respectively. The *baseline* model achieved the highest median DSC score in six out of nine of the scenarios.

The *late* model achieved a higher DSC score for the whole lesion mask when FLAIR scans were missing (DSC = 0.765), while the *early* model achieved higher DSC scores for the T2/FLAIR hyperintensity and contrast enhancement masks when T1 scans were absent (DSC = 0.936 and 0.913, respectively). The *early* model consistently displayed comparatively lower DSC values across scenarios, except for contrast enhancement masks (Fig. 5) and instances where only the T1-weighted scan was absent (Scenario 3) (see panel c, Figs. 3, 4, and 5).

Notably, the absence of FLAIR scans (Scenario 1 and Scenario 2) significantly reduced segmentation quality for the entire lesion and T2/FLAIR hyperintensity, resulting in median DSC scores ranging from 0.420 to 0.765 (see panel b, Figs. 3 and 4). We observed optimal performance across all models (*baseline*, *early*, and *late*) for the contrast enhancement mask (Fig. 5) and in cases where only the T1 scan was missing (Scenario 3) (see panel c, Figs. 3,4, and 5), achieving DSC scores ranging from 0.884 to 0.955 and 0.908 to 0.948, respectively.

In scenarios where both T1 and FLAIR scans were missing (Figs. 6 and 7), the *baseline* and *late* models showed significant statistical differences from the *early* models in most comparisons ($P < 0.0001$), except for the contrast enhancement mask (Table 2, Scenario 1). For the contrast mask in Scenario 1, a significant statistical difference existed between the *baseline* and *early* models ($P < 0.0001$) but not between the *late* and *early* models. When FLAIR scans were absent, the *baseline* and *late* models exhibited significant statistical differences from the early models across all comparisons ($P < 0.0001$). When T1 scans were absent, a statistical difference

Table 1 Mean Squared Error (MSE) and Structural Similarity Index (SSI) for Internal Test Set (n = 485)

	MSE		SSI	
Model Scenario	T1 Model	FLAIR Model	T1 Model	FLAIR Model
Baseline	0.006 (0.001–0.021)	0.016 (0.006–0.031)	0.957 (0.889–0.999)	0.924 (0.840–0.995)
Early	0.014 (0.002–0.035)	0.016 (0.006–0.030)	0.918 (0.820–0.999)	0.908 (0.800–0.995)
Late	0.008 (0.001–0.029)	0.019 (0.006–0.038)	0.947 (0.849–0.999)	0.915 (0.816–0.995)

Numbers are medians, with 25th–75th percentiles in parentheses. For each model scenario (baseline, early, and late), the mean squared error and the structural similarity index of the T1 models and FLAIR models were compared using the Mann–Whitney test ($P < 0.0001$ for all comparisons, Online Reference 1, E5). The performance of the baseline, early, and late models was compared using the Friedman test for the T1 and FLAIR medical synthesis task. Correction for multiple comparisons was performed with the Dunn test ($P < 0.0001$ for all comparisons, Online Reference 1, E6). *Baseline models*, trained on small dataset ($n = 135$ subjects); *Early models*, trained on large dataset ($n = 1{,}251$ subjects) and stopped early; *Late models*, trained on large dataset for 50 epochs

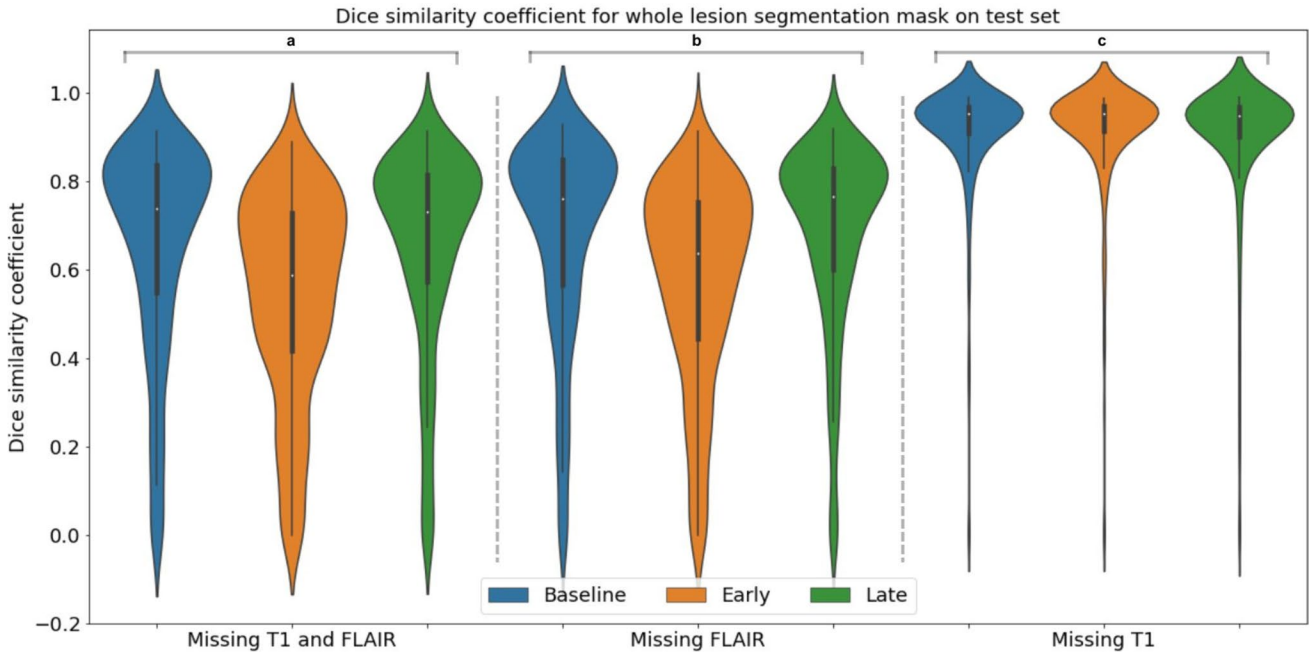

Fig. 3 Dice similarity coefficient for whole lesion segmentation mask on test set. (a) Missing T1 and FLAIR. (b) Missing FLAIR. (c) Missing T1. Blue: *baseline models*; orange: *early models*; green: *late models*. *Baseline models*: trained on small dataset ($n = 135$ subjects); *Early models*: trained on large dataset ($n = 1{,}251$ subjects) and stopped early; *Late models*: trained on large dataset for 50 epochs

emerged between the *late* and *early* models for the whole lesion and FLAIR/T2 hyperintensity mask ($P < 0.0001$).

We observed significant statistical differences between the *baseline* and *late* models in two out of nine segmentation comparisons. These differences appeared for contrast enhancement masks when both T1 and FLAIR scans were missing (Scenario 1, $P < 0.0001$) and when only T1 scans were missing (Scenario 3, $P = 0.0006$).

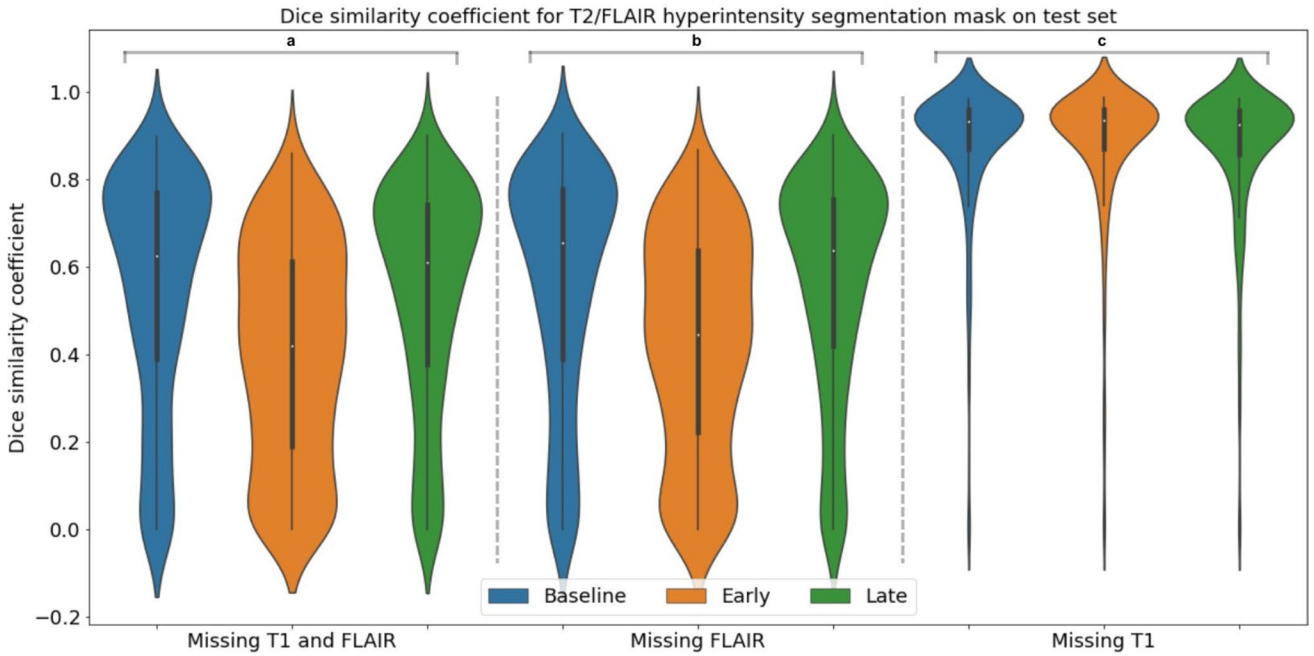

Fig. 4 Dice similarity coefficient for T2/FLAIR hyperintensity segmentation mask on test set. (a) Missing T1 and FLAIR. (b) Missing FLAIR. (c) Missing T1. Blue: *baseline models*; orange: *early models*; green: *late models*. *Baseline models*: trained on small dataset ($n = 135$ subjects); *Early models*: trained on large dataset ($n = 1{,}251$ subjects) and stopped early; *Late models*: trained on large dataset for 50 epochs

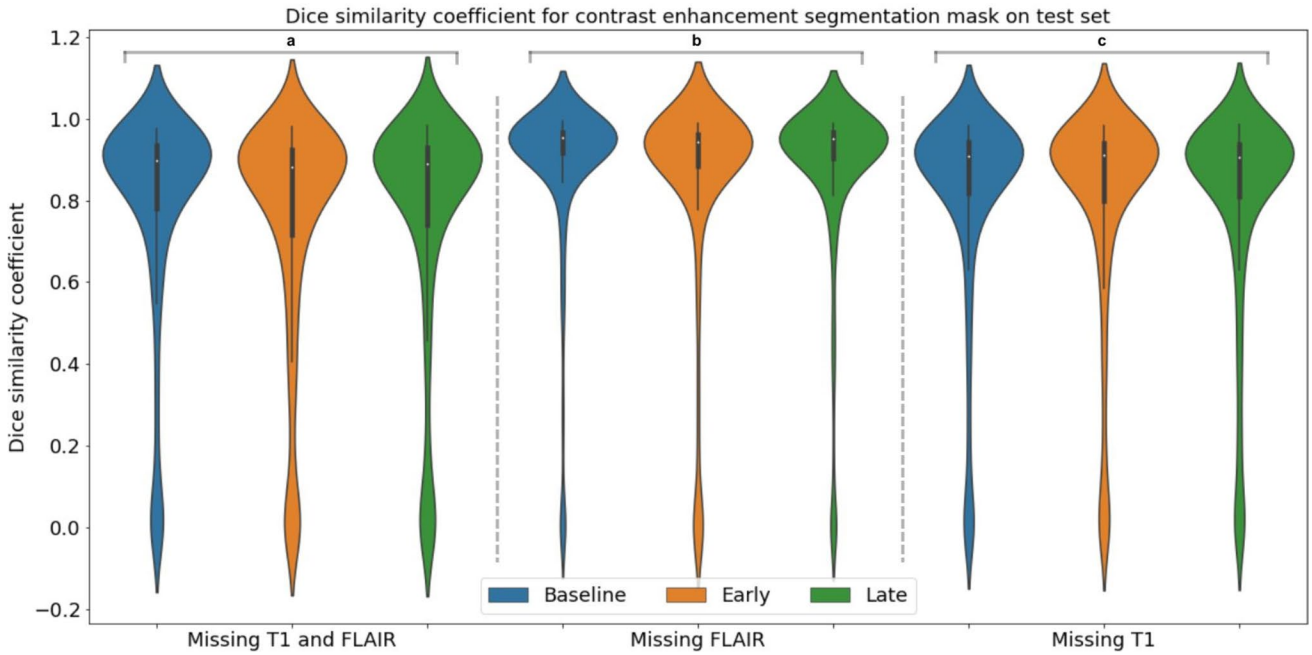

Fig. 5 Dice similarity coefficient for contrast enhancement segmentation mask on test set. (a) Missing T1 and FLAIR. (b) Missing FLAIR. (c) Missing T1. Blue: *baseline models*; orange: *early models*; green: *late models*. *Baseline models*: trained on small dataset (135 subjects); *Early models*: trained on large dataset (1,251 subjects) and stopped early; *Late models*: trained on large dataset for 50 epochs

Discussion

The sampling size is a crucial aspect in the training of deep learning models. However, it remains unclear how larger datasets impact the performance of GANs. Our previous research demonstrated that GANs can effectively generate missing MRI sequences for brain tumor segmentation [7]. In this study, we extend this work by investigating the influence of training set size on model performance. Overall, we observe that generative models trained on a relatively smaller cohort perform comparably to those trained on a dataset ten times larger.

Although statistically significant differences in mean squared error (MSE) and structural similarity index (SSI) existed between the *baseline* and *late* models for the MSE and SSI for both the T1 and FLAIR models ($P < 0.0001$ for both), the practical disparity was minimal (MSE: 0.002 and 0.003; SSI: 0.009 and 0.010 for T1 and FLAIR models, respectively). Consequently, the quality assessment of synthetic images in clinical contexts is crucial. Subsequently, we used gT1 and gFLAIR images as inputs for a pre-trained brain tumor segmentation model across various clinical scenarios (missing T1 and FLAIR, missing FLAIR, and missing T1) and segmentation masks (whole lesion, T2/FLAIR hyperintensity, and contrast enhancement). For the *baseline* and *late* models, we observed statistical differences in the DSC score in two out of nine segmentation comparisons. These differences only emerged for the contrast enhancement mask in Scenario 1 (missing T1 and FLAIR) and Scenario 3 (missing T1). However, the actual DSC scores were comparable and high-scoring for both of these models—0.902 vs 0.893 (Scenario 1); and 0.911 and 0.908 (Scenario 3) for *baseline* and *late* models, respectively—indicating analogous performance for the *baseline* and *late* models performed across all experiments.

Given the negligible differences between models trained on distinct cohort sizes, we evaluated the GAN model trained on the larger dataset at the loss function's convergence point (*early* model) to determine if earlier stopping epochs were viable (epoch 2 for the T1 Model and epoch 4 for the FLAIR model). We reasoned that if there was no difference in model performance based on dataset size, perhaps the model trained on the *large dataset* learned all the image features in fewer epochs.

Overall, the *early* model's performance was worse than the *baseline* and *late* models, particularly in FLAIR synthesis (see G, Figs. 3 and 4). The reduction in the quality of the FLAIR significantly decreased the quality of the segmentations produced in scenarios where the FLAIR was missing (Scenario 1 and Scenario 2), demonstrated by the significantly lower DSC scores with differences as large as 0.209.

T1 synthesis emerged as a robust scenario across models, attaining high DSC values (0.908–0.953) when only the T1 was missing (Scenario 3). Conversely, synthetic FLAIR generation proved challenging, as reflected in higher MSE, lower SSI, and reduced DSC scores, even with a larger

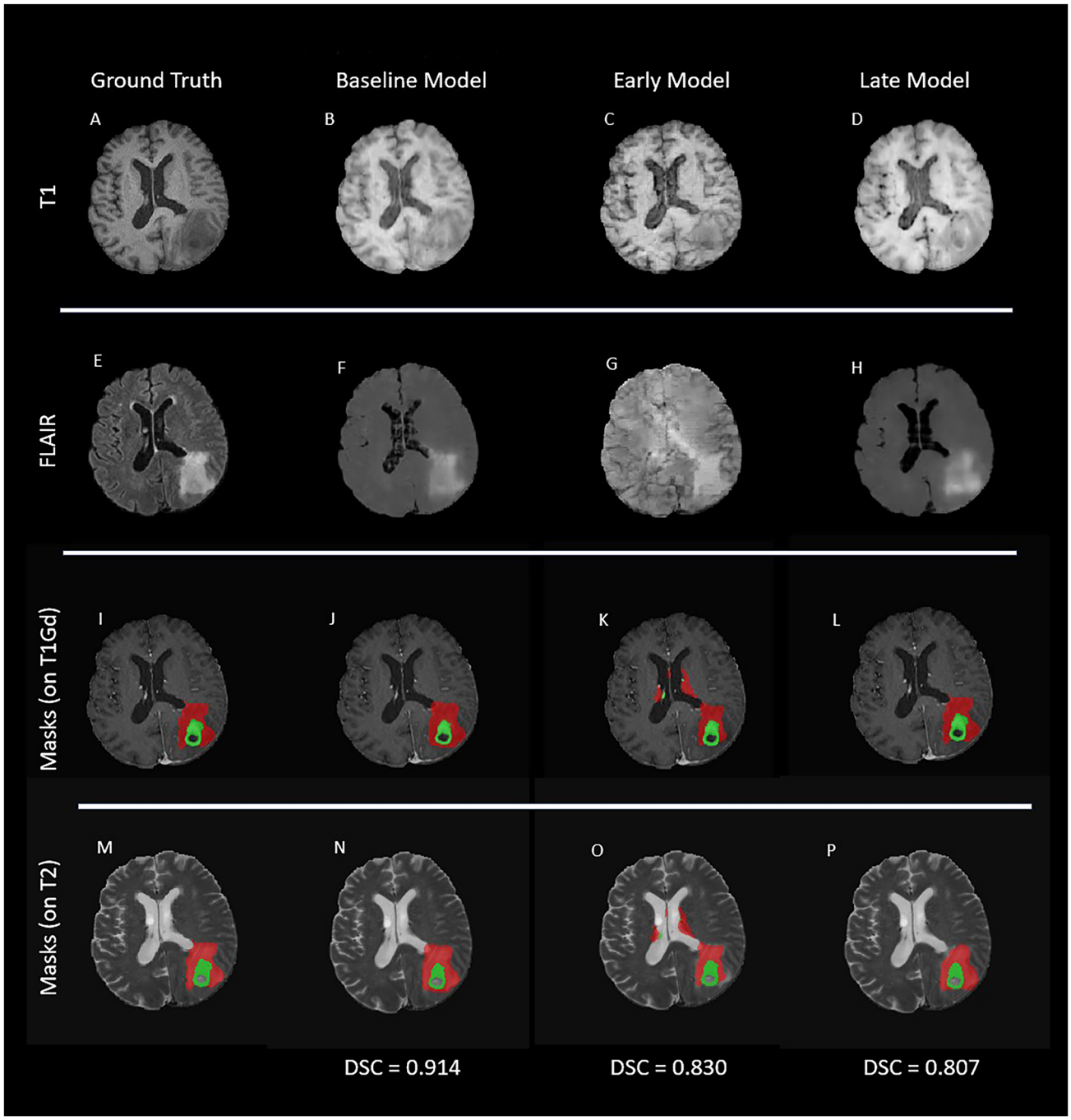

Fig. 6 Examples of high-quality segmentation when missing T1 and FLAIR. (A) Original T1-weighted MRI scan. (B–D) T1-weighted MRI scan generated with GAN (gT1) created by our models. (E) Original fluid attenuated inversion recovery (FLAIR) MRI scan. (F–H) FLAIR MRI scan generated with GAN (gFLAIR) created by our models. (I, M) Segmentation obtained using only original MRI scans that served as reference. (J–L) (N–P) Segmentation of lesion obtained with generated FLAIR image and generated T1-weighted image instead of original MRI scans. Red indicates segmentation of T2 and/or FLAIR hyperintensity. Green area indicates segmentation of the enhancing area of lesion

training set. In addition to T1 synthesis, the segmentation of the contrast-enhancing area performed well across all models (*baseline*, *early*, and *late*) and scenarios (missing T1 and FLAIR, missing FLAIR, and missing T1) compared to the whole lesion and T2/FLAIR hyperintensity masks. The quality of synthetic images directly impacted segmentations, with missing FLAIR images exerting the greatest influence on T2/FLAIR hyperintensity segmentation, while missing

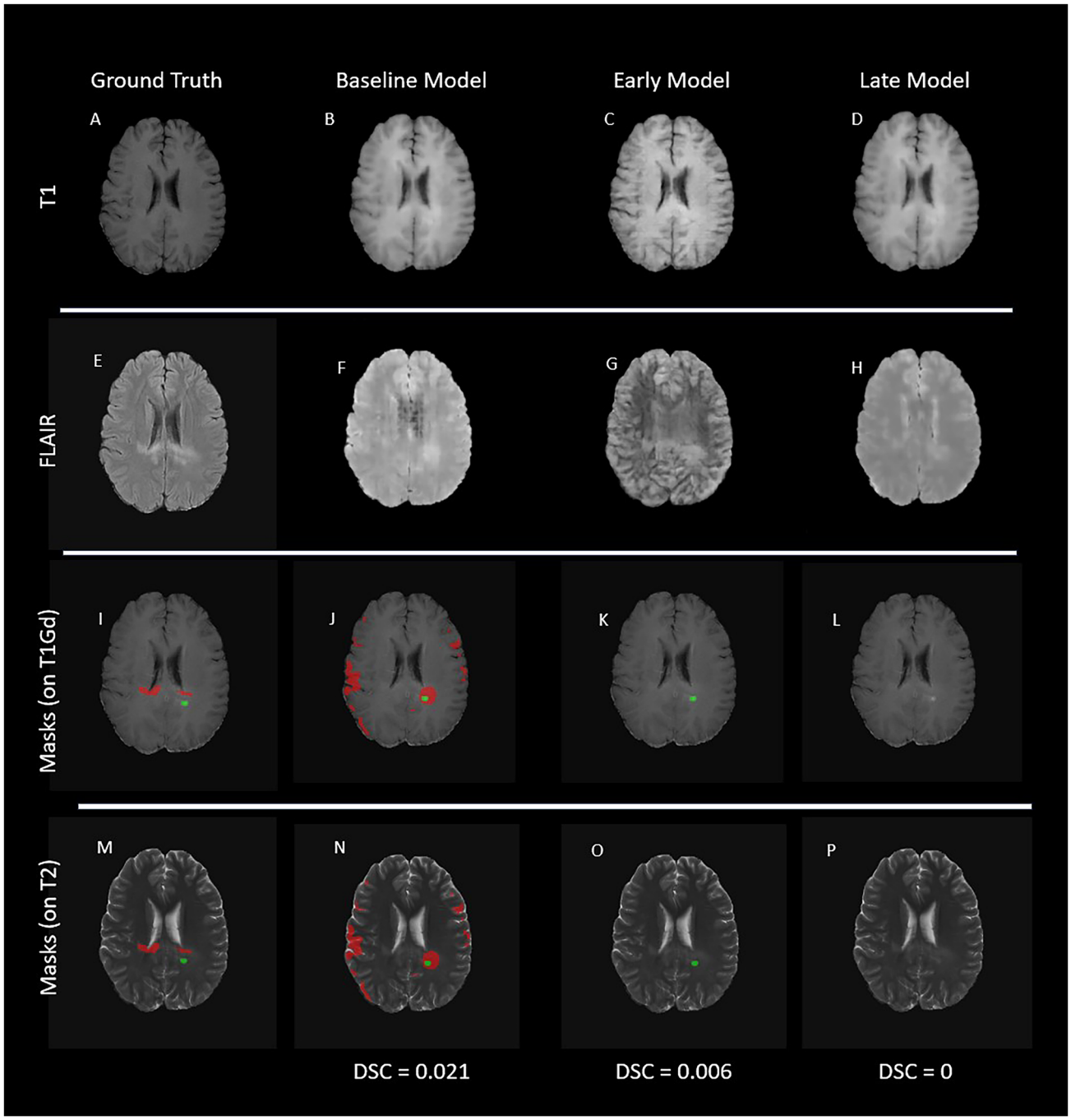

Fig. 7 Examples of low-quality segmentation when missing T1 and FLAIR. (A) Original T1-weighted MRI scan. (B–D) T1-weighted MRI scan generated with cGAN (gT1) created by our models. (E) Original fluid attenuated inversion recovery (FLAIR) MRI scan. (F–H) FLAIR MRI scan generated with cGAN (gFLAIR) created by our models. (I, M) Segmentation obtained using only original MRI scans that served as reference. (J–L) (N–P) Segmentation of lesion obtained with generated FLAIR image and generated T1-weighted image instead of original MRI scans. Red indicates segmentation of T2 and/or FLAIR hyperintensity. Green area indicates segmentation of the enhancing area of lesion

T1 images, to a lesser extent, affected the contrast enhancement segmentation.

There have been other works that have evaluated the impact of training set size. For convolutional neural networks, it was found that model performance improves as a power law relationship with respect to dataset size [23, 24]. However, previous work on cGANs has found that performance stabilized around 20 patients when used

Table 2 Dice similarity coefficient (DSC) values for internal test set ($n = 485$)

Segmentation masks	Scenario 1			Scenario 2			Scenario 3		
	Missing T1-weighted and FLAIR scans			Missing FLAIR scans			Missing T1-weighted scans		
	Models			Models			Models		
	Baseline	Early	Late	Baseline	Early	Late	Baseline	Early	Late
Whole lesion	0.738 (0.539–0.837)*	0.588 (0.416–0.727)	0.731 (0.573–0.814)*	0.760 (0.565–0.850)*	0.637 (0.445–0.752)	0.765 (0.598–0.830)*	0.953 (0.908–0.968)	0.952 (0.914–0.971)	0.948 (0.902–0.967)*
T2 and/or FLAIR hyperintensities	0.625 (0.387 – 0.770)*	0.420 (0.190–0.615)	0.610 (0.378–0.743)*	0.655 (0.389–0.778)*	0.446 (0.222–0.637)	0.637 (0.419–0.755)*	0.932 (0.870–0.959)	0.936 (0.871–0.960)	0.926 (0.858–0.957)*
Contrast enhancement	0.902 (0.782–0.939)*^	0.884 (0.724–0.929)	0.893 (0.750–0.934)^	0.955 (0.920–0.970)*	0.943 (0.885–0.962)	0.954 (0.908–0.968)*	0.911 (0.821–0.947)^	0.913 (0.815–0.944)	0.908 (0.818–0.941)^

Numbers are medians, with 25th–75th percentiles in parentheses. The results obtained for each segmentation type and each scenario are compared using the Friedman test. Correction for multiple comparisons was performed using the Dunn test (Online Reference 1, E7). The asterisk (*) denotes a comparison between the *early model* where the *P*-value was found to be statistically significant ($P < 0.01$). The carrot (^) denotes a comparison between the *baseline* and *late* models that was found to be statistically significant. This occurred in only 2 scenarios for the contrast enhancement mask: when T1 and FLAIR were missing (Scenario 1) and when T1 was missing (Scenario 3). *Baseline models*, trained on small dataset (135 subjects); *Early models*, trained on large dataset (1,251 subjects) and stopped early; *Late models*, trained on large dataset for 50 epochs

to synthesize segmentations for pelvic and thorax CT images [25, 26]. To the best of our knowledge, our work is the first attempt to assess the impact of training set size for brain MRI synthesis.

Our study had limitations. (1) We did not generate all possible MRI sequences since our aim was to reproduce the original study design of the *baseline* models. (2) We did not test every epoch between the early stopping point and the final checkpoint for the model trained on the *large* dataset. While it is likely that there is an epoch between these two checkpoints that learned all the features, it was beyond the scope of the study to assess all these model checkpoints. (3) We tested only two training set sizes. While testing more cohort size has definite value, it was out of scope for this study, which aims to be a proof-of-concept study to inspire more research on this field. (4) Finally, we did not test our approach on other disease or image types.

Conclusion

This study evaluated the impact of dataset size on cGAN performance for synthesizing T1-weighted (T1) and fluid attenuated inversion recovery scans (FLAIR) scans. We found that a tenfold dataset increase did not notably enhance performance. This suggests that cGANs can effectively generate missing MRIs when data is limited, which is beneficial for rare diseases and resource-limited settings.

We introduced an *early model* trained until GAN losses converged at an earlier epoch on the *large dataset*. While the *early model* performed less well than *baseline* and *late* models in FLAIR synthesis, it matched their performance in pre-contrast T1 synthesis and contrast enhancement segmentation. Therefore, if either of these aspects takes precedence in a medical or research setting, they can be synthetically generated using fewer computing resources.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10278-024-00976-4.

Author Contribution Study concepts/study design or data acquisition or data analysis/interpretation, all authors; manuscript drafting or manuscript revision for important intellectual content, all authors; approval of final version of submitted manuscript, all authors; agrees to ensure any questions related to the work are appropriately resolved, all authors; literature research, M.M.Z., G.M.C.; statistical analysis, M.M.Z.,G.M.C; and manuscript editing, all authors.

Funding Center for Individualized Medicine, Mayo Clinic

Data Availability The data used to train the models is publicly available. The BraTS 2017 dataset is available at https://www.med.upenn.edu/sbia/brats2017/data.html. The BraTS 2021 dataset is available at http://braintumorsegmentation.org/.

Declarations

Ethics Approval and Consent to Participate This project was granted an exemption from the requirement for IRB approval (45 CFR 46.104d, Category 4).

Competing Interest The authors declare no competing interests.

References

1. Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial networks. *Commun ACM*. 2020;63(11):139-144.
2. Mirza M, Osindero S. Conditional Generative Adversarial Nets. *arXiv [csLG]*. Published online November 6, 2014. http://arxiv.org/abs/1411.1784
3. Lan L, You L, Zhang Z, et al. Generative Adversarial Networks and Its Applications in Biomedical Informatics. *Front Public Health*. 2020;8:164.
4. Menze BH, Jakab A, Bauer S, et al. The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). *IEEE Trans Med Imaging*. 2015;34(10):1993-2024.
5. Li HB, Conte GM, Anwar SM, et al. The Brain Tumor Segmentation (BraTS) Challenge 2023: Brain MR Image Synthesis for Tumor Segmentation (BraSyn). *arXiv [eessIV]*. Published online May 15, 2023. http://arxiv.org/abs/2305.09011
6. Kofler F, Meissen F, Steinbauer F, et al. The Brain Tumor Segmentation (BraTS) Challenge 2023: Local Synthesis of Healthy Brain Tissue via Inpainting. *arXiv [eessIV]*. Published online May 15, 2023. http://arxiv.org/abs/2305.08992
7. Conte GM, Weston AD, Vogelsang DC, et al. Generative adversarial networks to synthesize missing T1 and FLAIR MRI sequences for use in a multisequence brain tumor segmentation model. Radiology. 2021;300(1):E319. https://doi.org/10.1148/radiol.2021203786
8. Baid U, Ghodasara S, Mohan S, et al. The RSNA-ASNR-MICCAI BraTS 2021 Benchmark on Brain Tumor Segmentation and Radiogenomic Classification. *arXiv [csCV]*. Published online July 5, 2021. http://arxiv.org/abs/2107.02314
9. Bakas S, Akbari H, Sotiras A, et al. Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features. *Scientific Data*. 2017;4(1):1-13.
10. S. Bakas, H. Akbari, A. Sotiras, M. Bilello, M. Rozycki, J. Kirby, et al. Segmentation Labels and Radiomic Features for the Pre-operative Scans of the TCGA-GBM collection (BraTS-TCGA-GBM). doi:https://doi.org/10.7937/K9/TCIA.2017.KLXWJJ1Q
11. Harris CR, Millman KJ, van der Walt SJ, et al. Array programming with NumPy. *Nature*. 2020;585(7825):357-362.
12. Brett M, Markiewicz CJ, Hanke M, et al. *Nipy/nibabel*.; 2022. https://nipy.org/nibabel/
13. Isola P, Zhu JY, Zhou T, Efros AA. Image-to-image translation with conditional adversarial networks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. ; 2017:1125–1134.
14. Biewald L. Experiment Tracking with Weights and Biases. Published online 2020. https://www.wandb.com/
15. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine Learning in Python. *arXiv [csLG]*. Published online January 2, 2012:2825–2830. Accessed June 1, 2023. https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf?ref=https:/
16. van der Walt S, Schönberger JL, Nunez-Iglesias J, et al. scikit-image: image processing in Python. *PeerJ*. 2014;2:e453.
17. Wang Z, Bovik AC. Mean squared error: Love it or leave it? A new look at Signal Fidelity Measures. *IEEE Signal Process Mag*. 2009;26(1):98-117.
18. 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.
19. Kickingereder P, Isensee F, Tursunova I, et al. Automated quantitative tumour response assessment of MRI in neuro-oncology with artificial neural networks: a multicentre, retrospective study. *Lancet Oncol*. 2019;20(5):728-740.
20. Isensee F, Jäger PF, Kohl SAA, Petersen J, Maier-Hein KH. Automated Design of Deep Learning Methods for Biomedical Image Segmentation. *arXiv [csCV]*. Published online April 17, 2019. http://arxiv.org/abs/1904.08128
21. Ghaffari M, Sowmya A, Oliver R. Automated Brain Tumor Segmentation Using Multimodal Brain Scans: A Survey Based on Models Submitted to the BraTS 2012–2018 Challenges. *IEEE Rev Biomed Eng*. 2020;13:156-168.
22. Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*. Springer International Publishing; 2015:234–241.
23. Sun C, Shrivastava A, Singh S, Gupta A. Revisiting unreasonable effectiveness of data in deep learning era. In: *Proceedings of the IEEE International Conference on Computer Vision*. ; 2017:843–852.
24. Hestness J, Narang S, Ardalani N, et al. Deep Learning Scaling is Predictable, Empirically. *arXiv [csLG]*. Published online December 1, 2017. http://arxiv.org/abs/1712.00409
25. Heilemann G, Matthewman M, Kuess P, et al. Can Generative Adversarial Networks help to overcome the limited data problem in segmentation? *Z Med Phys*. 2022;32(3):361-368.
26. Dong X, Lei Y, Wang T, et al. Automatic multiorgan segmentation in thorax CT images using U-net-GAN. *Med Phys*. 2019;46(5):2157-2168.