Decentralized Distributed Multi-institutional PET Image Segmentation Using a Federated Deep Learning Framework

Shiri, Isaac; Sadr, Alireza Vafaei; Amini, Mehdi; Salimi, Yazdan; Sanaat, Amirhossein; Akhavanallaf, Azadeh; Razeghi, Behrooz; Ferdowsi, Sohrab; Saberi, Abdollah; Arabi, Hossein

Published in:
Clinical Nuclear Medicine

DOI:
10.1097/RLU.0000000000004194

Publication date:
2022

Decentralized Distributed Multi-institutional PET Image Segmentation Using a Federated Deep Learning Framework

Isaac Shiri, MSc, Alireza Vafaei Sadr, MSc,†‡ Mehdi Amini, MSc,* Yazdan Salimi, MSc,* Amirhossein Sanaat, MSc,* Azadeh Akhavanallaf, MSc,* Behrooz Razeghi, PhD,§ Sohrab Ferdowsi, PhD,|| Abdollah Saberi, MSc,* Hossein Arabi, PhD,* Minerva Becker, MD,¶ Slava Voloshynovskiy, PhD,§ Deniz Gündüz, PhD,** Arman Rahmim, PhD,†††‡‡ and Habib Zaidi, PhD*§§||||¶¶*

Purpose: The generalizability and trustworthiness of deep learning (DL)–based algorithms depend on the size and heterogeneity of training datasets. However, because of patient privacy concerns and ethical and legal issues, sharing medical images between different centers is restricted. Our objective is to build a federated DL-based framework for PET image segmentation utilizing a multicentric dataset and to compare its performance with the centralized DL approach.

Methods: PET images from 405 head and neck cancer patients from 9 different centers formed the basis of this study. All tumors were segmented manually. PET images converted to SUV maps were resampled to isotropic voxels ($3 \times 3 \times 3$ mm^3) and then normalized. PET image subvolumes ($12 \times 12 \times 12$ cm^3) consisting of whole tumors and background were analyzed. Data from each center were divided into train/validation (80% of patients) and test sets (20% of patients). The modified R2U-Net was used as core DL model. A parallel federated DL model was developed and compared with the centralized approach where the data sets are pooled to one server. Segmentation metrics, including Dice similarity and Jaccard coefficients, percent relative errors (RE%) of SUV_{peak}, SUV_{mean}, SUV_{median}, SUV_{max}, metabolic tumor volume, and total lesion glycolysis were computed and compared with manual delineations.

Results: The performance of the centralized versus federated DL methods was nearly identical for segmentation metrics: Dice (0.84 ± 0.06 vs 0.84 ± 0.05) and Jaccard (0.73 ± 0.08 vs 0.73 ± 0.07). For quantitative PET parameters, we obtained comparable RE% for SUV_{mean} ($6.43\% \pm 4.72\%$ vs $6.61\% \pm 5.42\%$), metabolic tumor volume ($12.2\% \pm 16.2\%$ vs $12.1\% \pm 15.89\%$), and total lesion glycolysis ($6.93\% \pm 9.6\%$ vs $7.07\% \pm 9.85\%$) and negligible RE% for SUV_{max} and SUV_{peak}. No significant differences in performance ($P > 0.05$) between the 2 frameworks (centralized vs federated) were observed.

Conclusion: The developed federated DL model achieved comparable quantitative performance with respect to the centralized DL model. Federated DL models could provide robust and generalizable segmentation, while addressing patient privacy and legal and ethical issues in clinical data sharing.

Key Words: distributed deep learning, federated learning, multicenter studies, PET, segmentation

(*Clin Nucl Med* 2022;47: 606–617)

PET is an established noninvasive imaging modality capturing functional and metabolic information of the underlying tissues at the molecular level. ^{18}F-FDG PET imaging plays a major role for improved clinical diagnosis, evaluation of prognosis, treatment planning including external radiation therapy (RT), and for posttreatment follow-up.[1] Radiation therapy is a standard treatment modality of head and neck cancer (HNC),[2] and the precise delineation of tumor boundaries is crucial as segmentation accuracy not only affects survival of HNC patients but is also essential in avoiding irradiation of organs at risk.[3] A number of studies have demonstrated that the ability of ^{18}F-FDG PET to characterize tumor metabolism facilitates its segmentation for RT planning.[4] Currently, tumor segmentation is performed manually by the radiation oncologist. This task, however, is time-consuming and labor-intensive and also crucially suffers from interobserver and intraobserver variability because of the complex HN anatomy on the one hand and the required considerable operator experience on the other hand.[5] However, even in experienced hands, interobserver variability can be substantial. In a recent study,[6] the interobserver variability of PET/CT-based gross target volume (GTV) segmentation in HNC patients undergoing RT resulted in a mean GTV overlap, as reflected by the Dice similarity coefficient, of only 69%, although 3 experienced radiation oncologists had manually segmented the tumors. To segment HNCs, data from ^{18}F-FDG PET and CT acquisitions are usually used.[7] In this approach, it is assumed that ^{18}F-FDG tumor uptake and anatomic tumor boundaries correspond on coregistered PET and CT images.[8] However, anatomic and metabolic tumor boundaries may not coincide because of PET/CT mis-coregistration errors and peritumoral inflammation, which may lead to overestimation of tumor volume on morphologic images.[9]

In addition, accurate and precise delineation of tumor contours plays a critical role for the reliability of quantitative analysis of ^{18}F-FDG uptake, including texture analysis. Such analysis (referred to as radiomics) can be utilized to evaluate tumor changes during treatment and establish prognostic models for predicting survival and treatment outcome.[10] It has been shown that the variability of tumor contouring can jeopardize the robustness and reproducibility of quantitative metrics including radiomics features extracted from PET images.[10] Moreover, tumor segmentation has been recognized as a

Received for publication January 18, 2022; revision accepted February 12, 2022.

From the *Division of Nuclear Medicine and Molecular Imaging, Geneva University Hospital; †Department of Theoretical Physics and Center for Astroparticle Physics, University of Geneva, Geneva, Switzerland; ‡Institute of Pathology, RWTH Aachen University Hospital, Aachen, Germany; §Department of Computer Science, and ||HES-SO, University of Geneva, Geneva; ¶Division of Radiology, Geneva University Hospital, Geneva, Switzerland; **Faculty of Engineering, Department of Electrical and Electronic Engineering, Imperial College London, London, United Kingdom; ††Department of Radiology and Physics, University of British Columbia, Vancouver, BC; ‡‡Department of Integrative Oncology, BC Cancer Research Institute, Vancouver, BC, Canada; §§Geneva University Neurocenter, University of Geneva, Geneva, Switzerland; ||||Department of Nuclear Medicine and Molecular Imaging, University of Groningen, University Medical Center Groningen, Groningen, The Netherlands; and ¶¶Department of Nuclear Medicine, University of Southern Denmark, Odense, Denmark.

Conflicts of interest and sources of funding: This work was supported by the Swiss National Science Foundation under grants SNSF 320030_176052 and SNSF 320030_173091/1.

Correspondence to: Habib Zaidi, PhD, Division of Nuclear Medicine and Molecular Imaging, Geneva University Hospital, Rue Gabrielle-Perret-Gentil 4, CH-1211 Geneva, Switzerland. E-mail: habib.zaidi@hcuge.ch.

Supplemental digital content is available for this article. Direct URL citation appears in the printed text and is provided in the HTML and PDF versions of this article on the journal's Web site (www.nuclearmed.com).

ISSN: 0363-9762/22/4707–0606
DOI: 10.1097/RLU.0000000000004194

stumbling block and a time-consuming step in radiomics studies,[11] hindering the utilization of large multicenter data sets, an essential step for successful clinical implementation of texture analysis.[12]

Apart from the aforementioned reasons, newly developed RT techniques and radiomics models based on PET alone, without relying on other imaging modalities, hold promise for fully automated segmentations. Therefore, the development of accurate fully automated segmentation methods is highly desirable. Several artificial intelligence (AI) techniques, especially deep learning (DL) algorithms, have already been used for various tasks in nuclear medicine image analysis, including classification, dosimetry, image-to-image translation, and image segmentation.[13–15] An array of DL techniques have been developed for the task of medical image segmentation and proven to produce promising results for different modalities, especially PET.[16] Among the different AI techniques, DL-based methods have gained special attention because of their ability of automatically extracting high-throughput features and generating probability maps to segment and delineate normal and abnormal tissues.[17] In a PET segmentation study by Czakon et al,[18] 3 different AI-based methods, namely, a model based on spatial distance weighted fuzzy c-means, another based on dictionary learning, and a DL approach, were compared. The DL approach achieved the highest performance. Moreover, the MICCAI (Medical Image Computing and Computer Assisted Intervention)[19] and HECKTOR (HEad and neCK TumOR)[20] automated segmentation challenges, the best performing models, were all DL-based. Nevertheless, only a limited number of studies have so far investigated the potential of DL segmentation methods based solely on PET.[21] In a more recent study, 3 DL algorithms with a combination of 8 loss functions were assessed for HNC tumor segmentation from PET images, reporting promising results.[22] However, DL-based algorithms are known to be data-hungry, and as such, their generalizability is largely dependent on the size and the heterogeneity of the used datasets.

In a wide range of medical situations, this becomes a critical challenge, because sharing medical data between different centers is limited because of concerns over patients' privacy and ethical and legal issues.[23–25] To overcome this limitation, the concept of federated learning (FL) is being increasingly explored in the context of medical data and more recently in medical imaging. The idea is to build DL-based models and learning recipes to be applied on multi-institutional data sets without sharing the data between the centers to address patients' privacy and data sharing issues.[23–25]

Traditionally, DL models are developed in a single center, where the data owner trains the application-specific model using available local training data sets. However, this approach has 2 major limitations. First, the development of an accurate and robust DL model requires massive data sets, which are unlikely to be obtained from a single center. Second, data acquired in a single center may be homogeneous, resulting in poor generalizability and poor performance on independent unseen samples. To address these limitations, data owners (users) send their data sets to a central server, having significant computational power and storage capacity, to pool data for a meaningful model implementation. This approach is known as the centralized framework. In the last decade, an abundant range of applications have been proposed based on standalone and centralized frameworks. However, data sets often contain sensitive information that data owners (users) may prefer to keep private. Hence, sharing medical data between different centers is often limited because of concerns regarding patient privacy and ethical and legal issues.[23–25]

Privacy-preserving mechanisms can be utilized to ensure privacy of the data owners (users). Alternatively, we can train a centralized model in a decentralized fashion. The FL paradigm proposes approaches to update such decentralized models.[26–32] Alternatively, models can be trained using distributed or decentralized approaches. In the distributed framework, first proposed by Lu et al[28] and McMahan et al,[33] data owners collaborate by sharing their local model's updates to train a common model.[29–32] In the FL framework, data owners (users) train their networks locally and send the trained model updates to a central server. Then, the server can, for instance, iteratively aggregate the local updates into a global network. As a result, FL allows training collaborative DL models using localized data from different centers, addressing privacy concerns to some extent. Federated learning–based models have been developed for different medical imaging tasks, including abnormality detection and classification,[34] prognostic modeling,[35] and segmentation.[36,37] In the current study, we have developed a federated DL-based framework for PET image segmentation using multicenter data sets and compared its performance with a centralized model, which uses data pooling in a central server for model building.

PATIENTS AND METHODS

PET/CT Data Acquisition and Description

In this study, high-quality artifact-free anonymized PET images of 405 HNC patients from 9 different centers were used. The numbers of included patients were 60, 75, 58, 22, 20, 68, 32, 41, and 29 from centers 1 to 9, respectively. Data were acquired and reconstructed using different scanners and protocols; the information regarding data sets can be found in References [22,38–43]. The data from each center were divided into train/validation set (80% patients, 324 patients) and test set (20% patients, 81 patients).

Manual Image Segmentation and Preprocessing

Image quality of all PET images was first evaluated by an experienced nuclear medicine physician. Subsequently, manual segmentation of primary tumors was performed for each center on axial slices, starting from initial segmentation provided by each center, by an experienced nuclear medicine physician or in consensus by 2 radiologists, depending on the center. The delineations were used as standard of reference for evaluation. PET images were converted to SUV maps. Because the data sets were acquired at different centers with different scanners, image acquisition protocols, and reconstruction settings, PET images were interpolated to an isotropic voxel of $3 \times 3 \times 3$ mm^3, which resulted in rotationally invariant uniform (matrix size and voxel size) data sets. In addition, in order to make the computations tractable, all PET images were cropped to $12 \times 12 \times 12$-cm^3 subvolumes (uniform matrix size and voxel size) including whole tumor and background. Cropped PET images were normalized to the range [0,1]. These straightforward steps were adopted for easy implementation and to ensure reproducibility of image preprocessing in clinical setting.

Federated Learning Framework

Consider a centralized server that aims to train a DL model consisting of d parameters based on data samples available at K data centers (owners/users). The objective is to minimize some loss function $\mathcal{L}(.;.)$. Let $\left\{x_i^{(k)}, y_i^{(k)}\right\}_{i=1}^{N_k}$ denote the set of N_k labeled training data samples available at the kth data center (owner), where $k \in \{1,2,\ldots, K\}$. The vector $x_i^{(k)}$ represents the ith PET image at center k, where $i \in \{1, 2, ..., N_k\}$, and $y_i^{(k)}$ represents the corresponding label. The goal of the FL model is to find the vector θ° satisfying:

$$\theta^\circ = \underset{\theta}{\operatorname{argmin}} \left(F(\theta) \triangleq \sum_{k=1}^{K} \alpha_k F_k(\theta) \right), \quad (1)$$

where $F_k(\theta)$, $k \in \{1,\ldots,K\}$ are the local objective functions, and $\alpha_k \in \{1,\ldots,K\}$ are nonnegative weighting coefficients satisfying $\sum_{k=1}^{K} \alpha_k = 1$. Let us consider the collection of centers (hospitals)

to have a total of $N = \sum_{k=1}^{K} N_k$ data samples; then one can assign $\alpha_k = N_k / N$. The local objective functions are defined as empirical averages over the associated training sets as follows:

$$F_k(\theta) \triangleq \frac{1}{N_k} \sum_{i=1}^{N_k} \mathcal{L}\left(\theta; \left(x_i^{(k)}, y_i^{(k)}\right)\right). \quad (2)$$

In our framework, each center (data owner) minimizes its empirical loss function with respect to the local data samples, using the stochastic gradient descent (SGD) algorithm. Let t denote the global iteration and suppose each data owner performs a $\tau-$ step local SGD, for some $\tau \in N$. Upon receiving the global model parameter $\theta(t)$ from the server, the $|$th step of the local SGD at data owner k, $k \in \{1,2,\ldots,K\}$, corresponds to the following update:

$$\theta_k^{j+1}(t) = \theta_k^j(t) - \eta_k^j(t) \nabla F_k\left(\theta_k^j(t)\right), \quad j \in \{1, 2, \ldots, \tau\}, \quad (3)$$

where $\eta_k^j(t)$ denotes the learning rate, and the first local update is set as $\theta_k^1(t) = \theta(t)$. Depending on the strategy to update the parameters of the global and local models, various techniques have been proposed[44–46] to optimize the communication efficiency compared with the naive SGD method. We used federated averaging (*FedAvg*) in our framework. The schematic description of the FL process is presented in Figure 1. First, the global model developed by the server distributes data through different centers (A). Next, the models are trained separately in each center (B) using the local data set, and finally, trained models from all centers are returned to the server to aggregate and update the central global model (C). These steps are repeated until some convergence criteria are met, for example, until no significant loss descend is observed. The model learns from the data sets using SGD for all optimizations.[22] Similar to previous studies,[47–53] the FL process in this work was performed on a server with multiple local graphics processing units (GPUs), where each local GPU was considered as a different center.

Deep Neural Network

As for the DL architecture, we used a modified R2U-Net,[54] which is composed of recurrent residual connections as well as convolutional blocks (Fig. 1). The most established neural structure for image segmentation in the medical community is U-Net, based on which many further variations have been proposed. R2U-Net builds on top of this by adding recurrence to the convolutional residual blocks, which helps the network to increase its effective capacity without increasing the number of parameters. In fact, recurrence can be thought of as the operation of unrolling a network block through

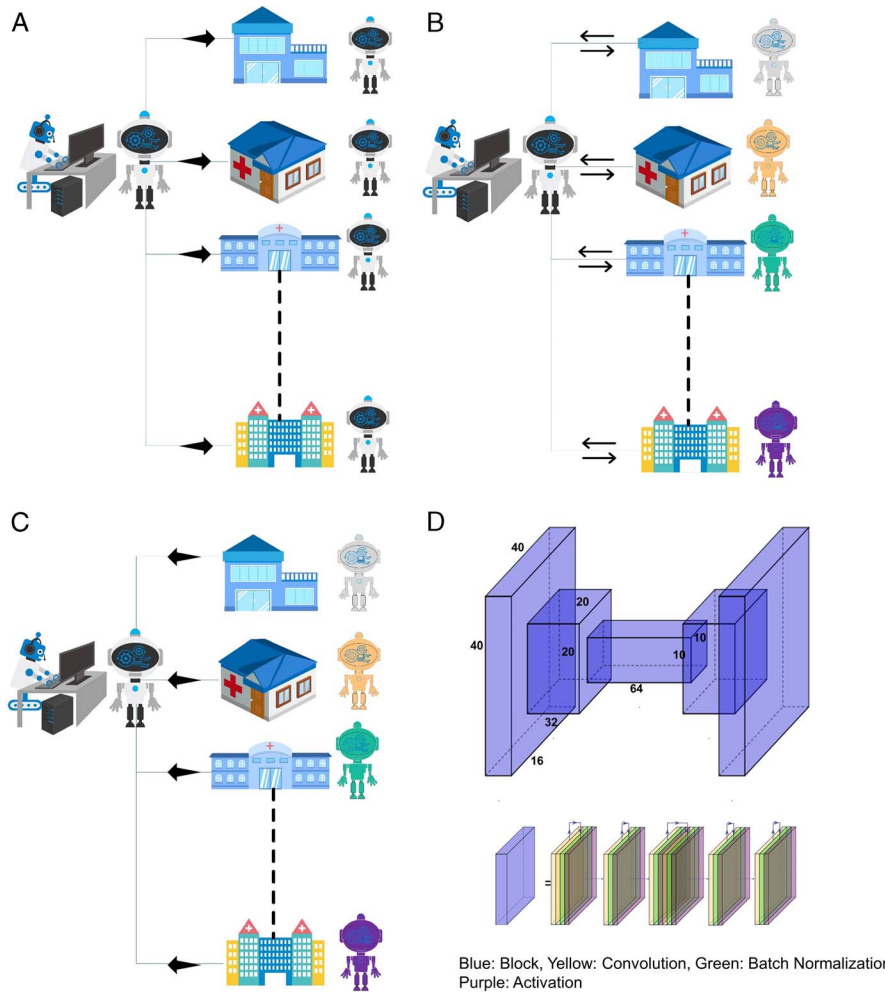

FIGURE 1. Federated learning flowchart. The global model generated by the server is distributed to different centers (**A**). Models are trained separately in each center (**B**). Trained models from all centers are returned to the server to be aggregated to form a new global model (**C**). Schematic of deep recurrent residual neural network (R2U-Net) (**D**).

time to provide more effective depth. Moreover, R2U-Net uses feature accumulation, which helps extracting low-level features. We used 3 down- and up-sampling levels with 16, 32, and 64 channels in our R2U-Net structure, as well as 2 recurrent convolutional layers with 2 iterations per down- and up-sampling, along with the batch normalization layers. As for the activation function, we used the standard ReLU, except for the sigmoid in the output layer. All implementations were performed in TensorFlow.

Training

All evaluations and reports were performed for 81 patients (20% of each center). All PET images were fed as input to the R2UNet in both the FL and centralized frameworks to generate the corresponding binary masks of tumors. We trained the DL model with axial slices as 1-channel images with a batch size of 64.

Quantitative Evaluation

To evaluate the performance of the 2 models, standard segmentation metrics, including the Dice similarity coefficient, Jaccard similarity coefficient, false-negative rate, false-positive rate, volume similarity, and mean and SD of surface distance were calculated with respect to manual segmentations. In addition, clinical evaluation of DL-guided segmentations using both centralized and FL frameworks was assessed through a number of image-derived PET metrics, including SUV_{peak}, SUV_{mean}, SUV_{median}, SUV_{max}, metabolic tumor volume (MTV), and total lesion glycolysis (TLG). In addition, we extracted a number of shape radiomic features, including sphericity, asphericity, elongation, and flatness using SERA package,[55] which is compliant with the Image Biomarker Standardization Initiative guidelines.[56] We calculated the mean relative error (RE%) and the mean absolute relative error (ARE%) with respect to manual segmentation.

Statistical Analysis

Descriptive statistics included mean ± SD and 95% confidence interval (CI) for different image quantification metrics. The Kolmogorov-Smirnov test showed that the data were not normally distributed. Therefore, pairwise comparison between parameters was performed using the nonparametric 2-sample Wilcoxon test

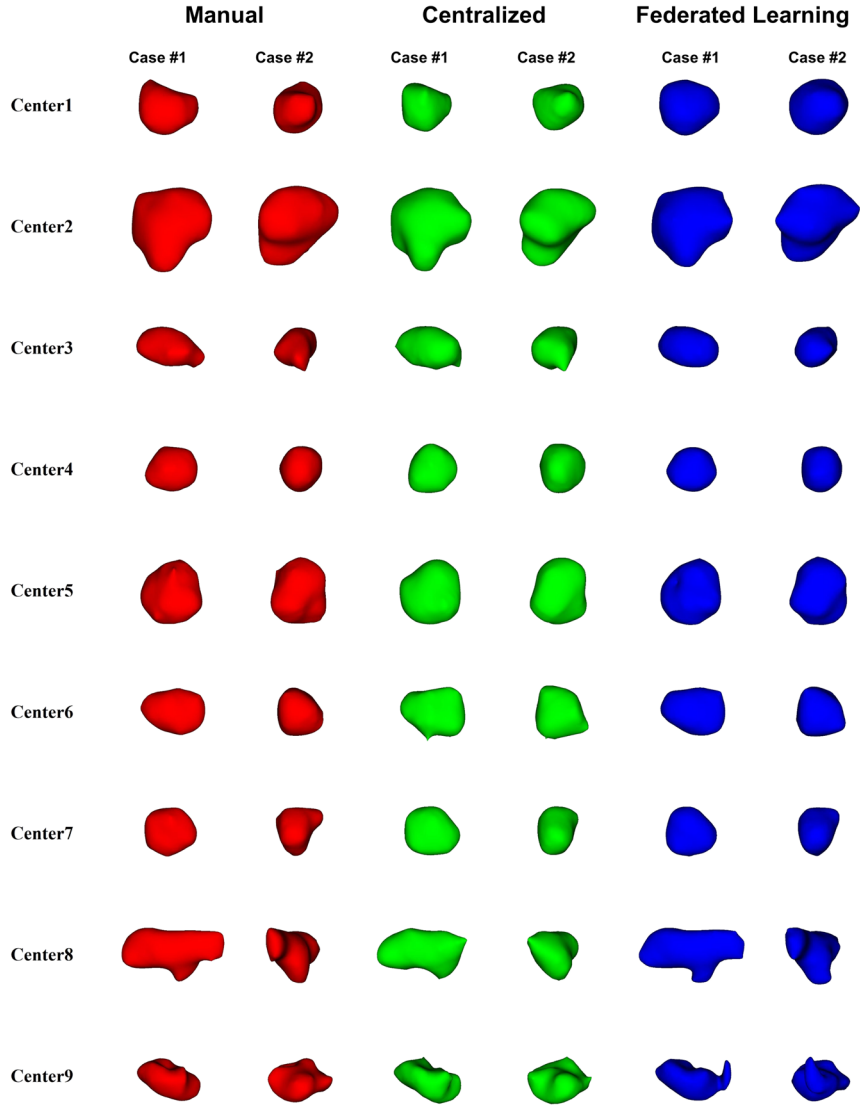

FIGURE 2. 3D views of PET segmentations obtained from manual (red), centralized learning (green), and federated learning (blue) methods, on representative patients from different centers (cases 1 and 2 are from 2 centers).

(Wilcoxon rank sum test or Mann-Whitney U test) with $P < 0.05$ defined as threshold for statistical significance.

RESULTS

Figure 2 illustrates the 3D-rendered volumes of segmented GTVs, sampled from different clinical centers for manual, centralized, and federated segmentation approaches. The model performance of centralized (green) and federated DL algorithm (blue) is visually compared against manual segmentation. Supplemental Figure 1 (http://links.lww.com/CNM/A378) represents additional cases of segmented GTVs categorized by clinical center.

Figure 3 presents a visual comparison between the 2 different learning strategies against manual segmentation (ground truth) via multiple 2D axial views of different lesions from each center. A magnified version of the same GTVs is illustrated in Figure 4. In Supplemental Figures 2–10 (http://links.lww.com/CNM/A378), different cases from various centers are shown, depicting the accuracy and complexity of the segmentation task according to textural GTV characteristics.

In separate center-by-center analyses, strong consistency was observed between the 2 approaches, in terms of quantitative image segmentation performance metrics. In Figure 5, model performance within centralized versus FL approaches is compared in terms of Dice similarity coefficient (0.84 ± 0.06 vs 0.84 ± 0.05), Jaccard similarity coefficient (0.73 ± 0.08 vs 0.73 ± 0.07), false-negative (0.16 ± 0.12 vs 0.17 ± 0.11), false-positive (0.14 ± 0.1 vs 0.11 ± 0.1), mean surface distance (0.19 ± 0.08 vs 0.19 ± 0.07), and SD surface distance (0.44 ± 0.12 vs 0.45 ± 0.14). Mann-Whitney U statistical analysis showed no significant differences ($P > 0.05$) between centralized compared with FL approach for almost all quantitative metrics. We provide additional details, including mean ± SD, 95% CI, and P value tables in Supplemental Tables 1–3 (http://links.lww.com/CNM/A378), respectively.

In addition, through center-based analysis, we observed consistency between centralized and FL models in terms of relative bias from ground truth segmentation in SUV_{mean} (6.43% ± 4.72% vs 6.61% ± 5.42%), MTV (12.23%–16.19% vs 12.1%–15.89%), and TLG (6.93%–9.6% vs 7.07%–9.85%). For some indices, such as SUV_{max} and SUV_{peak}, the difference between the prediction and ground truth could be ignored. In the case of shape features, almost the same consistency was observed between centralized versus FL (elongation = 5.5%–7.43% vs 5.67%–8.19% and sphericity = 3.39%–4.6% vs 3.81%–5.35%). Table 1 summarizes the mean ± SD of RE% and ARE% of quantitative PET metrics between centralized and FL

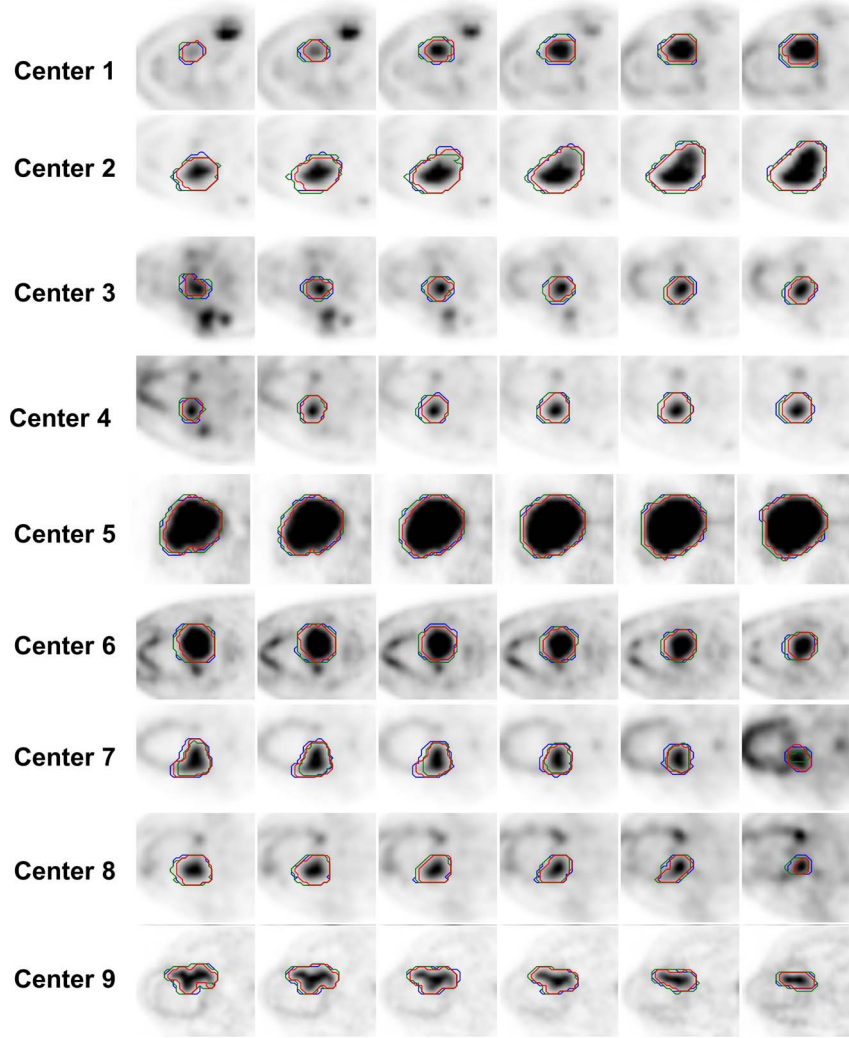

FIGURE 3. 2D views of segmentations obtained from manual (red), centralized learning (green), and federated learning (blue) methods on representative patients from the 9 different centers.

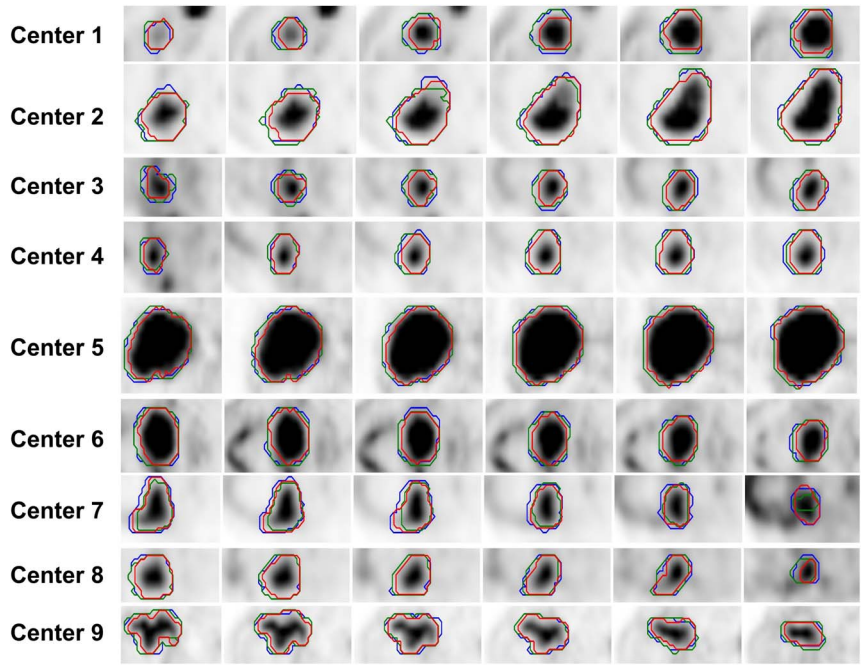

FIGURE 4. Magnified 2D views of segmentations obtained from manual (red), centralized learning (green), and federated learning (blue) methods on patients from the 9 different centers.

approaches compared with the ground truth using the center-based approach. Center-based statistical analysis revealed that differences between all derived quantitative metrics were not significant ($P > 0.05$). Selected image-derived features (signal intensity features and shape features) of the segmented volumes of the whole test set are illustrated in Figure 6. Furthermore, the AREs of these metrics between centralized and FL models are summarized in Table 2, as categorized by center. Lower and upper bands of 95% CI are summarized in Supplemental Tables 4 and 5 (http://links.lww.com/CNM/A378) along with statistical analysis, in which no significant differences between the 2 approaches were observed (Supplemental Tables 6 and 7, http://links.lww.com/CNM/A378).

DISCUSSION

Accurate and reproducible tumor segmentation from noisy PET images faces many challenges.[57,58] Inherent limitations in PET imaging, such as low spatial resolution, partial volume effect, high noise characteristics, and motion artifacts, result in blurred boundaries between tumor and background. In addition, different shapes, textures, and locations of tumors render the development of generalized segmentation methods difficult. Furthermore, the variability of PET scanners, imaging protocols, and reconstruction/correction algorithms challenges the reproducibility of segmentation results.[9]

Multiple computer-aided methods have been proposed for PET image segmentation that successfully addressed the aforementioned challenges to some extent.[58] Conventional segmentation methods range from simple algorithms, such as threshold-based,[59] region-growing, and active contours, to more sophisticated approaches based on clustering and classification algorithms trained on PET features, such as fuzzy locally adaptive Bayesian, atlas-based, fuzzy c-mean iterative clustering and Gaussian mixture models. Although these algorithms provided promising results, translation into the clinic faced multiple impediments. Some techniques require manual identification of the central tumor voxel or a bounding box encompassing the entire tumor,[60] some are limited by partial volume effect,[61] and some require additional tuning on different scanners.[62]

Compared with the aforementioned methods, DL-based methods have shown promising results. In the first MICCAI PET segmentation challenge,[19] the performance of conventional and machine learning algorithms was evaluated on a dataset of 176 PET images consisting of simulated, phantom, and clinical studies. Deep learning–based algorithms outperformed other techniques, achieving a Dice score of 0.80. Huang et al[63] applied a U-Net architecture for HNC segmentation from PET/CT images on dual-center datasets utilizing 22 patient studies evaluated using one-leave-out scheme and reported a Dice coefficient of 0.73. Andrearczyk et al[64] segmented HNC tumors using the V-Net architecture and evaluated their model using one-center-leave-out in a 4-center database and reported Dice coefficients of 0.58 and 0.60 for PET and fused PET/CT images, respectively. Leung et al[65] proposed a physics-guided tumor segmentation method from PET images using DL. They simulated realistic tumors and trained the model based on these data followed by fine-tuning on clinical datasets. They reported a Dice coefficient of 0.87 (95% CI, 0.86–0.88) and 0.73 (95% CI, 0.71–0.76) for simulated and clinical studies, respectively.

In a more recent study, Shiri et al[22] developed a fully automated tumor segmentation from HNC PET studies using DL algorithms and multicentric datasets. They evaluated 24 different models implemented through the combination of 3 DL algorithms and 8 different loss functions. Deep learning models were trained on 370 images and tested on 100 PET images on 12-cm^3 subvolumes that included both tumor and background. They reported a Dice coefficient (mean ± SD and 95% CI) for Res-Net with cross-entropy loss (0.86 ± 0.05 and 0.85–0.87), Dense-VNet with cross-entropy loss (0.85 ± 0.058 and 0.84–0.86), and NN-UNet with Dice plus XEnt (0.87 ± 0.05 and 0.86–0.88). There were no statistically significant differences between the 3 models for various quantitative segmentation metrics. In addition, they reported an RE% <5% for SUV_{max}, SUV_{mean}, and SUV_{median} in NN-UNet with Dice plus XEnt model.[22]

Despite the potential of DL-based segmentation models, their performance depends highly on the specific datasets used for

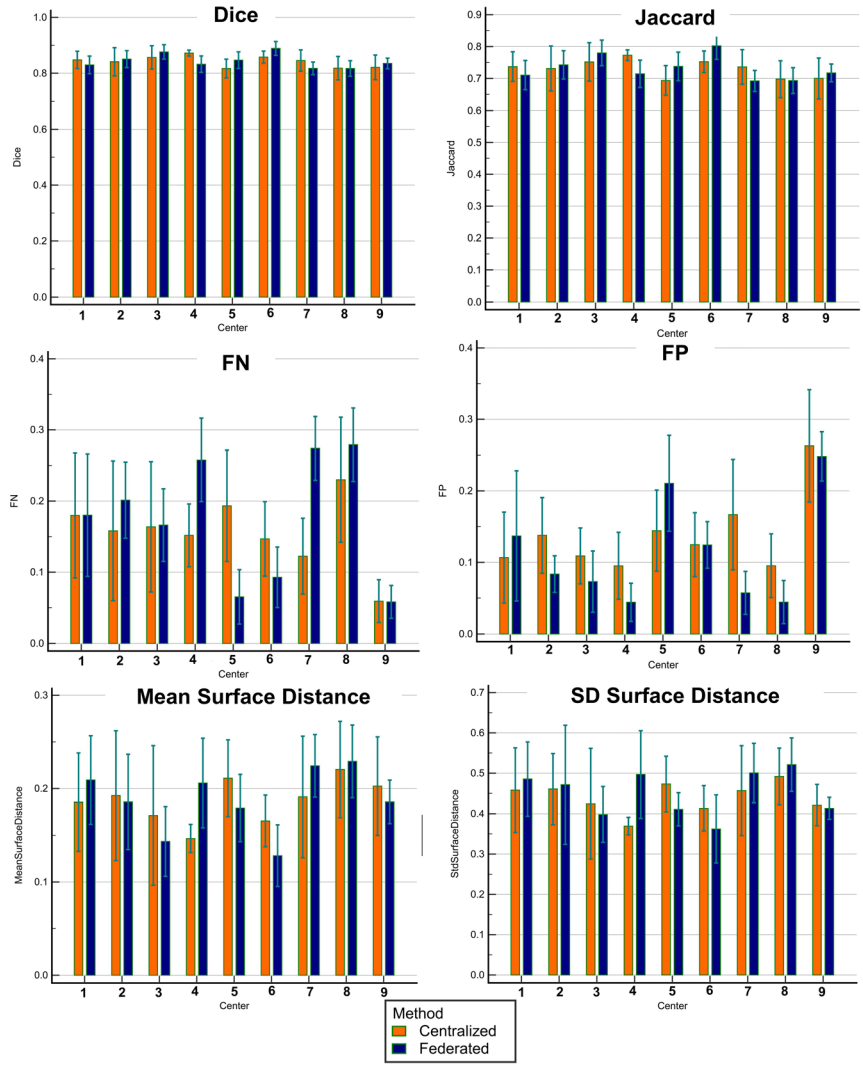

FIGURE 5. Comparison of the performance of the centralized versus federated learning frameworks in terms of quantitative metrics.

training. These algorithms require large/heterogeneous data sets to provide robust and generalizable models. Creating large data sets for data-hungry DL models requires collaboration among different centers. Meanwhile, owing to legal/ethical and privacy issues, direct data sharing between centers is not always feasible. The FL framework can address these challenges by providing decentralized training procedures for DL models. This approach preserves privacy and paves the way to train DL models collaboratively on large multicentric data sets without sharing data sets between centers.[23–25]

In the current study, we compared the performance of centralized and FL models for the segmentation of HNC PET images. Overall, a high consistency was observed between centralized and FL approaches in terms of quantitative image segmentation metrics, including Dice coefficient (0.84 ± 0.06 vs 0.84 ± 0.05) and Jaccard coefficient (0.73 ± 0.08 vs 0.73 ± 0.07). In terms of conventional PET image-derived quantitative metrics, consistency between FL versus centralized approach was confirmed with SUV_{mean} (6.43% ± 4.72% vs 6.61% ± 5.42%) and TLG (6.93%–9.6% vs 7.07%–9.85%). For SUV_{max} and SUV_{peak}, RE% and ARE% were almost zero. Overall, statistical analysis showed no significant differences ($P > 0.05$) between these 2 strategies for different quantitative metrics. Collaborative DL model training without sharing data sets between different hospitals and centers to preserve patients' privacy using FL has been reported in few studies.[23–25] Dayan et al[35] developed an FL-based model for oxygen requirements in COVID-19 patients using vital, laboratory, and chest x-ray images. They reported 16% and 38% improvement in average area under the curve and generalizability for FL-based models compared with center-based models. Gawali et al[66] reported an area under the receiver operating characteristic curve/F1 score of 0.95/0.72 and 0.93/0.62 with centralized and FL models for chest x-ray classification.

In a recent study by Feki et al,[53] FL-based models were evaluated for COVID-19 detection from chest x-ray images. They evaluated 2 different DL architectures for centralized and FL frameworks with different settings and reported that the FL-based method can achieve comparable results with respect to centralized methods and remain robust in the presence of not independent and identically distributed and unbalanced data. In another study by Lee et al,[67] an FL framework was tested for thyroid nodules malignancy classification using ultrasound images. They enrolled 8457 ultrasounds images from 6 different centers and compared the performance of 5 different DL-based networks. They reported areas under the receiver operating characteristic curve of 78.88% to 87.56% and 82.61% to 91.57% for FL and centralized-based learning methods,

TABLE 1. Summary of RE(%) in Quantitative PET Metrics (Mean ± SD) for Centralized and Federated learning models Among the Different Centers

		MTV	Sphericity	Asphericity	Elongation	Flatness	SUV_{peak}	SUV_{mean}	SUV_{median}	SUV_{max}	TLG
Center 1	Centralized	9.79 ± 16.02	−5 ± 7.23	18.76 ± 26.98	−3.64 ± 9.94	−15.2 ± 9.36	0 ± 0	−5.17 ± 7.82	−6.71 ± 10.41	0 ± 0	6.21 ± 10.16
	Federated	2.07 ± 13.18	−4.86 ± 7.69	16.97 ± 25.47	−5.26 ± 11.08	−11.75 ± 12.76	0 ± 0	−1.77 ± 6.03	−1.49 ± 8.74	0 ± 0	1.05 ± 6.95
Center 2	Centralized	12.07 ± 11.84	−2.59 ± 3.02	12.26 ± 12.96	−0.49 ± 11.38	−3.98 ± 12.85	0 ± 0	−5.05 ± 4.82	−6.51 ± 6.35	0 ± 0	8.12 ± 8.6
	Federated	9.34 ± 0.03	−1.04 ± 3.59	4.94 ± 14.83	−8.03 ± 9.19	−1.58 ± 5.66	0 ± 0	−4.35 ± 1.75	−6.27 ± 4.3	0 ± 0	5.39 ± 1.55
Center 3	Centralized	9.52 ± 8.11	−3.15 ± 4.31	14.84 ± 17.79	1.58 ± 7	−1.99 ± 6.39	0 ± 0	−5.79 ± 4.85	−8.59 ± 7.58	0 ± 0	4.6 ± 4.76
	Federated	14.48 ± 14.26	−4.53 ± 2.69	18.24 ± 10.95	−0.13 ± 5.97	−4.68 ± 13.61	0 ± 0	−7.78 ± 8.2	−11.02 ± 11.33	0 ± 0	8.8 ± 9.51
Center 4	Centralized	15.55 ± 13.46	−5.63 ± 3.38	26.49 ± 9.28	−2.01 ± 8.24	−5.54 ± 9.64	0 ± 0	−8.09 ± 7.04	−11.98 ± 10.8	0 ± 0	9.52 ± 9.03
	Federated	9.61 ± 13.12	−4.72 ± 4.05	18.59 ± 15.44	−0.2 ± 7.88	−3.22 ± 7.48	0 ± 0	−5.17 ± 6.69	−7.25 ± 8.29	0 ± 0	5.72 ± 7.43
Center 5	Centralized	−4.71 ± 16.99	−1.77 ± 4.51	8.26 ± 23.07	−1.45 ± 6.25	−2.08 ± 8.04	0 ± 0	2.22 ± 5.53	3.5 ± 8.06	0 ± 0	−1.55 ± 11.56
	Federated	−7.15 ± 18.1	−2.12 ± 3.77	9.44 ± 17.88	2.36 ± 8.09	−3.21 ± 6.69	0 ± 0	2.78 ± 6.33	4.05 ± 9.01	0 ± 0	−3.15 ± 11.32
Center 6	Centralized	−4.36 ± 11.17	−3.16 ± 2.53	13.94 ± 9.58	−3.04 ± 5.56	−6.59 ± 7.14	0 ± 0	2.19 ± 3.83	2.88 ± 5.76	0 ± 0	−1.69 ± 7.21
	Federated	1.25 ± 15.99	−2.34 ± 2.93	9.65 ± 12.43	−1.67 ± 5.87	−2.22 ± 8.7	0 ± 0	−1.44 ± 7.01	−2.52 ± 10.37	0 ± 0	0.82 ± 9.88
Center 7	Centralized	14.13 ± 16.02	−2.62 ± 3.95	10.38 ± 16.67	−2.09 ± 9.32	−3.43 ± 10.74	0 ± 0	−7.48 ± 7.39	−10.48 ± 10.87	0 ± 0	8.74 ± 10.68
	Federated	9.5 ± 7.75	−3.23 ± 4.47	10.72 ± 18.07	−3.74 ± 9.38	−3.45 ± 8.72	0 ± 0	−4.51 ± 4.61	−6.52 ± 7.16	0 ± 0	5.73 ± 4.3
Center 8	Centralized	4.03 ± 14.06	−3.72 ± 3.33	16.81 ± 12.55	−1.23 ± 7.36	−5.21 ± 9.77	0 ± 0	−2.1 ± 6.94	−2.63 ± 10.3	0 ± 0	2.89 ± 8.66
	Federated	12.12 ± 16.9	−4.07 ± 3.96	17.79 ± 16.77	−3.14 ± 6.67	−5.34 ± 12.46	0 ± 0	−7.89 ± 7.03	−11.16 ± 10.4	0 ± 0	6.12 ± 12.63
Center 9	Centralized	−18.41 ± 16.48	−2.31 ± 2.1	10.08 ± 10.16	−1.14 ± 8.37	−4.27 ± 10.66	0 ± 0	7.41 ± 7.42	9.26 ± 10.05	0 ± 0	−8.56 ± 7.79
	Federated	−6.49 ± 17.5	−2.69 ± 8.95	4.39 ± 21.98	−7.68 ± 14.75	−24.93 ± 57.03	0 ± 0	6.1 ± 10.74	7.76 ± 13.16	0.08 ± 0.27	0.54 ± 15.56
Overall	Centralized	3.62 ± 17.1	−3.22 ± 3.9	14.24 ± 16.27	−1.56 ± 7.98	−5.09 ± 9.8	0 ± 0	−2.08 ± 7.72	−2.95 ± 10.94	0 ± 0	2.88 ± 10.33
	Federated	4.03 ± 16.5	−3.39 ± 4.97	12.78 ± 17.7	−2.34 ± 9.13	−6.76 ± 21.44	0 ± 0	−2.34 ± 8.25	−3.37 ± 11.4	0.01 ± 0.09	2.91 ± 10.62

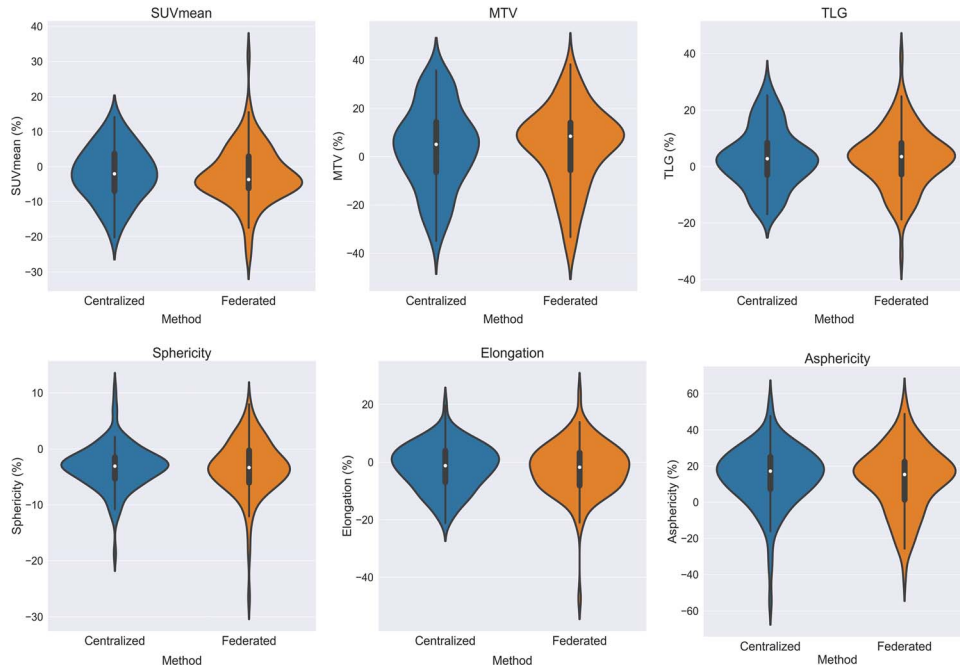

FIGURE 6. Violin plots of RE% for quantitative PET metrics (SUV_{mean}, MTV, and TLG) and radiomics shape features (sphericity, elongation, and asphericity) for centralized versus federated learning models.

respectively. It was concluded that FL-based techniques could potentially achieve the performance of centralized methods in the classification of benign and malignant lesions from ultrasound images. In Li et al,[68] multisite functional MRI analysis was performed using FL and domain adaptation for classification of autism spectrum disorders among healthy control subjects using brain function connectivity. To tackle the domain shift issue, they proposed 2 different methods for FL model performance boosting and showed that FL model performance could be improved by domain adaptation. The proposed method could potentially be implemented in nuclear medicine FL studies for improving models' performance.

In this work, we showed that building DL models from multiple decentralized data sets at multiple centers is possible via FL, where the local data sets remain in the respective centers.[69] Federated learning algorithms bear some inherent limitations. For instance, curious servers may be able to infer local sensitive data sets from trained models, via different types of attacks,[70,71] such as membership inference attacks[72] and model inversion attack.[73] Privacy-aware FL models have recently been introduced to address these additional privacy challenges at the expense of additional computational complexity and performance loss.[36,74,75] Another issue in FL would be malicious parties that could potentially perform data poisoning attacks during the training process,[76,77] that is, modifying the label of data and uploading random updates to the global model.[73,78] Two different categories of noise could arise during training. First was the inherent noise present in the data sets (either in PET images or segmentations) that could potentially cause the learning process to diverge in local centers and affect the whole learning process (in case of high magnitude of noise in the data that the network could not handle). Second, in FL, data poisoning could be induced by malicious parties performing data poisoning attacks during the training process,[76,77] that is, by modifying the labels of the data or by uploading random updates to the global model.[73,78]

Another challenge in FL is how to establish harmony in data preprocessing because it should be performed on data from different centers acquired with different protocols and settings. In the present study, all preprocessing was performed in a uniform manner, including converting to SUV, cropping, and normalizing to provide reproducibility across different centers. One main limitation of the current study was performing all computations on a server with different GPUs simulating different nodes (local computer GPUs were considered as different centers/hospitals) as performed in previous FL studies.[47–53] A number of challenges were linked to the training of the data sets for implementation of the FL approach, such as local computer capacity and communication between centers and local sites. Further studies should be carried out involving real multiple clinical centers (using one-center leave-out strategy) to tackle these challenges, specifically the communication bottleneck. Another limitation of this work is the lack of comparison of DL algorithms with conventional segmentation techniques. Yet, the main aim of the current study was the comparison of the FL approach with centralized training. In the current study, the data from each center were randomly divided into train/validation (80% of patients) and test sets (20% of patients) as a standard evaluation method owing to the computational burden associated with alternative, more complex data-splitting methods. Further studies should be performed through cross-validation strategies (ie, 10-fold) to assess the effect of permutations of data splitting on FL learning robustness compared with centralized DL models.

CONCLUSION

We evaluated the performance of a federated DL framework for PET image segmentation to enable robust decentralized learning without directly sharing data among clinical centers. We compared our proposed model with centralized models and achieved similar performance for an array of image segmentation metrics and quantitative PET features. Federated learning–based models provide robust and generalizable segmentation models while addressing the privacy concerns and legal and ethical issues in medical data sharing among clinical centers.

TABLE 2. Summary of ARE (%) in Quantitative PET Metrics (Mean ± SD) for Centralized and Federated Learning Models Within Different Centers

		MTV	Sphericity	Asphericity	Elongation	Flatness	SUV_{peak}	SUV_{mean}	SUV_{median}	SUV_{max}	TLG
Center 1	Centralized	14.98 ± 10.46	6.74 ± 5.37	27.88 ± 15.54	9.08 ± 4.45	15.7 ± 8.38	0 ± 0	7.62 ± 5.04	10.25 ± 6.31	0 ± 0	8.69 ± 7.8
	Federated	11.35 ± 5.94	6.86 ± 5.75	25.76 ± 15.22	10.5 ± 5.57	14.52 ± 9.05	0 ± 0	5.38 ± 2.76	7.6 ± 3.82	0 ± 0	6.03 ± 3
Center 2	Centralized	13.01 ± 10.69	2.76 ± 2.85	13.06 ± 12.08	9.09 ± 6.3	10.76 ± 7.49	0 ± 0	5.33 ± 4.48	6.93 ± 5.84	0 ± 0	8.84 ± 7.79
	Federated	9.34 ± 0.03	2.54 ± 1.47	10.49 ± 6.98	8.03 ± 9.19	4 ± 2.24	0 ± 0	4.35 ± 1.75	6.27 ± 4.3	0 ± 0	5.39 ± 1.55
Center 3	Centralized	10.18 ± 7.13	4.77 ± 1.96	21.75 ± 5.25	6.05 ± 3.17	5.37 ± 3.52	0 ± 0	5.94 ± 4.64	9.05 ± 6.94	0 ± 0	5.12 ± 4.1
	Federated	15.41 ± 13.12	4.66 ± 2.43	18.93 ± 9.54	4.24 ± 3.93	10 ± 9.87	0 ± 0	8.02 ± 7.93	11.4 ± 10.9	0 ± 0	9.47 ± 8.75
Center 4	Centralized	16.82 ± 11.63	5.63 ± 3.38	26.49 ± 9.28	6.95 ± 4.34	8.3 ± 7.11	0 ± 0	8.62 ± 6.29	12.84 ± 9.64	0 ± 0	10.25 ± 8.09
	Federated	14.3 ± 6.98	5.11 ± 3.48	21.53 ± 10.42	6.36 ± 4.2	6.33 ± 4.82	0 ± 0	7.28 ± 3.99	9.72 ± 4.74	0 ± 0	8 ± 4.56
Center 5	Centralized	14.33 ± 9.55	3.98 ± 2.59	20.01 ± 13.2	5.05 ± 3.72	6.51 ± 4.86	0 ± 0	5.1 ± 2.81	7.3 ± 4.55	0 ± 0	8.76 ± 7.33
	Federated	15.81 ± 10.75	3.7 ± 2.09	17.38 ± 9.63	5.64 ± 6.13	5.95 ± 4.24	0 ± 0	5.52 ± 3.97	8.04 ± 5.44	0 ± 0	9.42 ± 6.64
Center 6	Centralized	8.99 ± 7.63	3.16 ± 2.53	13.94 ± 9.58	4.44 ± 4.42	6.76 ± 6.97	0 ± 0	3.47 ± 2.62	5 ± 3.9	0 ± 0	5.23 ± 5.06
	Federated	12.72 ± 9.11	3.04 ± 2.13	13.1 ± 8.36	5.22 ± 2.84	7.04 ± 5.26	0 ± 0	5.55 ± 4.26	8.58 ± 5.92	0 ± 0	7.66 ± 5.93
Center 7	Centralized	17.81 ± 11.35	3.72 ± 2.81	16.57 ± 9.79	7.14 ± 6	6.64 ± 8.94	0 ± 0	8.43 ± 6.18	11.81 ± 9.26	0 ± 0	11.26 ± 7.7
	Federated	10.64 ± 5.92	4.1 ± 3.62	16.56 ± 12.36	7.04 ± 7.02	5.41 ± 7.56	0 ± 0	5.23 ± 3.68	7.56 ± 5.94	0 ± 0	6.1 ± 3.72
Center 8	Centralized	11.84 ± 8	3.96 ± 3.03	17.88 ± 10.84	5.23 ± 5.13	8.51 ± 6.82	0 ± 0	5.77 ± 4.12	8.67 ± 5.71	0 ± 0	6.68 ± 5.99
	Federated	16.96 ± 11.61	4.58 ± 3.3	20.37 ± 13.26	6.66 ± 2.74	9.8 ± 9.11	0 ± 0	8.38 ± 6.4	11.93 ± 9.44	0 ± 0	10.3 ± 9.27
Center 9	Centralized	21.77 ± 11.03	2.6 ± 1.67	12.28 ± 6.96	6.78 ± 4.53	8.65 ± 7.12	0 ± 0	9.53 ± 3.9	12.19 ± 5.6	0 ± 0	9.98 ± 5.61
	Federated	14.35 ± 11.23	6.04 ± 6.93	19.76 ± 8.67	10.58 ± 12.62	25.06 ± 56.97	0 ± 0	8.39 ± 8.88	10.56 ± 10.8	0.08 ± 0.27	10.51 ± 11
Overall	Centralized	14.21 ± 10.09	4 ± 3.09	18.29 ± 11.47	6.46 ± 4.9	8.35 ± 7.19	0 ± 0	6.43 ± 4.72	9.05 ± 6.76	0 ± 0	8.27 ± 6.79
	Federated	13.99 ± 9.54	4.58 ± 3.9	18.69 ± 11.21	6.93 ± 6.34	9.94 ± 20.15	0 ± 0	6.61 ± 5.42	9.33 ± 7.32	0.01 ± 0.09	8.46 ± 7.01

REFERENCES

1. Unterrainer M, Eze C, Ilhan H, et al. Recent advances of PET imaging in clinical radiation oncology. *Radiat Oncol*. 2020;15:88.
2. Alterio D, Marvaso G, Ferrari A, et al. Modern radiotherapy for head and neck cancer. *Semin Oncol*. 2019;46:233–245.
3. Chen AM, Chin R, Beron P, et al. Inadequate target volume delineation and local-regional recurrence after intensity-modulated radiotherapy for human papillomavirus-positive oropharynx cancer. *Radiother Oncol*. 2017;123: 412–418.
4. Zaidi H, Vees H, Wissmeyer M. Molecular PET/CT imaging–guided radiation therapy treatment planning. *Acad Radiol*. 2009;16:1108–1133.
5. Andrearczyk V, Oreiller V, Depeursinge A. Oropharynx detection in PET-CT for tumor segmentation. *Irish Mach Vis Image Proc*. 2020:109–112.
6. Gudi S, Ghosh-Laskar S, Agarwal JP, et al. Interobserver variability in the delineation of gross tumour volume and specified organs-at-risk during IMRT for head and neck cancers and the impact of FDG-PET/CT on such variability at the primary site. *J Med Imaging Radiat Sci*. 2017;48:184–192.
7. Moe YM, Groendahl AR, Tomic O, et al. Deep learning–based auto-delineation of gross tumour volumes and involved nodes in PET/CT images of head and neck cancer patients. *Eur J Nucl Med Mol Imaging*. 2021;48: 2782–2792.
8. Li L, Lu W, Tan S, et al. Variational PET/CT tumor co-segmentation integrated with PET restoration. *IEEE Trans Radiat Plasma Med Sci*. 2020;4:37–49.
9. Iantsen A, Ferreira M, Lucia F, et al. Convolutional neural networks for PET functional volume fully automatic segmentation: development and validation in a multi-center setting. *Eur J Nucl Med Mol Imaging*. 2021;48:3444–3456.
10. Guezennec C, Bourhis D, Orlhac F, et al. Inter-observer and segmentation method variability of textural analysis in pre-therapeutic FDG PET/CT in head and neck cancer. *PLoS One*. 2019;14:e0214299.
11. Hatt M, Le Rest CC, Tixier F, et al. Radiomics: data are also images. *J Nucl Med*. 2019;60:38S–44S.
12. Zwanenburg A. Radiomics in nuclear medicine: robustness, reproducibility, standardization, and how to avoid data analysis traps and replication crisis. *Eur J Nucl Med Mol Imaging*. 2019;46:2638–2655.
13. Bradshaw TJ, Boellaard R, Dutta J, et al. Nuclear medicine and artificial intelligence: best practices for algorithm development. *J Nucl Med*. 2021.
14. Sanaat A, Shooli H, Ferdowsi S, et al. DeepTOFSino: a deep learning model for synthesizing full-dose time-of-flight bin sinograms from their corresponding low-dose sinograms. *Neuroimage*. 2021;245:118697.
15. Sanaat A, Shiri I, Arabi H, et al. Deep learning–assisted ultra-fast/low-dose whole-body PET/CT imaging. *Eur J Nucl Med Mol Imaging*. 2021;48: 2405–2415.
16. Yousefirizi F, Jha AK, Brosch-Lenz J, et al. Toward high-throughput artificial intelligence–based segmentation in oncological PET imaging. *PET Clin*. 2021;16:577–596.
17. Zhong Z, Kim Y, Zhou L, et al. 3D fully convolutional networks for co-segmentation of tumors on PET-CT images. *Proc IEEE Int Symp Biomed Imaging*. 2018;2018:228–231.
18. Czakon J, Drapejkowski F, Zurek G, et al. Machine learning methods for accurate delineation of tumors in PET images. arXiv:161009493. https://doi.org/10.48550/arXiv.1610.09493. 2016.
19. Hatt M, Laurent B, Ouahabi A, et al. The first MICCAI challenge on PET tumor segmentation. *Med Image Anal*. 2018;44:177–195.
20. Andrearczyk V, Oreiller V, Jreige M, et al. Overview of the HECKTOR challenge at MICCAI 2020: automatic head and neck tumor segmentation in PET/CT. In: *3D Head and Neck Tumor Segmentation in PET/CT Challenge*. Lima, Peru: Springer; 2020:1–21.
21. Chen L, Shen C, Zhou Z, et al. Automatic PET cervical tumor segmentation by combining deep learning and anatomic prior. *Phys Med Biol*. 2019;64:085019.
22. Shiri I, Arabi H, Sanaat A, et al. Fully automated gross tumor volume delineation from PET in head and neck cancer using deep learning algorithms. *Clin Nucl Med*. 2021;46:872–883.
23. Rieke N, Hancox J, Li W, et al. The future of digital health with federated learning. *NPJ Digit Med*. 2020;3:119.
24. Kaissis GA, Makowski MR, Rückert D, et al. Secure, privacy-preserving and federated machine learning in medical imaging. *Nat Mach Intell*. 2020;2: 305–311.
25. Kirienko M, Sollini M, Ninatti G, et al. Distributed learning: a reliable privacy-preserving strategy to change multicenter collaborations using AI. *Eur J Nucl Med Mol Imaging*. 2021;48:3791–3804.
26. Wei K, Li J, Ding M, et al. Federated learning with differential privacy: algorithms and performance analysis. *IEEE Trans Inf Forensics Secur*. 2020;15: 3454–3469.
27. Mothukuri V, Parizi RM, Pouriyeh S, et al. A survey on security and privacy of federated learning. *Future Gener Comput Syst*. 2021;115:619–640.
28. Lu Y, Huang X, Dai Y, et al. Blockchain and federated learning for privacy-preserved data sharing in industrial IoT. *IEEE Trans Industr Inform*. 2019; 16:4177–4186.
29. Bonawitz K, Eichner H, Grieskamp W, et al. Towards federated learning at scale: system design. arXiv preprint arXiv:190201046. https://doi.org/10.48550/arXiv.1902.01046. 2019.
30. Li Q, Wen Z, Wu Z, et al. A survey on federated learning systems: vision, hype and reality for data privacy and protection. arXiv preprint arXiv: 190709693. https://doi.org/10.1109/TKDE.2021.3124599. 2019.
31. Li T, Sahu AK, Talwalkar A, et al. Federated learning: challenges, methods, and future directions. *IEEE Signal Proc Mag*. 2020;37:50–60.
32. Amiri MM, Gündüz D. Federated learning over wireless fading channels. *IEEE T Wirel Commun*. 2020;19:3546–3557.
33. McMahan B, Moore E, Ramage D, et al. Communication-efficient learning of deep networks from decentralized data. In: Aarti S, Jerry Z, eds. *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*. Fort Lauderdale, FL: PMLR; 2017:1273–1282.
34. Roth HR, Chang K, Singh P, et al. Federated learning for breast density classification: a real-world implementation. In: *Domain Adaptation and Representation Transfer, and Distributed and Collaborative Learning*. Lima, Peru: Springer; 2020:181–191.
35. Dayan I, Roth HR, Zhong A, et al. Federated learning for predicting clinical outcomes in patients with COVID-19. *Nat Med*. 2021;27:1735–1743.
36. Li W, Milletarì F, Xu D, et al. Privacy-preserving federated brain tumour segmentation. In: *International Workshop on Machine Learning in Medical Imaging*. Shenzhen, China: Springer; 2019:133–141.
37. Xia Y, Yang D, Li W, et al. Auto-FedAvg: learnable federated averaging for multi-institutional medical image segmentation. arXiv preprint arXiv: 210410195. https://doi.org/10.48550/arXiv.2104.10195. 2021.
38. Vallières M, Kay-Rivest E, Perrin LJ, et al. Radiomics strategies for risk assessment of tumour failure in head-and-neck cancer. *Sci Rep*. 2017;7:10117.
39. Clark K, Vendt B, Smith K, et al. The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. *J Digit Imaging*. 2013;26:1045–1057.
40. Fedorov A, Clunie D, Ulrich E, et al. DICOM for quantitative imaging biomarker development: a standards based approach to sharing clinical data and structured PET/CT analysis results in head and neck cancer research. *PeerJ*. 2016;4:e2057.
41. Grossberg AJ, Mohamed ASR, Elhalawani H, et al. Imaging and clinical data archive for head and neck squamous cell carcinoma patients treated with radiotherapy. *Sci Data*. 2018;5:180173.
42. MICCAI/M.D. Anderson Cancer Center Head and Neck Quantitative Imaging Working Group. Matched computed tomography segmentation and demographic data for oropharyngeal cancer radiomics challenges. *Sci Data*. 2017;4:170077.
43. Zuley ML, Jarosz R, Kirk S, et al. Radiology data from the Cancer Genome Atlas Head-Neck Squamous Cell Carcinoma [TCGA-HNSC] collection. *Cancer Imaging Arch*. 2016;10:K9.
44. Konečný J, McMahan HB, Yu FX, et al. Federated learning: strategies for improving communication efficiency. arXiv preprint arXiv:161005492. 2016.
45. Singh A, Vepakomma P, Gupta O, et al. Detailed comparison of communication efficiency of split learning and federated learning. arXiv preprint arXiv: 190909145. https://doi.org/10.48550/arXiv.1909.09145. 2019.
46. Luping W, Wei W, Bo L. CMFL: mitigating communication overhead for federated learning. In: *2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS),* Dallas, TX: IEEE; 2019:954–964.
47. Zhang M, Qu L, Singh P, et al. SplitAVG: A heterogeneity-aware federated deep learning method for medical imaging. arXiv preprint arXiv: 210702375. https://doi.org/10.48550/arXiv.2107.02375. 2021.
48. Stripelis D, Saleem H, Ghai T, et al. Secure neuroimaging analysis using federated learning with homomorphic encryption. In: *17th International Symposium on Medical Information Processing and Analysis,* Campinas, Brazil: SPIE; 2021:351–359.
49. Qu L, Zhou Y, Liang PP, et al. Rethinking architecture design for tackling data heterogeneity in federated learning. arXiv preprint arXiv:210606047. https://doi.org/10.48550/arXiv.2106.06047. 2021.

50. Liu Q, Yang H, Dou Q, et al. Federated semi-supervised medical image classification via inter-client relation matching. arXiv preprint arXiv:210608600. https://doi.org/10.48550/arXiv.2106.08600 . 2021.
51. Chakravarty A, Kar A, Sethuraman R, et al. Federated learning for site aware chest radiograph screening. In: *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI),* Nice, France: IEEE; 2021:1077–1081.
52. Linardos A, Kushibar K, Walsh S, et al. Federated learning for multi-center imaging diagnostics: a study in cardiovascular disease. arXiv preprint arXiv: 210703901. 2021.
53. Feki I, Ammar S, Kessentini Y, et al. Federated learning for COVID-19 screening from chest x-ray images. *Appl Soft Comput*. 2021;106:107330.
54. Alom MZ, Yakopcic C, Hasan M, et al. Recurrent residual U-Net for medical image segmentation. *J Med Imaging (Bellingham)*. 2019;6:014006.
55. Ashrafinia S. *Quantitative Nuclear Medicine Imaging Using Advanced Image Reconstruction and Radiomics*. Baltimore, MD: The Johns Hopkins University; 2019.
56. Zwanenburg A, Vallières M, Abdalah MA, et al. The Image Biomarker Standardization Initiative: standardized quantitative Radiomics for high-throughput image-based phenotyping. *Radiology*. 2020;295:328–338.
57. Foster B, Bagci U, Mansoor A, et al. A review on segmentation of positron emission tomography images. *Comput Biol Med*. 2014;50:76–96.
58. Hatt M, Lee JA, Schmidtlein CR, et al. Classification and evaluation strategies of auto-segmentation approaches for PET: report of AAPM task group no. 211. *Med Phys*. 2017;44:e1–e42.
59. Drever L, Roa W, McEwan A, et al. Iterative threshold segmentation for PET target volume delineation. *Med Phys*. 2007;34:1253–1265.
60. Hatt M, Cheze le Rest C, Turzo A, et al. A fuzzy locally adaptive Bayesian segmentation approach for volume determination in PET. *IEEE Trans Med Imaging*. 2009;28:881–893.
61. Brambilla M, Matheoud R, Secco C, et al. Threshold segmentation for PET target volume delineation in radiation treatment planning: the role of target-to-background ratio and target size. *Med Phys*. 2008;35:1207–1213.
62. Zaidi H, Abdoli M, Fuentes CL, et al. Comparative methods for PET image segmentation in pharyngolaryngeal squamous cell carcinoma. *Eur J Nucl Med Mol Imaging*. 2012;39:881–891.
63. Huang B, Chen Z, Wu PM, et al. Fully automated delineation of gross tumor volume for head and neck Cancer on PET-CT using deep Learning: a dual-center study. *Contrast Media Mol Imaging*. 2018;2018:8923028.
64. Andrearczyk V, Oreiller V, Vallières M, et al. Automatic segmentation of head and neck tumors and nodal metastases in PET-CT scans. In: *Proceedings of Machine Learning Research*. Montreal, QC, Canada: PMLR; 2020: 33–43.
65. Leung KH, Marashdeh W, Wray R, et al. A physics-guided modular deep-learning based automated framework for tumor segmentation in PET. *Phys Med Biol*. 2020;65:245032.
66. Gawali M, Arvind C, Suryavanshi S, et al. Comparison of privacy-preserving distributed deep learning methods in healthcare. In: *Annual Conference on Medical Image Understanding and Analysis*. Oxford, United Kingdom: Springer; 2021:457–471.
67. Lee H, Chai YJ, Joo H, et al. Federated Learning for thyroid ultrasound image analysis to protect personal information: validation study in a real health care environment. *JMIR Med Inform*. 2021;9:e25869.
68. Li X, Gu Y, Dvornek N, et al. Multi-site fMRI analysis using privacy-preserving federated learning and domain adaptation: ABIDE results. *Med Image Anal*. 2020;65:101765.
69. Melis L, Song C, De Cristofaro E, et al. Exploiting unintended feature leakage in collaborative learning. In: *2019 IEEE Symposium on Security and Privacy (SP)*. San Francisco, CA: IEEE; 2019:691–706.
70. Carlini N, Liu C, Erlingsson Ú, et al. The secret sharer: evaluating and testing unintended memorization in neural networks. In: *28th Security Symposium (Security 19)*. Santa Clara, CA; USENIX; 2019:267–284.
71. Duchi JC, Jordan MI, Wainwright MJ. Privacy aware learning. *Journal of the ACM (JACM)*. 2014;61:1–57.
72. Shokri R, Stronati M, Song C, et al. Membership inference attacks against machine learning models. In: *2017 IEEE Symposium on Security and Privacy (SP)*. San Jose, CA: IEEE; 2017:3–18.
73. Fredrikson M, Jha S, Ristenpart T. Model inversion attacks that exploit confidence information and basic countermeasures. In: *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*. 2015:1322–1333.
74. Malekzadeh M, Hasircioglu B, Mital N, et al. Dopamine: differentially private federated learning on medical data. *The Second AAAI Workshop on Privacy-Preserving Artificial Intelligence (PPAI-21)*. arXiv e-prints, pp. arXiv–2101, 2021.
75. Pfohl SR, Dai AM, Heller K. Federated and differentially private learning for electronic health records. arXiv preprint arXiv:191105861. https://doi.org/10.48550/arXiv.1911.05861. 2019.
76. Chen X, Liu C, Li B, et al. Targeted backdoor attacks on deep learning systems using data poisoning. arXiv preprint arXiv:171205526. https://doi.org/10.48550/arXiv.1712.05526. 2017.
77. Li B, Wang Y, Singh A, et al. Data poisoning attacks on factorization-based collaborative filtering. *Adv Neural Inform Proc Syst*. 2016;29:1885–1893.
78. Xie C, Huang K, Chen PY, et al. DBA: distributed backdoor attacks against federated learning. In: International Conference on Learning Representations. Addis Ababa, Ethiopia: ICLR; 2019.