Sahlsten, Jaakko; Wahid, Kareem A.; Glerean, Enrico; Jaskari, Joel; Naser, Mohamed A.; He, Renjie; Kann, Benjamin H.; Mäkitie, Antti; Fuller, Clifton D.; Kaski, Kimmo

Segmentation stability of human head and neck cancer medical images for radiotherapy applications under de-identification conditions: Benchmarking data sharing and artificial intelligence use-cases

Published in:
Frontiers in Oncology

DOI:
10.3389/fonc.2023.1120392

Published: 01/01/2023

Document Version
Publisher's PDF, also known as Version of record

Please cite the original version:
Sahlsten, J., Wahid, K. A., Glerean, E., Jaskari, J., Naser, M. A., He, R., Kann, B. H., Mäkitie, A., Fuller, C. D., & Kaski, K. (2023). Segmentation stability of human head and neck cancer medical images for radiotherapy applications under de-identification conditions: Benchmarking data sharing and artificial intelligence use-cases. *Frontiers in Oncology*, *13*, 1-10. Article 1120392. https://doi.org/10.3389/fonc.2023.1120392

frontiers | Frontiers in Oncology

TYPE Original Research
PUBLISHED 28 February 2023
DOI 10.3389/fonc.2023.1120392

OPEN ACCESS

EDITED BY
Jasper Nijkamp,
Aarhus University, Denmark

REVIEWED BY
Mark J. Gooding,
Inpictura Ltd, United Kingdom
Elisa D'Angelo,
University Hospital of Modena, Italy

*CORRESPONDENCE
Clifton D. Fuller
cdfuller@mdanderson.org
Kimmo Kaski
kimmo.kaski@aalto.fi

SPECIALTY SECTION
This article was submitted to Cancer Imaging and Image-directed Interventions, a section of the journal Frontiers in Oncology

RECEIVED 09 December 2022
ACCEPTED 13 February 2023
PUBLISHED 28 February 2023

CITATION
Sahlsten J, Wahid KA, Glerean E, Jaskari J, Naser MA, He R, Kann BH, Mäkitie A, Fuller CD and Kaski K (2023) Segmentation stability of human head and neck cancer medical images for radiotherapy applications under de-identification conditions: Benchmarking data sharing and artificial intelligence use-cases. *Front. Oncol.* 13:1120392. doi: 10.3389/fonc.2023.1120392

Segmentation stability of human head and neck cancer medical images for radiotherapy applications under de-identification conditions: Benchmarking data sharing and artificial intelligence use-cases

Jaakko Sahlsten[1], Kareem A. Wahid[2], Enrico Glerean[3], Joel Jaskari[1], Mohamed A. Naser[2], Renjie He[2], Benjamin H. Kann[4], Antti Mäkitie[5], Clifton D. Fuller[2*] and Kimmo Kaski[1*]

[1]Department of Computer Science, Aalto University School of Science, Espoo, Finland, [2]Department of Radiation Oncology, The University of Texas MD Anderson Cancer Center, Houston, TX, United States, [3]Department of Neuroscience and Biomedical Engineering, Aalto University, Espoo, Finland, [4]Artificial Intelligence in Medicine Program, Brigham and Women's Hospital, Dana-Farber Cancer Institute, Harvard Medical School, Boston, MA, United States, [5]Department of Otorhinolaryngology, Head and Neck Surgery, University of Helsinki and Helsinki University Hospital, Helsinki, Finland

Background: Demand for head and neck cancer (HNC) radiotherapy data in algorithmic development has prompted increased image dataset sharing. Medical images must comply with data protection requirements so that re-use is enabled without disclosing patient identifiers. Defacing, i.e., the removal of facial features from images, is often considered a reasonable compromise between data protection and re-usability for neuroimaging data. While defacing tools have been developed by the neuroimaging community, their acceptability for radiotherapy applications have not been explored. Therefore, this study systematically investigated the impact of available defacing algorithms on HNC organs at risk (OARs).

Methods: A publicly available dataset of magnetic resonance imaging scans for 55 HNC patients with eight segmented OARs (bilateral submandibular glands, parotid glands, level II neck lymph nodes, level III neck lymph nodes) was utilized. Eight publicly available defacing algorithms were investigated: afni_refacer, DeepDefacer, defacer, fsl_deface, mask_face, mri_deface, pydeface, and quickshear. Using a subset of scans where defacing succeeded (N=29), a 5-fold cross-validation 3D U-net based OAR auto-segmentation model was utilized to perform two main experiments: 1.) comparing original and defaced data for training when evaluated on original data; 2.) using original data for training and comparing the model evaluation on original and defaced data. Models were primarily assessed using the Dice similarity coefficient (DSC).

Results: Most defacing methods were unable to produce any usable images for evaluation, while mask_face, fsl_deface, and pydeface were unable to remove the face for 29%, 18%, and 24% of subjects, respectively. When using the original data for evaluation, the composite OAR DSC was statistically higher ($p \leq 0.05$) for the model trained with the original data with a DSC of 0.760 compared to the mask_face, fsl_deface, and pydeface models with DSCs of 0.742, 0.736, and 0.449, respectively. Moreover, the model trained with original data had decreased performance ($p \leq 0.05$) when evaluated on the defaced data with DSCs of 0.673, 0.693, and 0.406 for mask_face, fsl_deface, and pydeface, respectively.

Conclusion: Defacing algorithms may have a significant impact on HNC OAR auto-segmentation model training and testing. This work highlights the need for further development of HNC-specific image anonymization methods.

KEYWORDS

anonymization, radiotherapy, head and neck cancer, MRI, medical imaging, artificial intelligence (AI), autosegmentation, defacing

Introduction

The landscape of data democratization is rapidly changing. The rise of open science practices, inspired by coalitions such as the Center for Open Science (1), and the FAIR (Findable, Accessible, Interoperable, and Reusable) guiding principles (2), has spurred interest in public data sharing. Subsequently, the medical imaging community has increasingly adopted these practices through initiatives such as The Cancer Imaging Archive (3). Given the appropriate removal of protected health information through anonymization techniques, public repositories have democratized the access to medical imaging data such that the world at large can now help develop algorithmic approaches to improve clinical decision-making. Among the medical professions seeking to leverage these large datasets, radiation oncology has the potential to vastly benefit from these open science practices (4). Imaging is crucial to radiotherapy workflows, particularly for organ at risk (OAR) and tumor segmentation (5, 6). Moreover, in recent years public data competitions, such as the Head and Neck Tumor Segmentation and Outcome Prediction in positron emission tomography/computed tomography (PET/CT) Images (HECKTOR) challenge (7–9), have been targeted to improve the radiotherapy workflow. However, there is a particular facet of medical image dissemination for radiotherapy applications that has spurred controversy, namely the anonymization of head and neck cancer (HNC) related images.

While the public dissemination of HNC image data is invaluable to improve the radiotherapy workflow, concerns have been raised regarding readily identifiable facial features on medical imaging. Importantly, the U.S. Health Insurance Portability and Accountability Act references “full-face photographs and any comparable images” as a part of protected health information (10). This policy introduces some uncertainty in the dissemination of high-resolution images, where the intricacies of facial features can be reconstructed to generate similar or “comparable” visualizations with relative ease. Several studies have shown the potential danger in releasing unaltered medical images containing facial features, as they can often be easily recognized by humans and/or machines (11–15). For example, using facial recognition software paired with image-derived facial reconstructions, one study found up to 83% of research participants could be identified from their magnetic resonance imaging (MRI) scans (13). Similar alarming results have been demonstrated for CT images (14). While brain images are often processed such that obvious facial features are removed (i.e., skull stripping), these crude techniques remove large anatomic regions necessary for building predictive models with HNC imaging data. “Defacing” tools, where voxels that correspond to the areas of the patient’s facial features are either removed or altered, offer one solution. However, they may still engender the potential loss of voxel-level information needed for predictive modeling or treatment planning, thereby prohibiting their use in data resharing strategies for radiotherapy applications. While several studies have investigated the effects of defacing for neuroimaging (16–21), there have not yet been any systematic studies on the effects of defacing tools for radiotherapy applications.

Inspired by the increasing demand for public HNC imaging datasets and the importance of protecting the privacy of patients, a systematic analysis of a number of existing methods for facial anonymization on HNC MRI images was performed. Through qualitative and quantitative analysis using open-source datasets and tools, the efficacies of defacing approaches on whole images and structures relevant to radiation treatment planning were determined. Moreover, the effects of these approaches on auto-segmentation, a specific domain application that is increasingly relevant for HNC public datasets, were also examined. This study is an important first step towards the development of robust approaches for the safe and trusted democratization of HNC imaging data.

Methods

Dataset

For this analysis, a publicly available dataset hosted on the TCIA, the American Association of Physicists in Medicine RT-MAC Grand Challenge 2019 (AAPM) dataset (22), was utilized. The AAPM dataset consists of T2-weighted MRI scans of 55 HNC patients that are labeled for OAR segmentations of bilateral: i) submandibular glands, ii) level II neck lymph nodes, iii) level III neck lymph nodes, and iv) parotid glands. Structures were annotated as being on the right or left side of the patient anatomy. The spatial resolution of the scans is 0.5 mm × 0.5 mm with 2.0 mm spacing. Additional technical details on the AAPM images and segmentations can be found in the corresponding data descriptor (22). Defacing experiments were also attempted using the HECKTOR 2021 training dataset (8) containing 224 HNC patients with CT scans. Additional technical details on the HECKTOR dataset can be found in the corresponding overview papers (8, 9).

Defacing methods

For defacing the images, the same methods as taken into consideration by Schwartz et al. (16), as well as novel tools that benefit from recent advances in deep learning were used. The most popular tools use a co-registration to a template in order to identify face and ears and then identify those structures in the original image, which should be removed or blurred. The following 6 co-registration based methods: *afni_refacer*, *fsl_deface* (23), *mask_face* (24), *mri_deface* (18), *pydeface* (25), and *quickshear* were implemented. Two more recent methods using deep learning technology were also included: *defacer* (26) and *DeepDefacer* (27). These methods utilize pre-trained deep learning models using data from public neuroimaging datasets to identify facial features to be removed. An automated pipeline for applying all these defacing methods is available at https://github.com/eglerean/faceai_testingdefacing. Each defacing method was tested with all subjects such that, for each subject, a defaced volume was produced as well as a volumetric mask of which voxels were affected by defacing. All methods were run with the default parameters and standard reference images.

Defacing performance

After applying the defacing methods, the success or failure of a defacing method was determined by visually inspecting all the defaced volumes (i.e., performing scanwise quality control). Specifically, a binary categorization of each scan was implemented: “1” if the eyes, nose, and mouth were removed (i.e., defacing succeeded), “0” if the eyes, nose, or mouth were not removed (i.e., defacing failed). Subsequently, the amount of voxels present in the structures after application of the defacing algorithm were quantitatively measured.

Deep learning model for OAR segmentation reliability

To evaluate the OAR segmentation performance under different defacing schemes from volumetric MRI data, a convolutional neural network architecture, 3D U-net, which has found wide success in HNC-related segmentation tasks (28–33), was utilized. Both contractive and expansive pathways include four blocks, where each block consists of two convolutional layers with a kernel size of 3, and each convolution is followed by an instance normalization layer and a LeakyReLU activation with 0.1 negative slope. The max-pooling and transpose convolutional layers have a kernel size and stride of 2. The last convolutional layer has a kernel size and stride of 1 with 9 output channels and a softmax activation. The model architecture is shown in Figure 1. Experiments were developed in Python v. 3.6.10 (34) using Pytorch 1.8.1 (35) with a U-net model from Project MONAI 0.7.0 (36) and data preprocessing and augmentation with TorchIO 0.18.61 (37).

A subset of patients for which defacing was deemed successful were used for building the segmentation models. The subset was randomly split with 5-fold cross validation: for each cross-validation iteration one fold was used for model testing, one fold was used for model validation, and the remaining three folds were used for model training. The reported segmentation performance

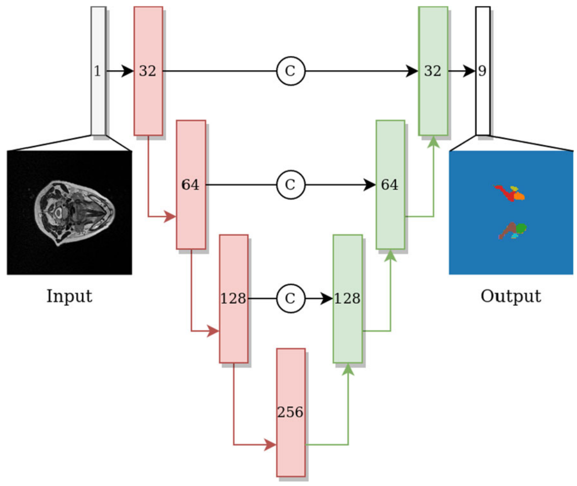

FIGURE 1
U-net network architecture with blocks on the contractive path colored in red and blocks on the expanding path colored in green. Each block includes two convolutions, each followed by instance normalization and Leaky ReLU activation, subsequently followed by a max-pool layer (red arrow) or transpose convolution layer (green arrow) on contractive and expanding paths, respectively. The number shown in each block indicates the number of channels of the feature map. Arrows with the letter C indicate concatenation.

was based on the test fold that was not used for model development. The same random splits were used for training and evaluating the models trained on original or defaced data.

Data preprocessing after the defacing included linear resampling to 2 mm isotropic resolution with the intensity scaled into a range of [-1,1]. The training data was augmented with random transforms that were applied with a probability (p), independently of each other. The used transforms were random elastic deformations (p=10%) for all axes, random flips for inferior-superior and anterior-posterior axes (p=50%), random rotation (-10° to 10°) of all axes (p=50%), random bias field (p=50%), and random gamma (p=50%). The model was trained using the cross-entropy loss for the 8 OAR classes and background with parameter updates computed using the Adam optimizer with (0.001 learning rate, 0.9 β_1, 0.999 β_2, and AMSGrad). The model training was stopped early after 60 epochs for non-improvement of the validation loss.

Segmentation evaluation

Two experiments to evaluate the impact of defacing on the resulting segmentations were performed. In order to determine the impact of defacing on algorithmic development, models were trained on original or defaced data using the original target data for evaluation. Subsequently, in order to determine the impact of defacing on algorithms not originally developed for defaced data, a model was trained using the original data and its performance was evaluated by using the original data or the defaced data.

For both experiments, the performance of the models were quantified primarily with the Dice similarity coefficient (DSC) and the mean surface distance (MSD), defined as follows:

$$DSC = \frac{2\ \ TP}{2\ \ TP + FP + FN},$$

$$MSD = \frac{1}{2}\left(\sum_{t\in T}\frac{d(t,\ P)}{|T|} + \sum_{p\in P}\frac{d(p,\ T)}{|P|}\right),$$

where TP denotes true positives, FP false positives, FN false negatives, P the set of segmentation surface voxels of the model output, and T the set of segmentation surface voxels of the annotation. The distance from the surface metric is defined as:

$d(a, B)=min_{b\in B}\{\|a-b\|_2\}$. These metrics were selected because of their ubiquity in literature and ability to capture both volumetric overlap and boundary distances (38, 39). The model output was resampled into the original resolution with the nearest-neighbor sampling and evaluated against the original resolution segmentations. MSD was measured in millimeters. When comparing the performance measures between the segmentation models, Wilcoxon signed rank tests (40) were implemented, with p-values less than or equal to 0.05 considered as significant. To correct for multiple hypotheses, a Benjamini-Hochberg false discovery rate procedure (41) was implemented by taking into account all the OARs and models compared. Statistical comparisons were performed using the statannotations 0.4.4 Python package (https://github.com/trevismd/statannotations). Notably, any ROI metrics that yielded empty outputs were omitted from the comparisons. Additional surface metric values (mean Hausdorff distance at 95% and Hausdorff distance at 95%) were also calculated as part of the supplementary analysis (details in Appendix A).

Results

Defacing performance

Five of the methods tested (afni_refacer, quickshear, mri_deface, DeepDefacer, and defacer) failed for all subjects in the AAPM dataset. Therefore, for all subsequent analyses only the mask_face, fsl_deface, and pydeface methods were considered. There was scanwise quality control to remove the defaced scans with poor quality from the analyses, which resulted in 16 (29%), 10 (18%), and 13 (24%) scans removed from mask_face, fsl_deface, and pydeface, respectively, with all these methods working on 29 patient scans. A barplot comparison of the ratio of remaining OAR voxels after defacing and quality control is depicted in Figure 2. In addition, the defacing methods removed some OARs completely, which were also omitted from the segmentation evaluation. After filtering unusable data, the total number of OARs available for use in segmentation experiments was 232 for the original data and mask_face, 231 for fsl_deface, and 169 for pydeface. A full comparison of omitted OARs is shown in Table 1.

All of the tested defacing methods were unable to provide sufficient data for segmentation analysis in the HECKTOR CT dataset. Specifically, fsl_deface and pydeface methods successfully defaced 18 (8%) and 102 (46%) scans, respectively. All other methods (afni_refacer, quickshear, mri_deface, DeepDefacer, defacer, and mask_face) failed to correctly deface any of the scans. Although pydeface had the highest success rate on defacing, it only preserved the brain. Thus, no further analysis was performed for this dataset.

Segmentation performance

The 29 patient scans for which the defacing was deemed successful were used to construct and evaluate segmentation models for the mask_face, fsl_deface, and pydeface methods. The model DSC performances pooled across all structures based on training input and valid evaluation target combinations are shown in Table 2. The models trained using the original, mask_face, and fsl_deface input data had the highest composite mean DSC when evaluated on the original target data with values of 0.760, 0.742, and 0.736, respectively, while the model trained on pydeface input data had the highest composite mean DSC of 0.653 when evaluated on pydeface target data. In contrast, the models trained using original mask_face, and fsl_deface input data had the lowest composite mean DSC when evaluated on pydeface target data with values of 0.406, 0.413, 0.465, respectively, while the model trained using pydeface input data had the lowest composite mean DSC of 0.395

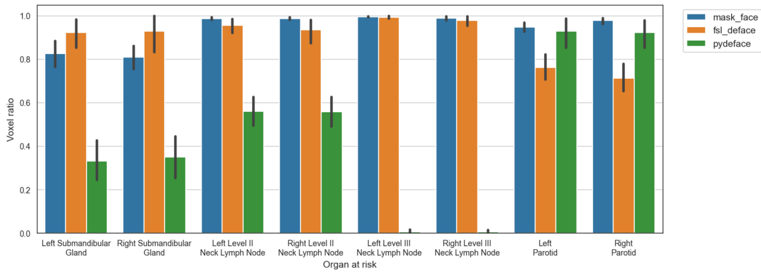

FIGURE 2
Ratio of preserved voxels in comparison to the original segmentation mask after defacing (mask_face, fsl_deface, and pydeface) for each of the organs at risk, where defacing was successful for N=39, N=42, and N=45, respectively. The mean and standard deviation are represented as the center and extremes of the error bars, respectively.

when evaluated on fsl_deface target data. All comparisons within the same evaluation data are statistically different from each other ($p \leq 0.05$) with the exception of mask_face and fsl_deface trained models evaluated on original data, and original as well as mask_face trained models evaluated on pydeface data.

Defacing impact on model training

The analysis was based on eight OAR structure segmentations from 29 patients totaling 232 evaluations. The MSD of left and right level III neck lymph nodes for pydeface trained models were omitted from the analysis as all the model outputs were empty. Full comparisons of the model performance for each OAR are depicted in Figure 3. Additional surface distance metrics are shown in Appendix A (Figure A1). Overall, the model trained with the original data performed better than the models trained with the defaced data for the majority of structures and evaluation metrics. Both metrics were significantly better for the model trained with the original data compared to the model trained with mask_face data for the left submandibular gland and right level II neck lymph node, while only the DSC was significantly better for the right submandibular gland and right level III neck lymph node. Similarly, both metrics were significantly better for the model trained with the original data compared to the model trained with fsl_deface data for the right level II neck lymph node, left parotid, and right parotid, while only the DSC was significantly better for the right level III neck lymph node. Moreover, both metrics were significantly better for the model trained with the original data compared to the model trained with pydeface data for all the structures.

Defacing impact on model testing

In these results, only valid target data with successful defacing on all three methods using non-empty segmentation structures were included. This was obtained using results from 26 left submandibular glands, 27 right submandibular glands, 1 left neck level III lymph nodes, 2 right neck level III lymph nodes, and 28 of each of the remaining structures. Due to the low number of cases for the right and left level III lymph nodes, they were omitted from the comparison. In addition, for the MSD metric, empty model output segmentations were discarded resulting in evaluation of 1 left submandibular gland for fsl_deface and mask_face and 14 for pydeface, 1 and 6 right submandibular glands on fsl_deface and

TABLE 1 Quantitative details on the number of organs at risk available after the defacing was applied for all 55 patient scans.

	Completely removed after successful defacing			Unavailable for segmentation analysis*		
Organ at risk/Defacing method	mask_face	fsl_deface	pydeface	mask_face	fsl_deface	pydeface
Left Submandibular Gland	0 (0%)	2 (4%)	6 (11%)	16 (29%)	14 (25%)	13 (24%)
Right Submandibular Gland	0 (0%)	1 (2%)	7 (13%)	16 (29%)	14 (25%)	14 (25%)
Left Neck Lymph Node Level II	0 (0%)	0 (0%)	4 (7%)	16 (29%)	13 (24%)	11 (20%)
Right Neck Lymph Node Level II	0 (0%)	0 (0%)	4 (7%)	16 (29%)	13 (24%)	11 (20%)
Left Neck Lymph Node Level III	0 (0%)	0 (0%)	45 (82%)	16 (29%)	13 (24%)	51 (93%)
Right Neck Lymph Node Level III	0 (0%)	1 (2%)	44 (80%)	16 (29%)	13 (24%)	50 (91%)
Left Parotid	0 (0%)	0 (0%)	6 (11%)	16 (29%)	13 (24%)	11 (20%)
Right Parotid	0 (0%)	0 (0%)	6 (11%)	16 (29%)	13 (24%)	11 (20%)
Total omitted	0 (0%)	4 (1%)	122 (28%)	128 (29%)	106 (24%)	172 (39%)

Only the mask_face, fsl_deface, and pydeface methods yielded usable data. The first group of columns correspond to the organs at risk that were completely removed from the cases with successful defacing. The second group of columns correspond to all items in the first group of columns plus incorporating any of the cases where defacing failed. Defacing success or failure was counted from scanwise quality control. *Organs at risk in these columns were omitted for all the subsequent segmentation-related experiments.

TABLE 2 Composite DSC performance - mean (standard deviation) - of all structures for all combinations of training data (rows) and evaluation data (columns).

	Evaluated on original (N =2 32)	Evaluated on mask_face (N = 232)	Evaluated on fsl_deface (N = 231)	Evaluated on pydeface (N = 169)
Trained on original	0.760 (0.112)	0.673 (0.181)	0.693 (0.140)	0.406 (0.304)
Trained on mask_face	0.742 (0.115)	0.733 (0.120)	0.668 (0.143)	0.413 (0.312)
Trained on fsl_deface	0.736 (0.108)	0.643 (0.185)	0.733 (0.122)	0.465 (0.293)
Trained on pydeface	0.449 (0.333)	0.417 (0.325)	0.395 (0.301)	0.653 (0.258)

The number of total segmentation maps evaluated is shown in brackets on the header. All comparisons within the same evaluation data are statistically different from each other ($p \leq 0.05$) with the exception of mask_face and fsl_deface trained models evaluated on original data, and original and mask_face trained models evaluated on pydeface data. Statistical significance was measured with Wilcoxon signed-rank tests corrected with Benjamini-Hochberg procedure comparisons within evaluation data.

pydeface, respectively, 1 left level II lymph node for pydeface, and 2 left parotids for pydeface. The model evaluated on the original data performed significantly better than the models evaluated on the defaced data for all of the structures and both evaluation metrics except in the case of left submandibular gland DSC for fsl_deface which exhibited a non-significant difference. The full comparison of the model performance for each of the OARs is shown in Figure 4. Additional surface distance metrics are shown in Appendix A (Figure A2).

Discussion

This study has systematically investigated the impact of a variety of defacing algorithms on structures of interest used for radiotherapy treatment planning. This study demonstrated that the overall usability of segmentations is heavily dependent on the choice of the defacing algorithm. Moreover, the results indicate that several OARs have the potential to be negatively impacted by the defacing algorithms, which is shown by the decreased performance of auto-segmentation algorithms trained and evaluated on defaced data in comparison to algorithms trained and evaluated on non-defaced data.

Defacing for HNC applications should be deemed optimal if the method simultaneously removes all recognizable facial features from the image and no voxels from structures of interest are affected. In this study, eight commonly available defacing algorithms developed by the neuroimaging community were applied: afni_refacer, mri_deface, defacer, DeepDefacer, mask_face, fsl_deface, pydeface, and quickshear. Unfortunately, for the investigated CT data, no defacing method was able to yield successful removal of facial features while preserving the OARs. This is not necessarily surprising given that the methods investigated were developed primarily with MRI in mind; these results echo previous similar work using CT data (42). Importantly, even when applied to MRI data of HNC patients, many of these

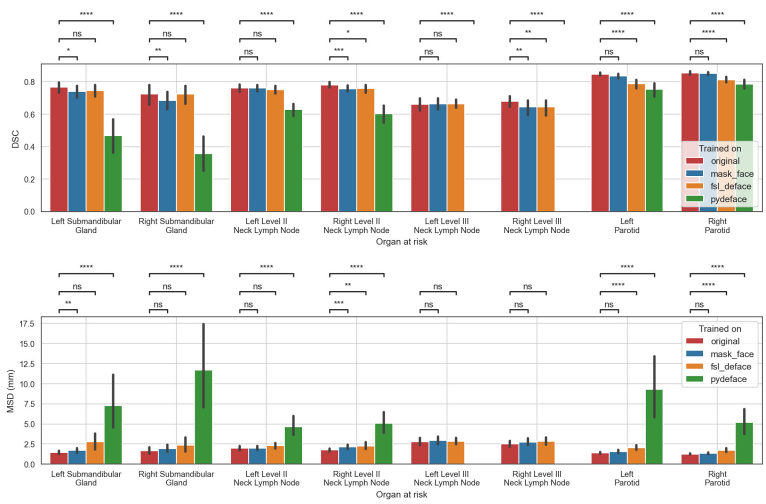

FIGURE 3
Performance of the models trained on original or defaced data and evaluated on the original data. The mean and standard deviation for each metric are represented as the center and extremes of the error bars, respectively. Statistical significance was determined using Wilcoxon signed-rank tests corrected with Benjamini-Hochberg procedure for all OARs and models. Comparison symbols: ns ($p > 0.05$), * ($p \leq 0.05$), ** ($p \leq 0.01$), *** ($p \leq$ 1e-4), **** ($p \leq$ 1e-5).

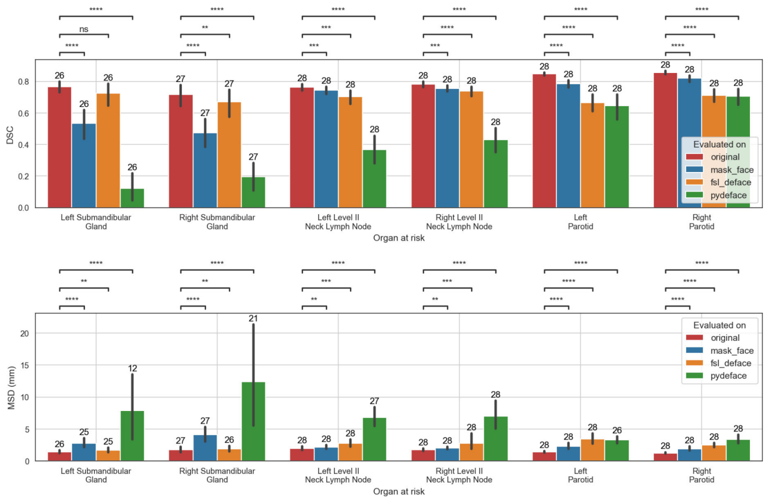

FIGURE 4
The performance of models trained on the original data when evaluated on the original, mask_face, fsl_deface, or pydeface data for the six organs at risk included in the analysis. Only cases that were available for all the methods were included: 28 segmentations were used for all structures except in the case of the left and right submandibular glands where 26 and 27 segmentations were used, respectively. In addition, for the MSD metric, empty model output segmentations were discarded, which resulted in a smaller number of evaluated structures. The number of evaluated structures is shown on top of the barplot. The mean and standard deviation for each metric are represented as the center and extremes of the error bars, respectively. Statistical significance was measured with Wilcoxon signed-rank tests corrected with Benjamini-Hochberg procedure for all OARs and models. Comparison symbols: ns ($p > 0.05$), * ($p \leq 0.05$), ** ($p \leq 0.01$), *** ($p \leq$ 1e-4), **** ($p \leq$ 1e-5).

defacing methods outright failed for most if not all patients. Therefore, despite extant studies demonstrating the acceptability of these methods to remove facial features from neuroimaging scans (16–21), these tools may not necessarily be robust to HNC-related imaging. Moreover, for those defacing algorithms that were able to successfully remove facial information in the MRI data, i.e. mask_face, fsl_deface, and pydeface, it was shown that regardless of the choice of the method, there was a loss of voxel-level information for all the OAR structures investigated. Importantly, pydeface leads to a greater number of lost voxels than mask_face and fsl_deface for all the OAR structures, with the exception of the parotid glands. While mask_face and fsl_deface lead to relatively minimal reduction of available voxels in many cases, the loss of topographic information in a radiotherapy workflow cannot be underscored enough. It is well known that even minor variations in the delineation of tumors and OARs can drastically alter the resulting radiotherapy dose delivered to a patient, which can impact important clinical outcomes such as toxicity and overall survival (43–46). Therefore, the loss of voxel-level information of OARs caused by the defacing algorithms, while potentially visibly imperceptible, can still affect downstream clinical workflows.

Relatively few studies have been conducted that determined the downstream analysis effects of defacing algorithms. For example, recent studies by Schwartz et al. (16) and Mikulan et al. (21) demonstrated that several defacing methods showed differences in specific neuroimaging applications, namely brain volume measurements and electroencephalography-related calculations. In this study, as a proxy for a clinically relevant task, an OAR auto-segmentation workflow was developed to investigate the impact of defacing-induced voxel-level information loss on downstream radiotherapy applications. As evident through both pooled analysis and investigation of individual OARs for auto-segmentation model training and evaluation, performance is often modestly decreased for fsl_deface and mask_face but greatly decreased for pydeface; these results were consistent with the overall voxel-level information loss. While pydeface has been shown to have favorable results for use with neuroimaging data (19, 21), its negative impact on HNC imaging is apparent. Therefore, in cases where defacing is unavoidable, mask_face or fsl_deface should likely be preferred for HNC image anonymization. Regardless, this study demonstrates existing approaches to anonymize facial data may not be sufficient for implementation on HNC-related datasets, particularly for deep learning model training and testing.

This study has several limitations. Firstly, to examine defacing methods as they are currently distributed ("out-of-the-box"), modifications to the templates or models utilized in any methods were not performed. Further preprocessing either of the CT and MRI data as well as subject specific settings could have helped some of the methods to better identify the face. In addition, more suitable templates for the HNC images (for both CT and MRI) would likely improve the defacing performance; for the registration-based methods, algorithms likely expected scans to cover the whole brain, while the field-of-view of the images for HNC mostly covered the neck and mouth, leaving the top of the brain excluded. Notably, additional deep learning model training schemes (i.e., transfer learning) may potentially allow for eventual implementation of existing deep learning methods on domain-specific datasets (i.e., HNC radiotherapy), but this negates the immediate interoperability of these tools. Furthermore, no additional image processing other than what was integrated into the defacing methods was implemented; it may be possible alternative processing could change these results. Secondly, while a robust analysis utilizing multiple relevant metrics established in existing literature (38) was performed to evaluate OAR auto-segmentation, there is not always a perfect correlation between spatial similarity metrics and radiotherapy plan acceptability (39). This study has not tested the downstream effects of defacing on radiotherapy plan generation, which may lead to different results from what was observed for the OAR segmentation. Thirdly, this study was limited to public data with no modifications. Only structures that were already available in existing datasets were analyzed. Moreover, as an initial exploration of defacing methods for radiotherapy applications, only a single imaging modality on a relatively limited sample size, namely T2-weighted MRI, was investigated for auto-segmentation experiments, despite the HNC radiotherapy workflow commonly incorporating additional modalities (47). Thus, experiments on additional imaging modalities and larger diverse HNC patient populations should be the subject of future investigations. Fourthly, the current analysis does not thoroughly explore possible performance confounding related to phenotypical and individual variables such as sex, ethnicity, and age of the measured individuals. Finally, this study has focused on defacing methods as an avenue for public data sharing for training and evaluating machine learning models, but privacy-preserving modeling approaches, e.g., through federated learning (48), may also act as a potential alternative solution.

Conclusion

In summary, by using publicly available data, the effects of eight established defacing algorithms, afni_refacer, mask_face, mri_deface, defacer, DeepDefacer, quickshear, fsl_deface, and pydeface, have been systematically investigated for radiotherapy applications. Specifically, the impact of defacing directly on ground-truth HNC OARs was determined and a deep learning based OAR auto-segmentation workflow to investigate the use of defaced data for algorithmic training and evaluation was developed. All methods failed to properly remove facial features on the CT dataset investigated. Moreover, it was observed that only fsl_deface, mask_face, and pydeface yielded usable images from the MRI dataset, but still decreased the total number of voxels in OARs and negatively impacted the performance of OAR auto-segmentation, with pydeface having more severe negative effects than mask_face or fsl_deface. This study is an important step towards ensuring widespread privacy-preserving dissemination of HNC imaging data without endangering data usability. Given that current defacing methods remove critical data, future larger studies should investigate alternative approaches for anonymizing facial data that preserve radiotherapy-related structures. Moreover, studies on the impact of these methods on radiotherapy plan generation, the inclusion of a greater number of OARs and target structures, and the incorporation of additional imaging modalities are also warranted.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: https://wiki.cancerimagingarchive.net/display/Public/AAPM+RT-MAC+Grand+Challenge+2019.

Ethics statement

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

Author contributions

Study concepts: all authors; Study design: JS, EG, and JJ; Data acquisition: KW, MN, and RH; Quality control of data and algorithms: JS and EG; Data analysis and interpretation: JS, KW, EG, JJ, BK, AM, and KK; Manuscript editing: JS, KW, EG, JJ, BK, AM, KK, and CF. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by the National Institutes of Health (NIH)/National Cancer Institute (NCI) through a Cancer Center Support Grant (CCSG; P30CA016672-44). MN is supported by an NIH grant (R01DE028290-01). KW is supported by a training fellowship from The University of Texas Health Science Center at Houston Center for Clinical and Translational Sciences TL1 Program (TL1TR003169), the American Legion Auxiliary Fellowship in Cancer Research, and an NIH/National Institute for Dental and Craniofacial Research (NIDCR) F31 fellowship (1 F31DE031502-01). CF received funding from the NIH/NIDCR (1R01DE025248-01/R56DE025248); an NIH/NIDCR Academic-Industrial Partnership Award (R01DE028290); the National Science Foundation (NSF), Division of Mathematical Sciences, Joint NIH/NSF Initiative on Quantitative Approaches to Biomedical Big Data (QuBBD) Grant (NSF 1557679); the NIH Big Data to Knowledge (BD2K) Program of the NCI Early Stage Development of Technologies in Biomedical Computing, Informatics, and Big Data Science Award (1R01CA214825); the NCI Early Phase Clinical Trials in Imaging and Image-Guided Interventions Program (1R01CA218148); an NIH/NCI Pilot Research Program Award from the UT MD Anderson CCSG Radiation Oncology and Cancer Imaging Program (P30CA016672); an NIH/NCI Head and Neck Specialized Programs of Research Excellence (SPORE) Developmental Research Program Award (P50CA097007); and the National Institute of Biomedical Imaging and Bioengineering (NIBIB) Research Education Program (R25EB025787).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fonc.2023.1120392/full#supplementary-material

References

1. Foster ED, Deardorff A. Open science framework (OSF). *J Med Libr Assoc JMLA* (2017) 105:203. doi: 10.5195/jmla.2017.88

2. Wilkinson MD, Dumontier M, IjJ A, Appleton G, Axton M, Baak A, et al. The FAIR guiding principles for scientific data management and stewardship. *Sci Data* (2016) 3:1–9. doi: 10.1038/sdata.2016.18

3. Clark K, Vendt B, Smith K, Freymann J, Kirby J, Koppel P, et al. The cancer imaging archive (TCIA): Maintaining and operating a public information repository. *J Digit Imaging* (2013) 26:1045–57. doi: 10.1007/s10278-013-9622-7

4. Wahid KA, Glerean E, Sahlsten J, Jaskari J, Kaski K, Naser MA, et al. Artificial intelligence for radiation oncology applications using public datasets. In: *Seminars in radiation oncology*. Elsevier (2022). p. 400–14.

5. Press RH, Shu H-KG, Shim H, Mountz JM, Kurland BF, Wahl RL, et al. The use of quantitative imaging in radiation oncology: A quantitative imaging network (QIN) perspective. *Int J Radiat Oncol Biol Phys* (2018) 102:1219–35. doi: 10.1016/j.ijrobp.2018.06.023

6. Beaton L, Bandula S, Gaze MN, Sharma RA. How rapid advances in imaging are defining the future of precision radiation oncology. *Br J Cancer* (2019) 120:779–90. doi: 10.1038/s41416-019-0412-y

7. Andrearczyk V, Oreiller V, Jreige M, Vallières M, Castelli J, Elhalawani H, 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*. Springer (2020). p. 1–21.

8. Andrearczyk V, Oreiller V, Boughdad S, Rest CCL, Elhalawani H, Jreige M, et al. Overview of the HECKTOR challenge at MICCAI 2021: Automatic head and neck tumor segmentation and outcome prediction in PET/CT images. In: *3D head and neck tumor segmentation in PET/CT challenge.* (Strasbourg, France: Springer) (2021). p. 1–37.

9. Oreiller V, Andrearczyk V, Jreige M, Boughdad S, Elhalawani H, Castelli J, et al. Head and neck tumor segmentation in PET/CT: The HECKTOR challenge. *Med Image Anal* (2022) 77:102336. doi: 10.1016/j.media.2021.102336

10. Meystre SM, Friedlin FJ, South BR, Shen S, Samore MH. Automatic de-identification of textual documents in the electronic health record: A review of recent research. *BMC Med Res Methodol* (2010) 10:1–16. doi: 10.1186/1471-2288-10-70

11. Prior FW, Brunsden B, Hildebolt C, Nolan TS, Pringle M, Vaishnavi SN, et al. Facial recognition from volume-rendered magnetic resonance imaging data. *IEEE Trans Inf Technol BioMed* (2008) 13:5–9. doi: 10.1109/TITB.2008.2003335

12. Mazura JC, Juluru K, Chen JJ, Morgan TA, John M, Siegel EL. Facial recognition software success rates for the identification of 3D surface reconstructed facial images: Implications for patient privacy and security. *J Digit Imaging* (2012) 25:347–51. doi: 10.1007/s10278-011-9429-3

13. Schwarz CG, Kremers WK, Therneau TM, Sharp RR, Gunter JL, Vemuri P, et al. Identification of anonymous MRI research participants with face-recognition software. *N Engl J Med* (2019) 381:1684–6. doi: 10.1056/NEJMc1908881

14. Parks CL, Monson KL. Automated facial recognition of computed tomography-derived facial images: patient privacy implications. *J Digit Imaging* (2017) 30:204–14. doi: 10.1007/s10278-016-9932-7

15. Delbarre DJ, Santos L, Ganjgahi H, Horner N, McCoy A, Westerberg H, et al. Application of a convolutional neural network to the quality control of MRI defacing. *Comput Biol Med* (2022) 151:106211. doi: 10.1016/j.compbiomed.2022.106211

16. Schwarz CG, Kremers WK, Wiste HJ, Gunter JL, Vemuri P, Spychalla AJ, et al. Changing the face of neuroimaging research: Comparing a new MRI de-facing technique with popular alternatives. *NeuroImage* (2021) 231:117845. doi: 10.1016/j.neuroimage.2021.117845

17. Schimke N, Kuehler M, Hale J. Preserving privacy in structural neuroimages. In: *IFIP annual conference on data and applications security and privacy*. (Richmond, VA, USA: Springer) (2011). p. 301–8.

18. Bischoff-Grethe A, Ozyurt IB, Busa E, Quinn BT, Fennema-Notestine C, Clark CP, et al. A technique for the deidentification of structural brain MR images. *Hum Brain Mapp* (2007) 28:892–903. doi: 10.1002/hbm.20312

19. Theyers AE, Zamyadi M, O'Reilly M, Bartha R, Symons S, MacQueen GM, et al. Multisite comparison of MRI defacing software across multiple cohorts. *Front Psychiatry* (2021) 12:189. doi: 10.3389/fpsyt.2021.617997

20. De Sitter A, Visser M, Brouwer I, Cover K, van Schijndel R, Eijgelaar R, et al. Facing privacy in neuroimaging: removing facial features degrades performance of image analysis methods. *Eur Radiol* (2020) 30:1062–74. doi: 10.1007/s00330-019-06459-3

21. Mikulan E, Russo S, Zauli FM, d'Orio P, Parmigiani S, Favaro J, et al. A comparative study between state-of-the-art MRI deidentification and AnonyMI, a new method combining re-identification risk reduction and geometrical preservation. (Hoboken, USA: John Wiley & Sons, Inc.) (2021) 42(17):5523–34. doi: 10.1101/2021.07.30.454335

22. Cardenas CE, Mohamed ASR, Yang J, Gooding M, Veeraraghavan H, Kalpathy-Cramer J, et al. Head and neck cancer patient images for determining auto-segmentation accuracy in T2-weighted magnetic resonance imaging through expert manual segmentations. *Med Phys* (2020) 47:2317–22. doi: 10.1002/mp.13942

23. Alfaro-Almagro F, Jenkinson M, Bangerter NK, Andersson JL, Griffanti L, Douaud G, et al. Image processing and quality control for the first 10,000 brain imaging datasets from UK biobank. *Neuroimage* (2018) 166:400–24. doi: 10.1016/j.neuroimage.2017.10.034

24. Milchenko M, Marcus D. Obscuring surface anatomy in volumetric imaging data. *Neuroinformatics* (2013) 11:65–75. doi: 10.1007/s12021-012-9160-3

25. Gulban O, Nielson D, Poldrack R, Gorgolewski C. *Poldracklab/pydeface: v2.0.0* . Available at: https://github.com/poldracklab/pydeface.

26. Jeong YU, Yoo S, Kim Y-H, Shim WH. De-identification of facial features in magnetic resonance images: Software development using deep learning technology. *J Med Internet Res* (2020) 22:e22739. doi: 10.2196/22739

27. Khazane A, Hoachuck J, Gorgolewski KJ, Poldrack RA. DeepDefacer: Automatic removal of facial features *via* U-net image segmentation. *arXiv* (2022) arXiv:2205.15536. doi: 10.48550/arXiv.2205.15536

28. Wahid KA, Ahmed S, He R, van Dijk LV, Teuwen J, McDonald BA, et al. Evaluation of deep learning-based multiparametric MRI oropharyngeal primary tumor auto-segmentation and investigation of input channel effects: Results from a prospective imaging registry. *Clin Transl Radiat Oncol* (2022) 32:6–14. doi: 10.1016/j.ctro.2021.10.003

29. McDonald BA, Cardenas C, O'Connell N, Ahmed S, Naser MA, Wahid KA, et al. Investigation of autosegmentation techniques on T2-weighted MRI for off-line dose reconstruction in MR-linac adapt to position workflow for head and neck cancers. *medRxiv* (2021). doi: 10.1101/2021.09.30.21264327

30. Taku N, Wahid KA, van Dijk LV, Sahlsten J, Jaskari J, Kaski K, et al. Auto-detection and segmentation of involved lymph nodes in HPV-associated oropharyngeal cancer using a convolutional deep learning neural network. *Clin Transl Radiat Oncol* (2022) 36:47–55. doi: 10.1016/j.ctro.2022.06.007

31. Naser MA, van Dijk LV, He R, Wahid KA, Fuller CD. *Tumor segmentation in patients with head and neck cancers using deep learning based-on multi-modality PET/CT images*. Springer (2020) p. 85–98.

32. Naser MA, van Dijk LV, He R, Wahid KA, Fuller CD. Tumor segmentation in patients with head and neck cancers using deep learning based-on multi-modality PET/CT images. *In Head and Neck Tumor Segmentation: First Challenge, HECKTOR 2020, Held in Conjunction with MICCAI 2020, Proceedings 1 2021*; Lima, Peru: Springer International Publishing (2020) p. 85–98.

33. Naser MA, Wahid KA, Grossberg AJ, Olson B, Jain R, El-Habashy D, et al. Deep learning auto-segmentation of cervical skeletal muscle for sarcopenia analysis in

patients with head and neck cancer. *Front Oncol* (2022) 12. doi: 10.3389/fonc.2022.930432

34. Van Rossum G, Drake FLJr. Python Reference manual. *Centrum voor Wiskunde en Informatica Amsterdam* (1995).

35. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. Pytorch: An imperative style, high-performance deep learning library. *Adv Neural Inf Process Syst* (2019) 32:8026–37.

36. *The MONAI consortium*. Project MONAI (2020). doi: 10.5281/zenodo.4323059

37. Pérez-García F, Sparks R, Ourselin S. TorchIO: a Python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning. *Comput Methods Programs BioMed* (2021) 208:106236. doi: 10.1016/j.cmpb.2021.106236

38. Taha AA, Hanbury A. Metrics for evaluating 3D medical image segmentation: Analysis, selection, and tool. *BMC Med Imaging* (2015) 15:1–28. doi: 10.1186/s12880-015-0068-x

39. Sherer MV, Lin D, Elguindi S, Duke S, Tan L-T, Cacicedo J, et al. Metrics to evaluate the performance of auto-segmentation for radiation treatment planning: A critical review. *Radiother Oncol* (2021) 160:185–91. doi: 10.1016/j.radonc.2021.05.003

40. Wilcoxon F. Individual comparisons by ranking methods. In: *Breakthroughs in statistics*. (New York, USA: Springer) (1992). p. 196–202.

41. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Ser B Methodol* (1995) 57:289–300. doi: 10.1111/j.2517-6161.1995.tb02031.x

42. Muschelli J. Recommendations for processing head CT data. *Front Neuroinformatics* (2019) 13:61. doi: 10.3389/fninf.2019.00061

43. Lin D, Lapen K, Sherer MV, Kantor J, Zhang Z, Boyce LM, et al. A systematic review of contouring guidelines in radiation oncology: Analysis of frequency, methodology, and delivery of consensus recommendations. *Int J Radiat Oncol Biol Phys* (2020) 107:827–35. doi: 10.1016/j.ijrobp.2020.04.011

44. Abrams RA, Winter KA, Regine WF, Safran H, Hoffman JP, Lustig R, et al. Failure to adhere to protocol specified radiation therapy guidelines was associated with decreased survival in RTOG 9704–a phase III trial of adjuvant chemotherapy and chemoradiotherapy for patients with resected adenocarcinoma of the pancreas. *Int J Radiat Oncol Biol Phys* (2012) 82:809–16. doi: 10.1016/j.ijrobp.2010.11.039

45. Peters LJ, O'Sullivan B, Giralt J, Fitzgerald TJ, Trotti A, Bernier J, et al. Critical impact of radiotherapy protocol compliance and quality in the treatment of advanced head and neck cancer: Results from TROG 02.02. *J Clin Oncol* (2010) 28:2996–3001. doi: 10.1200/JCO.2009.27.4498

46. Ohri N, Shen X, Dicker AP, Doyle LA, Harrison AS, Showalter TN. Radiotherapy protocol deviations and clinical outcomes: a meta-analysis of cooperative group clinical trials. *J Natl Cancer Inst* (2013) 105:387–93. doi: 10.1093/jnci/djt001

47. Salzillo TC, Taku N, Wahid KA, McDonald BA, Wang J, van Dijk LV, et al. Advances in imaging for HPV-related oropharyngeal cancer: Applications to radiation oncology. In: *Seminars in radiation oncology*. Elsevier (2021). p. 371–88.

48. Kaissis G, Ziller A, Passerat-Palmbach J, Ryffel T, Usynin D, Trask A, et al. End-to-end privacy preserving deep learning on multi-institutional medical imaging. *Nat Mach Intell* (2021) 3:473–84. doi: 10.1038/s42256-021-00337-8