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

Authors: Jaakko Sahlsten[1], Kareem A. Wahid[2], Enrico Glerean[3,4], Joel Jaskari[1], Mohamed A. Naser[2], Renjie He[2], Clifton D. Fuller[2]*, 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 USA
[3] Department of Neuroscience and Biomedical Engineering, Aalto University, Espoo, Finland
[4] International Laboratory of Social Neurobiology, Institute for Cognitive Neuroscience, HSE University, Moscow, Russian Federation
* co-corresponding authors

Corresponding authors contact information: cdfuller@mdanderson.org; kimmo.kaski@aalto.fi.

Funding Acknowledgements: This work is supported by the International Laboratory of Social Neurobiology ICN HSE RF Government grant ag. No. 075–15- 2019–1930 (to E.G.). This work was also supported by the National Institutes of Health (NIH)/National Cancer Institute (NCI) through a Cancer Center Support Grant (CCSG; P30CA016672-44). M.A.N. is supported by an NIH grant (R01DE028290-01). K.A.W. 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). C.D.F. 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).

Abstract

Background and Purpose: Increased demand for head and neck cancer (HNC) radiotherapy data for algorithmic development has prompted an escalation of head and neck image dataset sharing. Medical images must comply with data protection requirements so that data re-use is enabled without disclosing direct or indirect identifiers of study participants. Defacing, i.e., the removal of facial features, is often considered an acceptable compromise between data protection and re-usability. While existing defacing tools have been developed by the neuroimaging community, their acceptability for preserving data that is crucial to radiotherapy applications have not been previously explored. Therefore, we systematically investigated the impact of currently available defacing algorithms on HNC organs at risk (OARs) used for radiotherapy planning.

Materials and Methods: We utilized a publicly available dataset of T2-weighted magnetic resonance imaging (MRI) scans for 55 HNC patients with eight previously segmented OARs (right/left submandibular glands, right/left parotid glands, right/left level II neck lymph nodes, right/left level III neck lymph nodes). Four publicly available defacing algorithms were investigated: mri_deface, defacer, fsl_deface, and pydeface. We implemented a 5-fold cross-validation 3D U-net based OAR auto-segmentation model 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 model evaluation on original and defaced data. These models were primarily investigated using the Dice similarity coefficient (DSC). The mean surface distance (MSD), and mean Hausdorff distance at 95% (MHD_{95}) were also explored.

Results: Of the studied algorithms, mri_deface and defacer were unable to produce any usable images for evaluation, while fsl_deface and pydeface were unable to remove the face for 18 and 24% of subjects, respectively. All the remaining subjects had voxels removed from all OARs. Composite OAR DSC using defaced data for model training and evaluation yielded statistically worse mean DSC performance ($p \leq 0.05$) for both fsl_deface (training = 0.770, evaluation = 0.709) and pydeface (training = 0.631, evaluation = 0.453) compared to the original data (0.780). Significance testing demonstrated that the DSC and MHD_{95} were significantly better ($p \leq 0.05$) for the model trained with the original data for both submandibular glands and for the right parotid gland compared to the model trained with fsl_deface data. Similarly, the MSD was significantly better ($p \leq 0.05$) for the model trained with the original data for the left submandibular gland and both parotid glands. When trained with pydeface data, the model had significantly worse performance ($p \leq 1e\text{-}5$) compared to the model trained with the original data for all structures and metrics.

Conclusion: We demonstrate the choice of defacing algorithm has a significant impact on HNC OAR auto-segmentation model development and deployment. Our work incentivizes the further development of HNC radiotherapy-specific defacing methods.

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 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. Imaging is crucial to radiotherapy workflows, particularly for organ at risk (OAR) and tumor segmentation [4,5]. Moreover, in recent years public data competitions, such as the HECKTOR challenge [6], have targeted improving the radiotherapy workflow. However, there is a particular facet of medical image dissemination for radiotherapy applications that has spurred controversy, namely anonymization of head and neck cancer (HNC) imaging.

While the public dissemination of HNC data is invaluable to improvement of 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 [7]. This policy introduces some uncertainty in the dissemination of high-resolution images where the intricacies of facial features can be reconstructed to generate similar “comparable” visualizations with relative ease. Several studies have shown the potential danger in releasing unaltered medial images containing facial features, as they can often be easily recognized by humans and/or machines [8–10]. For example, using facial recognition software paired with the MRI-derived facial reconstructions, one study found up to 83% of research participants could be identified from their MRI scan [10]. 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 applications in neuro-imaging [11–14], there have yet to be 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 maintaining patient anonymity, we conducted a systematic analysis of existing methods for facial anonymization on HNC MRI images. Through qualitative and quantitative analysis using open-source datasets and tools, we determine the effects of defacing approaches on whole images and structures relevant to radiation treatment planning and clinical endpoints. Moreover, we test the effects of these approaches on auto-segmentation, a specific domain application that is increasingly relevant for HNC public datasets. Our 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 we used a publicly available dataset hosted on the TCIA, the AAPM RT-MAC Grand Challenge 2019 (AAPM) dataset [15]. The AAPM dataset consists of T2-weighted MRI scans of 55 patients that are labeled for OAR segmentations of: 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 images and segmentations can be found in the corresponding data descriptor [15]. For segmentation model building the data 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. Segmentation performance was reported 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.

Defacing methods

For defacing the images we used the same methods as taken into consideration by Schwartz et al. [11], as well as novel tools which benefit from recent advances in deep learning. The most popular tools use a co-registration to a template in order to identify face + ears and then identify those structures in the original image that should be removed or blurred. We tested the following co-registration based methods: *fsl_deface* [16], *pydeface* [17], and *mri_deface* [13]. A more recent method using deep learning technology (*defacer*) was also included [18]; in this method facial features (i.e., eyes, nose, mouth, ears) are automatically drawn and then blurred. An automated pipeline for applying all these defacing methods is available at https://github.com/eglerean/faceai_testingdefacing. Each defacing method was implemented such that a defaced volume was produced as well as a volumetric mask of which voxels were affected by defacing. After applying the defacing methods, we evaluated: 1) the performance of the defacing method by visually inspecting all defaced volumes (“1” if eyes, nose, and mouth were removed, “0” if the defacing failed); 2) the amount of voxels preserved in the 8 structures. Ideally, defacing is successful if no voxels from the structures are affected, while all facial features from the image are successfully removed.

Deep learning model

To evaluate OAR segmentation performance under different defacing schemes from volumetric MRI data, we utilized 3D U-net, a convolutional neural network architecture that is commonly used in the medical domain and has found wide success in HNC segmentation

tasks [19–23]. The model includes a contractive pathway, where feature maps are down-sampled with max-pooling, and an expanding pathway, where feature maps are upscaled with transpose convolutional layers. Both pathways include four blocks, where each block consists of two convolutional layers with kernel size 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. In addition, the output feature maps of each contractive block are concatenated to the input feature maps of each expanding pathway block. Layers on the contractive pathway blocks have 32, 64, 128 and 256 feature maps and layers on the expanding pathway blocks have 128, 64 and 32 feature maps. The last convolutional layer has 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 [24] using Pytorch 1.8.1 [25] with a U-net model from Project MONAI 0.7.0 [25] and data preprocessing and augmentation with TorchIO 0.18.61 [27].

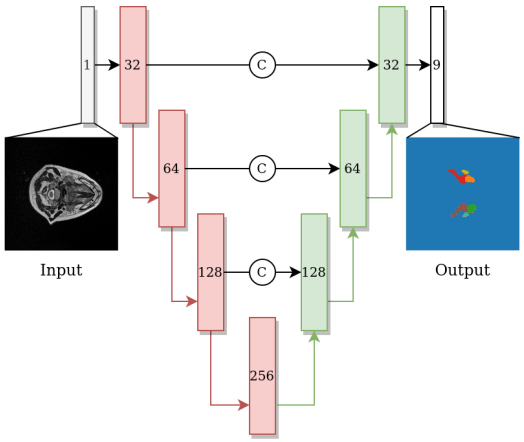

Figure 1. U-net network architecture with blocks on the contractive path colored in red and blocks on the expanding path colored in green. Each block includes two convolutions, each followed by instance normalization and Leaky ReLU activation, followed by a max-pool layer (red arrow) or transpose convolution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

Data preprocessing included linear resampling to 2 mm isotropic resolution with intensity scaled into a range of [-1,1]. The training data was augmented with 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). Model training was stopped early after 30 epochs of non-improvement of the validation loss.

Segmentation evaluation

We evaluated model segmentation performance with the Dice similarity coefficient (DSC), mean Surface Distance (MSD), and mean Hausdorff distance at 95% (MHD_{95}) defined as follows:

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

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

$$MHD_{95} = \frac{1}{2}\left(\{max_{P_{95}}\, d(t,P) \mid t \in T\} + max_{P_{95}}\{d(p,T)) \mid p \in P\}\right),$$

with the distance from 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 [28,29]. The model output was resampled into original resolution with nearest-neighbor sampling and evaluated against original resolution segmentations. The performance of MSD and MHD_{95} was evaluated in millimeters. When comparing the segmentation model metrics, we implemented Wilcoxon signed rank tests [30], with p-values less than or equal to 0.05 considered as significant. Statistical comparisons were performed using the statannotations Python package (https://github.com/trevismd/statannotations).

Results

Defacing performance

Two of the methods tested (mri_deface and defacer) failed for all subjects in the dataset we considered. Therefore, for all subsequent analysis only the fsl_deface and pydeface were considered. There was scanwise quality control for each defacing method with 10 (18%) scans removed from fsl_deface and 13 (24%) with pydeface. Poor quality defaced scans were ignored in the evaluation of defacing and segmentation performance. A barplot comparison of the ratio of remaining OAR voxels after defacing and quality control is depicted in **Figure 2**. In addition, defacing methods removed some OARs completely which were also omitted from segmentation evaluation. After filtering unusable data, the total number of OARs evaluated in segmentation was 440 for the original data, 334 for fsl_deface, and 268 for pydeface. A full comparison of omitted OARs is shown in **Table 1**.

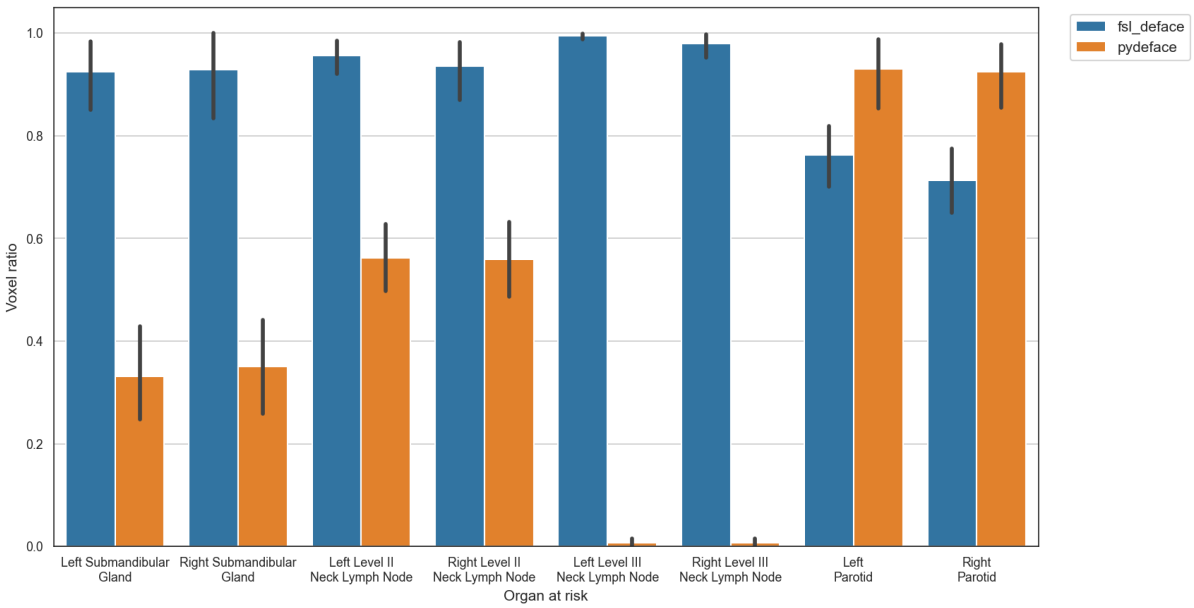

Figure 2. Ratio of voxels in comparison to the original segmentation mask after defacing (fsl_deface and pydeface) for each of the organs at risk, where defacing was successful. The mean and standard deviation are represented in the middle and extremes of the error bars, respectively.

	fsl_deface			pydeface		
Organ at risk	Defacing failed*	Completely removed	Either	Defacing failed*	Completely removed	Either
Left Submandibular Gland	13	2	14	10	6	13
Right Submandibular Gland	13	1	14	10	7	14
Left Neck Lymph Node Level II	13	0	13	10	4	11
Right Neck Lymph Node Level II	13	0	13	10	4	11
Left Neck Lymph Node Level III	13	0	13	10	45	51
Right Neck Lymph Node Level III	13	1	13	10	44	50
Left Parotid	13	0	13	10	6	11
Right Parotid	13	0	13	10	6	11
Total omitted	**104**	**4**	**106**	**80**	**122**	**172**

Table 1. Number of organs at risk that were labeled with failed defacing and were completely removed with defacing or either. *Counted from scanwise quality control.

Segmentation performance

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 original and

fsl_deface input data had the highest composite mean DSC when evaluated on the original target data (0.780 and 0.770, respectively), while the model trained on pydeface input data had the highest composite mean DSC when evaluated on pydeface target data (0.678). In contrast, the models trained using original and fsl_deface input data has the lowest composite mean DSC when evaluated on pydeface target data (0.453, 0.531, respectively), while the model trained using pydeface input data has the lowest composite mean DSC when evaluated on fsl_deface target data (0.577). All results on each evaluation target are significantly different ($p \leq 0.05$) based on the Wilcoxon signed-rank test.

	Evaluated on original (N=440)	**Evaluated on fsl_deface (N=334)**	**Evaluated on pydeface (N=268)**
Trained on original	0.780 (0.106)	0.709 (0.153)	0.453 (0.322)
Trained on fsl_deface	0.770 (0.104)	0.755 (0.123)	0.531 (0.307)
Trained on pydeface	0.631 (0.236)	0.577 (0.231)	0.678 (0.251)

Table 2. Composite DSC performance - mean (standard deviation) - of all structures for all combinations of training data (rows) and evaluation data (columns). 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 ($p \leq 0.05$) as determined by the Wilcoxon signed-rank test.

To determine the impact of defacing on algorithmic development, we compared models trained on original vs. defaced data using the original target data for evaluation. Our analysis was based on 8 OAR structure segmentations from 55 patients totaling 440 evaluations. We omitted the evaluation of MSD and MHD_{95} for 4 cases with empty output segmentation maps for the models trained on the original data (right submandibular gland = 1) and pydeface (right level III lymph node = 3). Full comparisons of the model performance for each OAR are depicted in **Figure 3**. Overall, with a few exceptions, the model trained with original data performed better than the model trained with the defaced data for the majority of structures and all evaluation metrics. Specifically, the DSC and MHD_{95} performance was significantly better ($p \leq 0.05$) for the model trained with the original data for both the right and left submandibular glands and the right parotid gland compared to the model trained with fsl_deface data; similarly, the MSD performance was significantly better ($p \leq 0.05$) for the model trained with the original data for the left submandibular gland and both the right and left parotid glands. When trained with pydeface data, the model had significantly worse performance ($p \leq 1e\text{-}5$) compared to the model trained with original data for all structures and metrics.

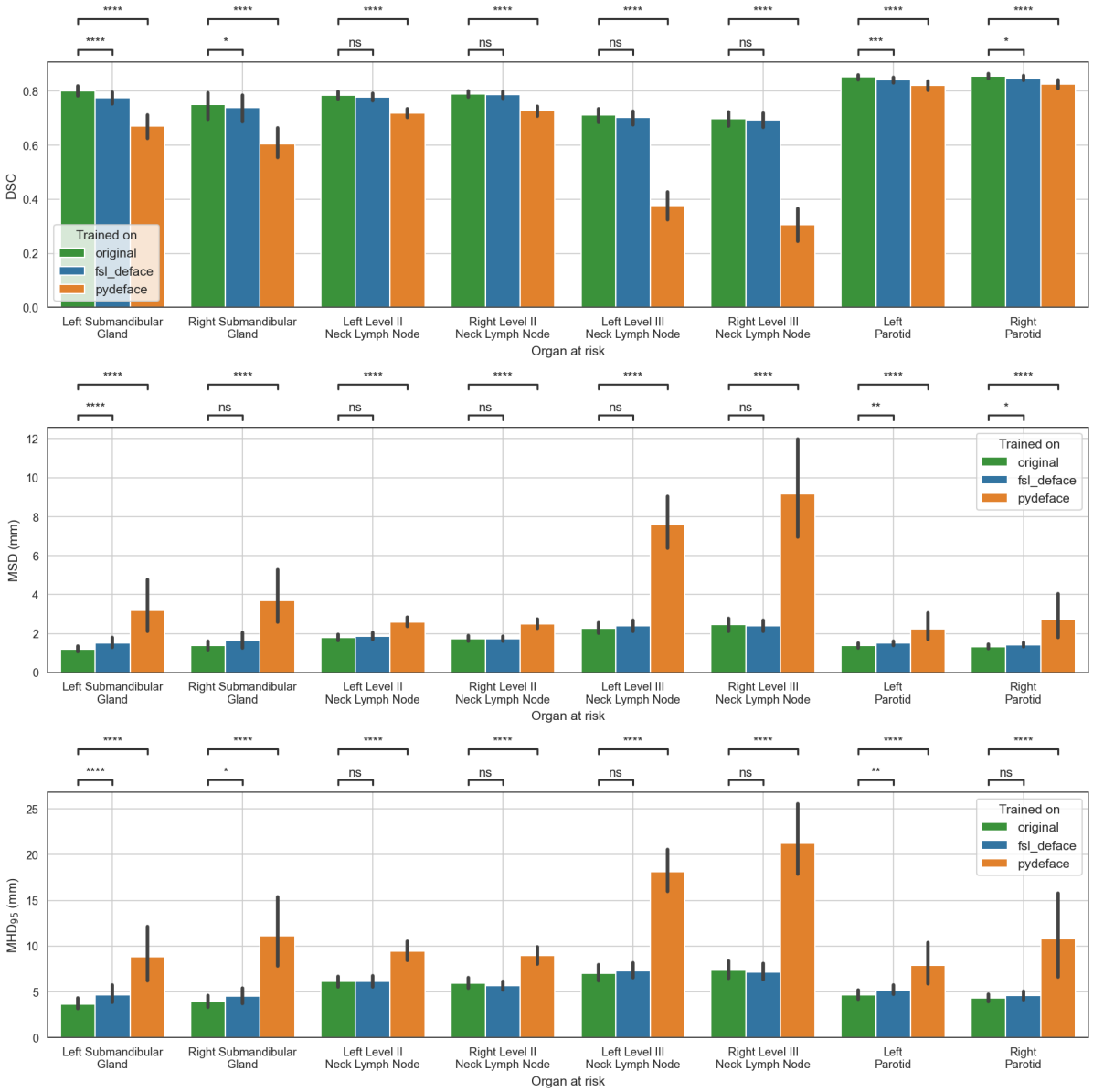

Figure 3. Performance of models trained on original or defaced data and evaluated on the original data. The mean and standard deviation for each metric are represented in the middle and extremes of the error bars, respectively. Statistical significance was determined using the Wilcoxon signed-rank test. Comparison symbols: ns ($p > 0.05$), * ($p \leq 0.05$), ** ($p \leq 0.01$), *** ($p \leq 1e-4$), **** ($p \leq 1e-5$).

To determine the impact of defacing on algorithmic deployment, we trained a model on original data and compared model evaluation based on original and defaced data. In these results, we only included valid target data by filtering images where defacing failed and empty segmentation structures for all the defacing methods. This resulted in using results from 34 left submandibular glands, 34 right submandibular glands, 36 left neck level II lymph nodes, 36 right neck level II lymph nodes, 3 left neck level III lymph nodes, 4 right neck level III lymph nodes, 36 left parotid glands, and 36 right parotid glands. Due to the low number of cases for right and left level III lymph nodes they were removed from the comparison. In addition, for the MSD and MHD_{95} metrics, empty model output segmentations were discarded resulting in evaluation of 31 and 21 left submandibular glands and 32 and 22 right

submandibular glands on fsl_deface and pydeface, respectively. The model evaluated on the original data performed better on all structures for all the evaluation metrics when compared to evaluation on defaced data. Specifically, the model evaluated on original data had significant differences ($p \leq 0.01$) on all the metrics and structures (except submandibular glands) compared evaluation on fsl_deface data and significant differences ($p \leq 1e\text{-}5$) on all metrics and structures when compared to evaluation on pydeface data. The full comparison of model performance for the 6 OARs is shown in **Figure 4**.

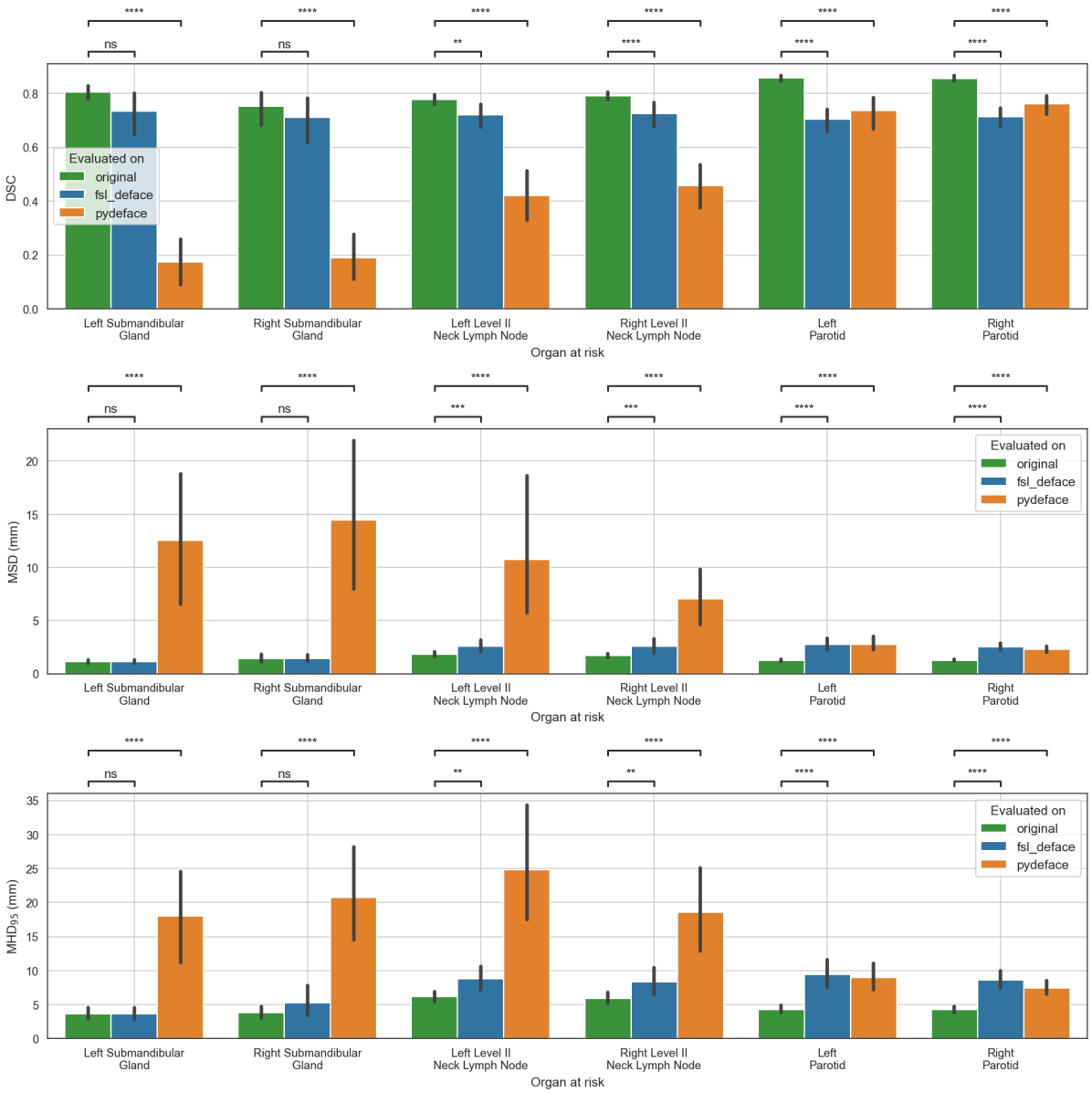

Figure 4. The performance of models trained on original data when evaluated on original, fsl_deface, or pydeface data for 6 organs-at-risk. We only include cases that were available for all methods: 36 segmentations were used for all structures except the submandibular glands which used 34 segmentations. In addition, for the MSD and MHD_{95} metrics, empty model output segmentations were discarded resulting in evaluation of 31 and 21 left submandibular glands and 32 and 22 right submandibular glands on fsl_deface and pydeface, respectively. The mean and standard deviation for each metric are represented as the middle and extremes of the error bars, respectively. Statistical significance was

measured with the Wilcoxon signed-rank test. Comparison symbols: ns ($p > 0.05$), * ($p \leq 0.05$), ** ($p \leq 0.01$), *** ($p \leq 1e-4$), **** ($p \leq 1e-5$).

Discussion

In this study we have systematically investigated the impact of a variety of defacing algorithms on structures of interest used for radiotherapy treatment planning. We demonstrate that the overall usability of segmentations is heavily dependent on the choice of the defacing algorithm. Moreover, we demonstrate that several OARs have the potential to be negatively impacted by defacing algorithms such that developed auto-segmentation algorithms severely underperform, compared to algorithms trained and evaluated on non-defaced data. To our knowledge, this is the first investigation on the impact of data anonymization in radiotherapy-related applications.

We investigated four commonly available defacing algorithms: mri_deface, defacer, fsl_deface, and pydeface. It is noted that two algorithms (mri_deface and defacer) out of four investigated, were unable to yield any usable images with our dataset. For mri_deface, the most likely reason was that the template used expected scans to cover the whole brain, while the field of view of the images of the AAPM dataset were mostly covering the neck and mouth, often leaving the top of the brain excluded. For defacer, we used the trained model provided with the algorithm, based on a public neuroimaging dataset; the field-of-view of AAPM dataset images were different from those used for training the model, which then failed for all subjects. For those defacing algorithms that were able to generate usable images (fsl_deface, pydeface), we demonstrate that regardless of the choice of the algorithm, 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 fsl_deface for all the OAR structures with the exception of the parotid glands, indicating its potential for greater downstream impact of radiotherapy algorithmic development and evaluation. While fsl_deface in many cases leads to relatively minimal reduction of available voxels, the loss of topographic information in a radiotherapy workflow cannot be underscored. 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 [31–34]. Therefore, voxel-level information loss of OARs secondary to the defacing algorithms, while potentially visibly imperceptible, as noted for the fsl_deface, has the capability to radically hamper the development of algorithms aimed for clinical translation.

In this study we developed an OAR auto-segmentation workflow to highlight the potential impact of defacing-induced voxel-level information loss on downstream radiotherapy applications. Through this workflow we demonstrate that using defaced data as training inputs to a deep learning model decreases performance when evaluated on non-defaced data compared to using original non-defaced data as training inputs. Moreover, even in cases where models trained with defaced input data are evaluated on corresponding defaced target data, the performance is still on average comparatively worse than the evaluation on original non-defaced data. This indicates a clear deficit when using these

defaced data for algorithmic development. Moreover, the issues of using defaced data in radiotherapy algorithmic workflows are not limited to model training, but also translate to deficiencies in model deployment, as we demonstrate decreased performance of an algorithm developed on original non-defaced data evaluated on defaced data.

Most studies aimed at evaluating defacing algorithms naturally focus on their ability to properly anonymize data. Relatively few studies have been conducted that were aimed at determining the downstream analysis effects of defacing algorithms. For example, recent studies by Schwartz et al. [11] and Mikulan et al. [35] demonstrated that several defacing methods showed differences in the downstream neuroimaging effects, namely brain measurements and electroencephalography-related calculations, respectively. As we demonstrate on both pooled analysis and investigation of individual OARs for auto-segmentation model training and evaluation, performance is often modestly decreased for the fsl_deface and greatly decreased for pydeface. While pydeface is often shown to have favorable outcomes for facial anonymization as shown in previous studies [14,35], the trade-off for certain applications, such as radiotherapy imaging, becomes apparent. Therefore, in cases where defacing is unavoidable, fsl_deface should likely be preferred for radiotherapy segmentation applications.

Our study has several limitations. To examine defacing methods as they are currently distributed ("out of the box"), we did not perform any modifications to the templates or models utilized in any methods. More suitable templates for head-and-neck images would improve the defacing performance and in the case of the method utilizing the deep learning (defacer), re-training of the model with head-and-neck images could make the method more applicable to the data considered in this study. Moreover, we have not investigated newer more recently developed defacing methods. Future work should incorporate newer methods; the development of further methods that are better equipped to deface data used for the radiotherapy workflow are also warranted. Additionally, while we ensure a robust analysis by utilizing multiple relevant metrics established in existing literature [28] to evaluate our OAR auto-segmentation workflow, our workflow only acts as a proxy for clinically-related radiotherapy tasks. Importantly, there is not always a perfect correlation between spatial similarity metrics and radiotherapy plan acceptability [29]. In our study, we have not tested the downstream effects of defacing on radiotherapy plan generation, which may lead to different results from what we have observed for OAR segmentation. Furthermore, we have chosen to focus on OARs that were already available in the AAPM dataset and deemed important for radiotherapy applications; additional OARs and target volumes should be investigated in future studies. Finally, as an initial exploration of defacing methods for radiotherapy applications, we have only investigated a single imaging modality (T2-weighted MRI). The HNC radiotherapy workflow commonly incorporates computed tomography and positron emission tomography [36], so the effect of available defacing methods on these imaging modalities should be the subject of future investigations.

Conclusions

In summary, by using a publicly available HNC MRI dataset, we have systematically investigated the effects of four established defacing algorithms, mri_deface, defacer,

fsl_deface, and pydeface, for radiotherapy applications. Specifically, we determine the impact of defacing directly on ground-truth HNC OARs and develop a deep learning based OAR auto-segmentation workflow to investigate the use of defaced data for algorithmic training and evaluation. We determine that mri_deface and defacer were unable to yield usable images from our datasets, and fsl_deface and pydeface decrease the total number of voxels in OARs and decrease performance of OAR auto-segmentation, with pydeface having more severe negative effects than fsl_deface. Our study is an important step towards ensuring widespread privacy-preserving dissemination of HNC imaging data without endangering data usability. Future studies should investigate the development of HNC radiotherapy-specific defacing methods, the impact of defacing on radiotherapy plan generation, the inclusion of a greater number of OARs and target structures, and the incorporation of additional imaging modalities.

References

1. Foster, E. D. & Deardorff, A. Open science framework (OSF). *J. Med. Libr. Assoc. JMLA* **105**, 203 (2017).
2. Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **3**, 1–9 (2016).
3. Clark, K. *et al.* The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. *J. Digit. Imaging* **26**, 1045–1057 (2013).
4. Press, R. H. *et al.* The use of quantitative imaging in radiation oncology: a quantitative imaging network (QIN) perspective. *Int. J. Radiat. Oncol. Biol. Phys.* **102**, 1219–1235 (2018).
5. Beaton, L., Bandula, S., Gaze, M. N. & Sharma, R. A. How rapid advances in imaging are defining the future of precision radiation oncology. *Br. J. Cancer* **120**, 779–790 (2019).
6. Andrearczyk, V. *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* 1–21 (Springer, 2020).
7. Meystre, S. M., Friedlin, F. J., South, B. R., Shen, S. & Samore, M. H. Automatic de-identification of textual documents in the electronic health record: a review of recent research. *BMC Med. Res. Methodol.* **10**, 1–16 (2010).

8. Prior, F. W. *et al.* Facial recognition from volume-rendered magnetic resonance imaging data. *IEEE Trans. Inf. Technol. Biomed.* **13**, 5–9 (2008).
9. Mazura, J. C. *et al.* Facial recognition software success rates for the identification of 3D surface reconstructed facial images: implications for patient privacy and security. *J. Digit. Imaging* **25**, 347–351 (2012).
10. Schwarz, C. G. *et al.* Identification of anonymous MRI research participants with face-recognition software. *N. Engl. J. Med.* **381**, 1684–1686 (2019).
11. Schwarz, C. G. *et al.* Changing the face of neuroimaging research: Comparing a new MRI de-facing technique with popular alternatives. *NeuroImage* **231**, 117845 (2021).
12. Schimke, N., Kuehler, M. & Hale, J. Preserving privacy in structural neuroimages. in *IFIP annual conference on data and applications security and privacy* 301–308 (Springer, 2011).
13. Bischoff-Grethe, A. *et al.* A technique for the deidentification of structural brain MR images. *Hum. Brain Mapp.* **28**, 892–903 (2007).
14. Theyers, A. E. *et al.* Multisite Comparison of MRI Defacing Software Across Multiple Cohorts. *Front. Psychiatry* **12**, 189 (2021).
15. Cardenas, C. E. *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.* **47**, 2317–2322 (2020).
16. Alfaro-Almagro, F. *et al.* Image processing and Quality Control for the first 10,000 brain imaging datasets from UK Biobank. *Neuroimage* **166**, 400–424 (2018).
17. Gulban, O., Nielson, D., Poldrack, R. & Gorgolewski, C. poldracklab/pydeface: v2.0.0 [Internet].
18. Jeong, Y. U., Yoo, S., Kim, Y.-H. & Shim, W. H. De-identification of facial features in magnetic resonance images: software development using deep learning technology. *J. Med. Internet Res.* **22**, e22739 (2020).
19. Wahid, K. A. *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.* **32**, 6–14 (2022).

20. Naser, M. A. *et al.* Head and Neck Cancer Primary Tumor Auto Segmentation using Model Ensembling of Deep Learning in PET-CT Images. *medRxiv* (2021).
21. Naser, M. A., van Dijk, L. V., He, R., Wahid, K. A. & Fuller, C. D. Tumor segmentation in patients with head and neck cancers using deep learning based-on multi-modality PET/CT images. in 85–98 (Springer, 2020).
22. Naser, M. A. *et al.* Deep Learning Auto-Segmentation of Cervical Neck Skeletal Muscle for Sarcopenia Analysis Using Pre-Therapy CT in Patients with Head and Neck Cancer. *medRxiv* 2021.12.19.21268063 (2021).
23. McDonald, B. A. *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).
24. Van Rossum, G. & Drake Jr, F. L. *Python reference manual*. (Centrum voor Wiskunde en Informatica Amsterdam, 1995).
25. Paszke, A. *et al.* Pytorch: An imperative style, high-performance deep learning library. *Adv. Neural Inf. Process. Syst.* **32**, 8026–8037 (2019).
26. The MONAI Consortium. *Project MONAI*. http://doi.org/10.5281/zenodo.4323059 (2020).
27. 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.* 106236 (2021).
28. Taha, A. A. & Hanbury, A. Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. *BMC Med. Imaging* **15**, 1–28 (2015).
29. Sherer, M. V. *et al.* Metrics to evaluate the performance of auto-segmentation for radiation treatment planning: A critical review. *Radiother. Oncol.* (2021).
30. Wilcoxon, F. Individual comparisons by ranking methods. in *Breakthroughs in statistics* 196–202 (Springer, 1992).
31. Lin, D. *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.* **107**, 827–835 (2020).

32. Abrams, R. A. *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.* **82**, 809–816 (2012).
33. Peters, L. 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.* **28**, 2996–3001 (2010).
34. Ohri, N. *et al.* Radiotherapy protocol deviations and clinical outcomes: a meta-analysis of cooperative group clinical trials. *J. Natl. Cancer Inst.* **105**, 387–393 (2013).
35. Mikulan, E. *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*. (2021).
36. Salzillo, T. C. *et al.* Advances in Imaging for HPV-Related Oropharyngeal Cancer: Applications to Radiation Oncology. in *Seminars in radiation oncology* vol. 31 371–388 (Elsevier, 2021).