ORIGINAL RESEARCH

Empowering Data Sharing in Neuroscience: A Deep Learning De-identification Method for Pediatric Brain MRIs

Ariana M. Familiar[1,2], Neda Khalili[1,2], Nastaran Khalili[1,2], Cassidy Schuman[3], Evan Grove[3], Karthik Viswanathan[1,2], Jakob Seidlitz[4,5,6], Aaron Alexander-Bloch, MD[4,5,6], Anna Zapaishchykova[7,8], Benjamin H. Kann[7,8], Arastoo Vossough[1,9,10], Phillip B. Storm[1,2], Adam C. Resnick[1,2], Anahita Fathi Kazerooni[1,2,11], Ali Nabavizadeh[1,10] *

ABSTRACT

BACKGROUND AND PURPOSE: Privacy concerns, such as identifiable facial features within brain scans, have hindered the availability of pediatric neuroimaging datasets for research. Consequently, pediatric neuroscience research lags adult counterparts, particularly in rare disease and under-represented populations. The removal of face regions (image defacing) can mitigate this, however existing defacing tools often fail with pediatric cases and diverse image types, leaving a critical gap in data accessibility. Given recent NIH data sharing mandates, novel solutions are a critical need.

MATERIALS AND METHODS: To develop an AI-powered tool for automatic defacing of pediatric brain MRIs, deep learning methodologies (nnU-Net) were employed using a large, diverse multi-institutional dataset of clinical radiology images. This included multi-parametric MRIs (T1w, T1w-contrast enhanced, T2w, T2w-FLAIR) with 976 total images from 208 brain tumor patients (Children's Brain Tumor Network, CBTN) and 36 clinical control patients (Scans with Limited Imaging Pathology, SLIP) ranging in age from 7 days to 21 years old.

RESULTS: Face and ear removal accuracy for withheld testing data was the primary measure of model performance. Potential influences of defacing on downstream research usage were evaluated with standard image processing and AI-based pipelines. Group-level statistical trends were compared between original (non-defaced) and defaced images. Across image types, the model had high accuracy for removing face regions (mean accuracy, 98%; *N*=98 subjects/392 images), with lower performance for removal of ears (73%). Analysis of global and regional brain measures (SLIP cohort) showed minimal differences between original and defaced outputs (mean r_S=0.93, all $p < 0.0001$). AI-generated whole brain and tumor volumes (CBTN cohort) and temporalis muscle metrics (volume, cross-sectional area, centile scores; SLIP cohort) were not significantly affected by image defacing (all r_S>0.9, p<0.0001).

CONCLUSIONS: The defacing model demonstrates efficacy in removing facial regions across multiple MRI types and exhibits minimal impact on downstream research usage. A software package with the trained model is freely provided for wider use and further development (pediatric-auto-defacer; https://github.com/d3b-center/pediatric-auto-defacer-public). By offering a solution tailored to pediatric cases and multiple MRI sequences, this defacing tool will expedite research efforts and promote broader adoption of data sharing practices within the neuroscience community.

ABBREVIATIONS: AI = artificial intelligence; CBTN = Children's Brain Tumor Network; CSA = cross-sectional area; SLIP = Scans with Limited Imaging Pathology; TMT = temporalis muscle thickness; NIH = National Institute of Health; LH = left hemisphere; RH = right hemisphere.

Received month day, year; accepted after revision month day, year.

[1] Center for Data-Driven Discovery in Biomedicine (D3b), Children's Hospital of Philadelphia, Philadelphia, PA, USA.
[2] Department of Neurosurgery, Children's Hospital of Philadelphia, Philadelphia, PA, USA.
[3] School of Engineering and Applied Science, University of Pennsylvania, Philadelphia, PA, USA.
[4] Department of Child and Adolescent Psychiatry and Behavioral Science, The Children's Hospital of Philadelphia, Philadelphia, PA, USA.
[5] Department of Psychiatry, University of Pennsylvania, Philadelphia, PA, USA.
[6] Lifespan Brain Institute at the Children's Hospital of Philadelphia and University of Pennsylvania, Philadelphia, PA, USA.
[7] Artificial Intelligence in Medicine (AIM) Program, Mass General Brigham, Harvard Medical School, Boston, MA, USA.
[8] Department of Radiation Oncology, Dana-Farber Cancer Institute and Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA.
[9] Division of Radiology, Children's Hospital of Philadelphia, Philadelphia, PA, USA.
[10] Department of Radiology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA.
[11] AI2D Center for AI and Data Science for Integrated Diagnostics, University of Pennsylvania, Philadelphia, PA, USA.

Author Dr. Aaron Alexander-Bloch has equity in Centile Biosciences, receives consulting fee for Octave bioscience. Author Dr. Jakob Seidlitz has equity in Centile Biosciences and is Board Member of Centile Biosciences.

Please address correspondence to Ali Nabavizadeh, MD, Department of Radiology, Hospital of the University of Pennsylvania, 3400 Spruce St., Philadelphia, PA 19104, USA; Ali.Nabavizadeh@pennmedicine.upenn.edu .

SUMMARY SECTION

PREVIOUS LITERATURE: Scientific data sharing promotes reproducibility of research and translation of findings into clinical care. Several centralized repositories have enabled broad sharing of large-scale imaging datasets; however, pediatric datasets have lagged their adult counterparts, and neuro-imaging data is particularly challenging to share due to privacy concerns, as brain scans can reveal identifiable features. Existing "defacing" tools to remove face regions are primarily designed for adult scans, and often struggle with pediatric images and do not generalize to a variety of sequence types. This work introduces the first tool (pediatric-auto-defacer) specifically for removing facial features from multi-parametric pediatric MRIs, addressing a critical gap in data sharing for neuroscience research.

KEY FINDINGS: A model was developed to automatically remove facial regions from brain MRIs for anonymization purposes. It performs well on several sequence types across various acquisition parameters, and does not over-remove brain tissue. Based on testing, defacing does not affect downstream analytical pipelines (e.g., image pre-processing or measured group-level trends).

KNOWLEDGE ADVANCEMENT: To facilitate broad sharing of pediatric neuro-imaging datasets, a robust, automatic de-identification tool is provided to ease the burden on research teams to prepare and release imaging data while protecting patient privacy. This will accelerate neuroscience research and clinical trials in pediatrics and ultimately empower scientific discoveries.

INTRODUCTION

Data sharing is a critical component of research endeavors as it lends to scientific transparency and data reuse. For the study of rare diseases, data sharing is crucial for gathering a meaningful group of samples to enable statistical comparisons in the given patient population. Due to calls to action across disciplines, data sharing plans have recently become a mandate for NIH (National Institute of Health)-funded projects and deposit of data files to centralized repositories is now a requirement by many scientific journals for publication. Such efforts will facilitate the reproducibility of research studies and consequently their translation into real-world applications such as clinical care contexts, as well as bolster the inclusion of historically under-represented populations, which can mitigate bias in developed models and support fair AI in healthcare[1].

In alignment with FAIR[2] principles, several imaging data repositories have been established such as the Alzheimer's Disease Neuroimaging Initiative (ADNI)[3] and the National Cancer Institute's The Cancer Imaging Archive (TCIA)[4] and Imaging Data Commons (IDC), which provide effective data discovery and accessibility. While several large-scale, multi-institutional imaging datasets exist, such as the NLST for lung cancer (chest CTs from over 26,000 patients)[5] and the Breast Cancer Screening Digital Breast Tomosynthesis (breast mammograms from 5,060 patients)[6], comparable radiology datasets in neuroscience fields have lagged their counterparts, primarily due to greater difficulty of removing identifying information from brain (head and neck) scans. Brain images can be inherently identifiable due to the presence of an individual's face, and their release can jeopardize patient privacy. Studies have shown brain MRIs can be used to identify subjects by matching to their photograph[7,8], even after face regions have been blurred[9]. "Defacing", or the removal of face regions in an image, is one way to mitigate this issue, and several defacing software tools for structural brain MRIs have been developed (e.g., mri_deface[10], pydeface[11], fsl_deface[12] and others[13,14]), some of which have less impact on downstream processing than others[15,16]. That said, existing tools do not typically perform well on pediatric cases[17], particularly in young children and infants, likely due to differences in brain and face anatomy across developmental stages. For example, one study found that FSL's defacing removed brain tissue in the majority of children (ages 8-11) and in some young adult (ages 19-31) cases, and had worse performance for eyes and mouth removal compared to adults[18]. FreeSurfer had better performance for face removal without impacting brain tissue in the same cases, however it was more invasive in removing intraorbital and brain stem structures. Many tools rely on alignment to standardized face or brain atlases created with adult MRIs, and therefore fail to properly deface pediatric scans. Additionally, most are developed for T1-weighted (T1w) sequences, and there remains a need for accessible tools for defacing additional sequence types collected under standard clinical imaging protocols (e.g., T2-weighted (T2w)).

Pediatric data sharing has been significantly hindered by regulatory barriers related to privacy concerns, creating a critical unmet need for public imaging datasets. Herein, we build a tool to enable automatic removal of face regions from multiple types of pediatric MRIs, with the goal of facilitating data sharing across neuroscience fields. This is, to the best of our knowledge, the first available pediatric defacing tool. To address the need for a tool that can operate across multi-parametric MRIs, we use a large, multi-institutional clinical radiology dataset (Children's Brain Tumor Network; CBTN[19]) with deep learning AI methods to develop a model for minimally invasive defacing. Our model was trained and validated with 208 pediatric brain tumor subjects (832 total images) and 36 clinical control subjects (144 images from the Scans with Limited Imaging Pathology (SLIP) cohort[20]), with four image sequences included per subject (T1w, T1w contrast-enhanced (T1w-CE), T2w, and T2w Fluid Attenuated Inversion Recovery (FLAIR) sequences). Images were acquired through clinical protocols, and thus capture real-world heterogeneity in scanner and image acquisition properties.

MATERIALS AND METHODS

Patient cohorts

Retrospective data was collected from the CBTN[19]; a large-scale, multi-institutional repository of longitudinal clinical, imaging, genomic, and other paired data[21]. 208 subjects were selected based on imaging availability and inclusion of a range of ages at the time of imaging (median age 8; min=0.35, max=21.71 years) and cancer histologies (Figures 1, S1, & S2; Tables 1 & S1). MRI scans were unprocessed images from treatment-naïve clinical exams (T1w, T1w-CE, T2w, and T2w-FLAIR). All subjects had histologically confirmed pediatric brain tumors.

To test generalizability to non-brain tumor patients (clinical control group), a cohort of 40 subjects with available images from the SLIP[20] dataset were selected to match the general distributions of age and sex of the CBTN cohort. 36 subjects had sufficient images and were included in the main analyses.

Ground truth creation with semi-automated face mask segmentation

Preliminary face masks were generated for each image using the MiDeface[22] algorithm and then were manually edited. 507 of the 976 images (52%) were found to be inaccurately defaced and were manually revised using the ITK-SNAP[23] software (by authors C.S., E.G.; *Supp. Methods*). The criteria for an accurate face mask was that any brain region or temporalis muscle (given potential implications as a biomarker[24]) were not affected and identifiable facial features, including eyes, nose, mouth, and ears were fully included. Common corrections included restoring brain voxels, particularly in the right prefrontal cortex, and properly realigning the face mask to the subject's face.

Table 1: Patient characteristics in the studied cohorts.

Patient Characteristics	Training/Validation *CBTN*	Internal Testing *CBTN*	External Testing *CBTN*	Clinical Control Testing *SLIP*
Multicenter	Yes	No	Yes	No
Total Patients	146	37	25	36
Total Images	584	148	100	144
Age at imaging, range (years)	0.35 - 19.7	0.84 - 21.71	1.08 – 17.69	0.23 – 17.33
Age at imaging, median (years)	7.8	11.13	5.94	7.16
Legal Sex (No. (%))				
Male	79 (54%)	18 (49%)	14 (56%)	19 (53%)
Female	66 (45%)	19 (51%)	11 (44%)	17 (47%)
Unknown	1 (1%)			
Race (No. (%))				
White	100 (68%)	24 (65%)	16 (64%)	25 (69%)
Black or African American	20 (14%)	4 (11%)	4 (16%)	6 (17%)
Asian	2 (1%)	2 (5%)		1 (3%)
Native Hawaiian or Other Pacific Islander	1 (1%)			
American Indian or Alaska Native	1 (1%)			
More than one race	1 (1%)			
Other/Unavailable/Not Reported	21 (14%)	7 (19%)	5 (20%)	4 (11%)
Ethnicity (No. (%))				
Not Hispanic or Latino	130 (89%)	30 (81%)	22 (88%)	9 (25%)
Hispanic or Latino	8 (5%)	5 (14%)	2 (8%)	3 (8%)
Unavailable	8 (5%)	2 (5%)	1 (4%)	24 (67%)
Histology (No. (%))				
Low-Grade Glioma/Astrocytoma	87 (60%)	22 (59%)	21 (84%)	N/A
Medulloblastoma	40 (27%)	8 (22%)		
High-Grade Glioma/Astrocytoma	9 (6%)	3 (8%)	4 (16%)	
High-Grade Glioma/DIPG	9 (6%)	3 (8%)		
Ganglioglioma	1 (1%)			
Unknown/Not Available		1 (3%)		
Scanner Magnetic Field Strength (T) (No. (%))				
3	95 (65%)	26 (70%)	9 (36%)	36 (100%)
1.5	51 (35%)	11 (30%)	16 (64%)	
Scanner Manufacturer (No. (%))				
Siemens	134 (92%)	33 (89%)	16 (64%)	36 (100%)
GE	10 (7%)	4 (11%)	9 (36%)	
Philips	1 (1%)			
Toshiba	1 (1%)			

AI deep learning model development

CBTN images were stratified into training/validation and testing sets (80-20 split) based on demographics (age, sex, race) and histology (Table 1). nnU-Net[25] v1 (https://github.com/MIC-DKFZ/nnUNet/tree/nnunetv1; 3D full resolution; *Supp. Methods*) was used with 5-fold cross-validation, initial learning rate 0.01, stochastic gradient descent (SGD) with Nesterov momentum (μ=0.99), and number of epochs=1000 x 250 minibatches. Each unprocessed T1w/T1w-CE/T2w/FLAIR sequence was treated as a separate input. The set of 4 images for each subject could be used for either training or validation but not both (i.e., images from a single subject could not be split into training and validation within a given fold). Given a large percentage of the CBTN scans were from CHOP, we additionally split the testing cohort into "internal" (CHOP) and "external" (4 separate institutions) testing datasets.

Defacing accuracy

Model performance was evaluated with (previously unseen) images in the testing cohorts. Traditional performance scores such as the Sørensen-Dice score (spatial overlap between model predicted mask and ground truth mask), sensitivity (percent of pixels correctly identified by the model), and 95% Hausdorff distance metrics (distances between nearest voxels in the predicted and ground truth masks, of which 95% of voxels fell within) were generated.

As an additional assessment of defacing accuracy, two raters (authors Ne.K. and Na.K.) evaluated model performance in the testing cohorts. For each image, they rated coverage of the eyes and ears (separately for left and right), mouth, and nose with either: 1 (fully covered), 0.75 (approximately 75% masked), 0.5 (50% masked), 0.25 (25% masked), or 0 (not masked at all); and whether any brain tissue was removed (Yes/No). After initial independent review, images with disagreement were reviewed until a consensus was reached.

Impact of defacing on downstream analytics

Given the overarching aim to facilitate data sharing of brain MRIs for research purposes, it is essential any modification of the images by defacing minimally impacts downstream analysis. Several methods were used to assess this using standard image processing steps, in both the brain tumor (CBTN) and non-brain tumor (SLIP) groups separately.

Pre-processing and application of pre-trained AI models: For each subject in the CBTN testing cohorts, T1w, T2w, and FLAIR sequence images were co-registered with their corresponding T1w-CE sequence and resampled to an isotropic resolution of 1 mm^3 based on the anatomical SRI24 atlas[26] using the Greedy algorithm (https://github.com/pyushkevich/greedy)[27] in the Cancer Imaging Phenomics Toolkit open-source software v.1.8.1 (CaPTk, https://www.cbica.upenn.edu/captk)[28]. Accuracy of co-registration was confirmed by visual assessment of the 4 images.

Pre-processed data for each subject was then input into existing pretrained AI models for automatic brain tissue extraction and tumor subregion segmentation (https://github.com/d3b-center/peds-brain-seg-pipeline-public)[29,30]. This was performed once using the original images (non-defaced), and once using the defaced images. Resulting brain and tumor segmentation masks were compared between these conditions.

Cortical and subcortical volumetric measures: For 31 subjects in the SLIP testing cohort, their T1w scan was input to FreeSurfer's reconstruction pipeline (recon-all; https://surfer.nmr.mgh.harvard.edu/fswiki/recon-all)[31] to generate cortical and subcortical structure parcellations (5 subjects were excluded due to insufficient T1w image quality). This was performed once with original images and once with defaced images. Resulting volumetric measurements based on the parcellations were compared between these conditions.

We additionally used an existing AI-powered pipeline to estimate the thickness (TMT) and cross-sectional area (CSA) of the temporalis muscle (https://doi.org/10.5281/zenodo.8428986)[24] for 28 SLIP subjects (5 subjects excluded for insufficient quality T1w images, 3 subjects excluded for being younger than 3 years of age as required by the tool).

Please see *Supplemental Materials* for a description of all statistical comparisons and a CLAIM checklist to indicate alignment with the proposed methodological guidelines recommended for AI in medical imaging[32–34].

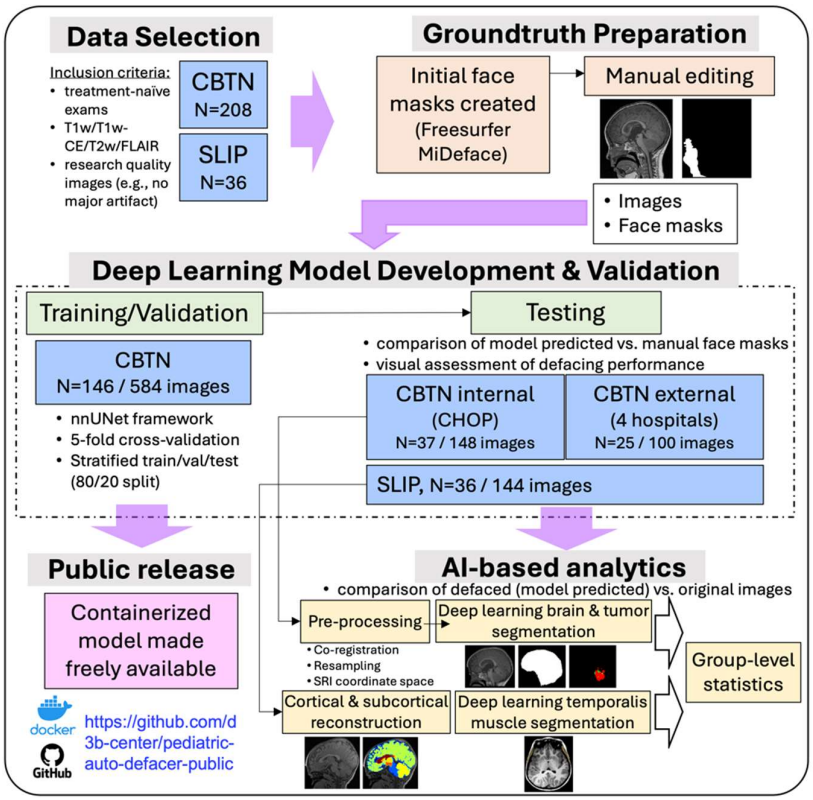

FIG 1. Diagram of overall study workflow. Data cohorts included brain tumor (CBTN) and non-brain tumor control (SLIP). Initial ground truth face masks were created with MiDeface and manually edited. A 3D deep learning model was trained with the nnUNet framework, using a single image as input, and tested on withheld data. The impact of defacing on downstream image processing and AI-based pipelines was evaluated with CBTN and SLIP testing data. The trained model is provided in an open-source software container on GitHub.

RESULTS

Defacing Accuracy

Across images, dice scores indicated decent spatial overlap between manual ground truth and model-predicted face masks in the internal (Mean=0.78, Median=0.8, Standard Error of the Mean (SEM)=0.008), external (Mean=0.75, Median=0.78, SEM=0.02), and clinical control (Mean=0.75, Median=0.77, SEM=0.01) groups (Fig. 2). Repeated-measures ANOVAs confirmed there was no effect of image type (T1w/T1w-CE/T2w/FLAIR) on dice scores in the internal ($F(3,108)=0.38$, $p=0.77$) and external ($F(3,72)=1.8$, $p=0.16$) cohorts, however there was a significant effect in the clinical control group ($F(3,105)=6.14$, $p=0.007$) with better model performance for T2w and FLAIR compared to T1w and T1w-CE (Table S2). Pearson correlations showed no effect of age on dice scores averaged across image types (internal: $r(35)=0.19$, $p=0.25$; external: $r(23)=0.29$, $p=0.17$; control: $r(34)=0.28$, $p=0.095$; Fig. S3). One-way ANOVAs indicated no effect of sex (internal: $F(1,35)=2.0$, $p=0.17$; external: $F(1,23)=0.28$, $p=0.6$; control: $F(1,34)=3.17$, $p=0.08$) or race (internal: $F(3,33)=0.18$, $p=0.911$; external: $F(2,22)=0.61$, $p=0.551$; control: $F(2,32)=1.07$, $p=0.356$) on dice scores, and no effect of histopathological diagnosis (internal: $F(4, 32) = 0.442$, $p = 0.777$; external: $F(1, 23) = 0.377$, $p = 0.545$) or general tumor location (internal: $F(4,32) = 0.837$, $p = 0.512$; external: $F(3,21) = 0.1$, $p = 0.959$) in the CBTN testing cohorts.

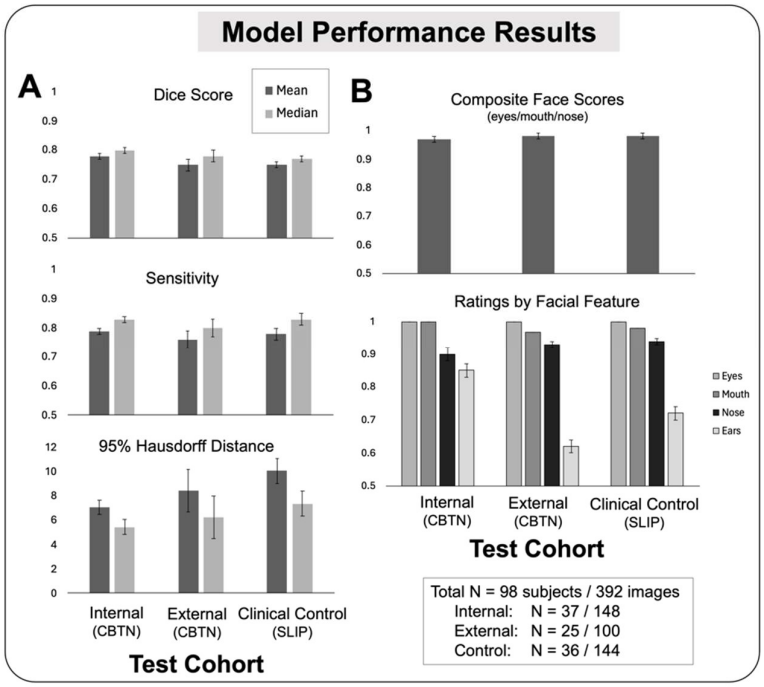

FIG 2. Model performance results. Plots show aggregate metrics across image types for each testing cohort (see Table S1 for results for image type separately); error bars represent standard error of the mean. (A) Standard metrics for segmentation evaluation including dice similarity, sensitivity, and 95% Hausdorff distance; (B) Average performance ratings based on visual inspection by two raters (1=fully covered, 0.75=approximately 75% masked, 0.5=50% masked, 0.25=25% masked, 0=not masked at all).

On further review, it was determined that the spatial metrics were not an ideal measure of defacing performance due to variability in extension of the face mask into the air in front of the face in the ground truth segmentations (Figures 3 & S4). To more accurately assess model performance, two raters (Ne.K., Na.K.) reviewed each defaced image in the internal, external, and clinical control testing groups. After applying the model-predicted face masks to the corresponding images, the raters were instructed to score the model's accuracy in masking (coverage of) the left eye, right eye, nose, mouth, left ear, and right ear separately (1=fully masked, 0.75/0.5/0.25=% partially masked, 0=not masked) for each image separately.

Across facial features, the average rated accuracy of model defacing was high for each testing set (Means: internal=0.93, external=0.86, control=0.89). Composite scores combining the eyes, mouth, and nose ratings indicated high masking performance for these features (Fig. 2, Table S2; internal=0.97, external=0.98, control=0.98), while performance for masking the ears was lower (internal=0.85, external=0.62, control=0.72). For every image, both raters reported no brain voxels were impacted by defacing in the internal, external, or clinical control groups. Repeated-measures ANOVAs showed a significant effect of image type on defacing performance in the clinical control group ($F(3,75)=10.8$, $p<0.0001$), with higher average ratings for T1w ($M=0.91$) and T1w-CE ($M=0.91$) compared to T2w ($M=0.89$) and FLAIR ($M=0.86$); but no effect of image type in the internal ($F(3,108)=1.17$, $p=0.33$) or external ($F(3,72)=0.32$, $p=0.81$) groups. Average rating across subjects and image types for each feature is displayed in Figure S5.

Assessing impact of defacing on downstream analytics

Pre-processing and application of pre-trained AI models: Defaced and original (non-defaced) images underwent pre-processing and were input to pretrained AI tools to assess any impact of defacing on standard downstream analysis using all four image sequences (T1w/T1w-CE/T2w/FLAIR). Visual inspection showed equivalent co-registration performance between defaced and original images. For the pediatric brain tumor test datasets, the volumes of AI-generated brain masks were equivalent between defaced and non-defaced images (internal: $r_S(35) > 0.99$, $p < 0.0001$; external: $r_S(23) > 0.99$, $p < 0.0001$; Fig. 4 upper and middle). AI-generated tumor segmentations were also unaffected by defacing, indicated by equivalent volumes of contrast-enhancing tumor, non-enhancing tumor, cystic, and edema subregions (internal: all subregions $r_S(35) > 0.99$, $p < 0.0001$; external: all subregions $r_S(23) > 0.99$, $p < 0.0001$; Fig. 4; Table S3).

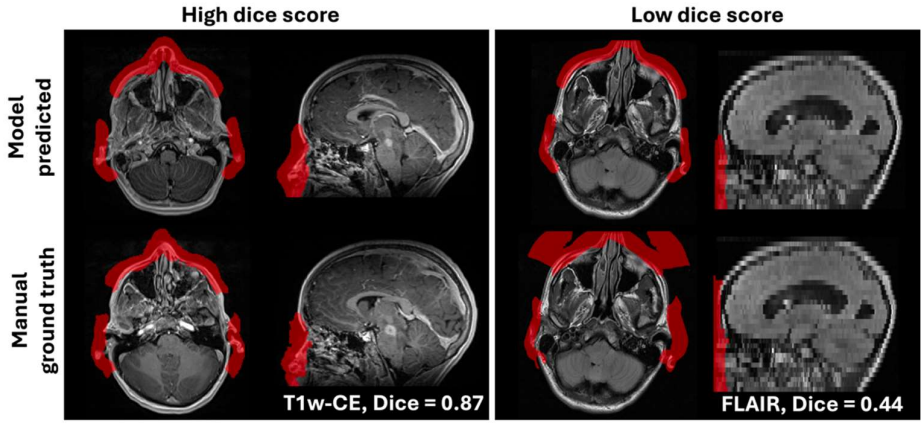

FIG 3. Representative example images of model predicted versus manual ground truth segmentation masks. Subjects shown with high (left box; T1w-CE sequence) and low (right box; FLAIR sequence) dice similarity scores between the model predicted (upper row) and manual ground truth (lower row) face masks. This illustrates how dice score, although a common metric for such segmentation tasks, was not an accurate measure of model performance in the present study, as ground truth masks were variable in their extension into space in front of the face (particularly due to "MiDeface" lettering imposed by the MiDeface Freesurfer tool that was used to generate initial face masks).

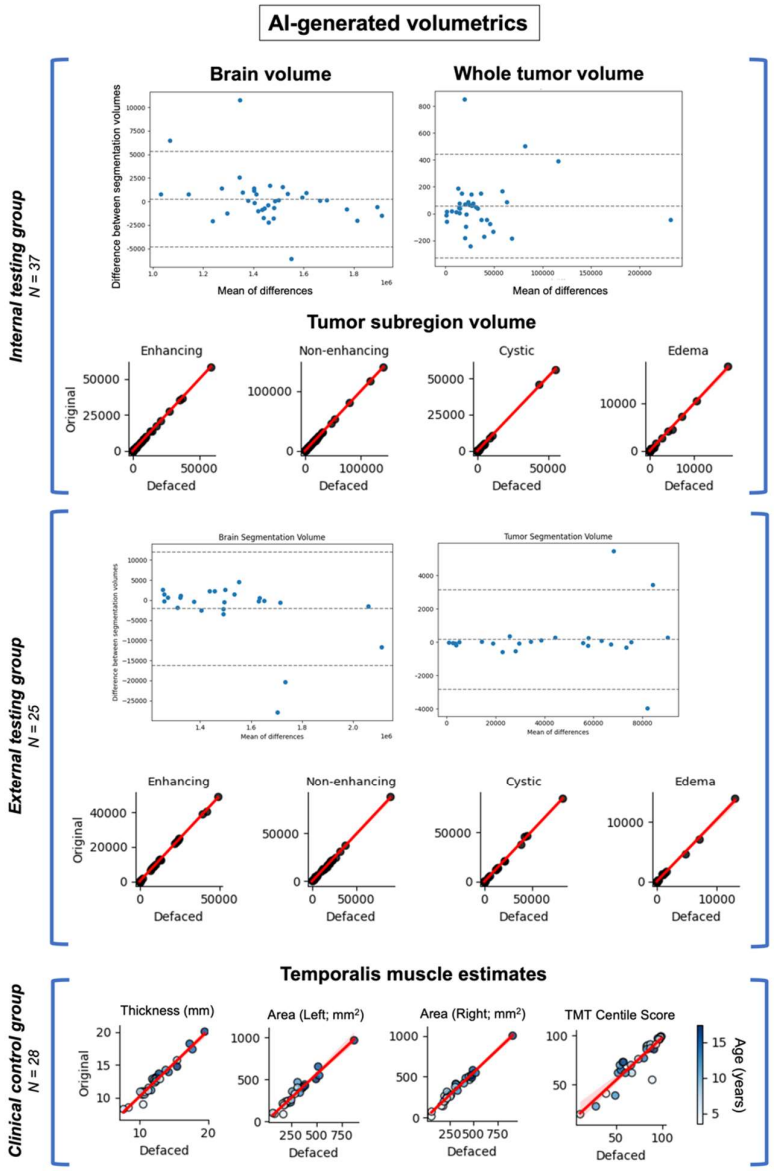

FIG 4. Testing the impact of defacing on AI-generated volumetrics. Each point represents one subject; the red line indicates a linear trend. Upper/middle: Comparison of tumor subregion volumes between defaced (x-axis) and original (y-axis) images in pediatric brain tumor subjects. There was very high agreement between brain and tumor segmentation volumes. Lower: Comparison of estimated temporal muscle thickness (TMT), area (CSA), and TMT centile scores between defaced (x-axis) and original (y-axis) T1w images from the clinical control group (point colors indicate age). Correlations indicated very high agreement between TM thickness, cross-sectional area, and resulting TMT centile scores.

Cortical and subcortical volumetric measures: For 31 subjects in the clinical control (SLIP) cohort, we further investigated any impact of defacing on derived brain measures from T1w images using a standard anatomical reconstruction pipeline (FreeSurfer recon-all). There was very high agreement between estimated global and regional measures, with all comparisons between original and defaced images being positively significant (mean $r_S(29)=0.93$, all $p<0.0001$; Table S4; Fig. S6). Correlations were above 0.9 for 48 out of 58 measures. Regions with the lowest agreement were the left and right cerebellum white matter (left: $r_S(29)=0.71$, $p<0.0001$; right: $r_S(29)=0.69$, $p<0.0001$). 9 global measurements (cortex, cerebral white matter, subcortical gray matter, total gray matter, supratentorial, brain segmentation, CSF, and total intracranial volumes) were equivalent between original and defaced ($r_S(29)>0.86$). Paired t-tests indicated no significant differences between original and defaced brain measures (Table S4; Fig. S6), with the exception of the right vessel (original $M=11.3$, $SEM=1.38$; defaced $M=14.7$, $SEM=2.19$; $t(30)=-2.32$, $p=0.03$) and the right hippocampus (original $M=3940.8$, $SEM=101$; defaced $M=3972.8$, $SEM=101$; $t(30)=-2.36$, $p=0.03$), which were estimated to be slightly larger on average in the defaced compared to original images. Overall, these results indicate defacing had minimal impact on cortical and subcortical volumetric assessments using a standard processing pipeline, which aligns with previous report of minimal effects of defacing tools on global FreeSurfer measurements[17].

To examine the impact of defacing on regional measurements in close proximity to the face, we extracted temporalis muscle thickness (TMT; mm) and cross-sectional area (CSA) measurements (SLIP cohort ages > 3 years; $N=28$) using an existing AI-powered pipeline[24] with T1w images. Notably, TM scores have been implicated as a predictive marker for sarcopenia across patient populations[35–38]. Spearman

correlations showed high agreement of estimated TMT ($r_S(26)=0.96$, all $p<0.0001$) and CSA (LH: $r_S(26)=0.96$, $p<0.0001$; RH: $r_S(26)=0.97$, $p<0.0001$; Fig. 4 lower) between defaced and original images. Paired t-tests indicated no difference in TMT volumes between original and defaced images ($t(27)=-1.8$, $p=0.08$), but a significant difference in CSA (LH: $t(27)=-3.74$, $p<0.0001$; RH: $t(27)=-4.79$, $p=0.0009$) with lower surface area estimates for the defaced (LH: *M*=306.2, *SEM*=30; RH: *M*=314.7, *SEM*=33) compared to original (LH: *M*=339.9, *SEM*=35; RH: *M*=350.5, *SEM*=37) images. Resulting centile scores based on TMT, age, and sex (compared with TMT distributions estimated from large-scale datasets[24]) were not significantly affected by defacing ($r_S(26)= 0.9$, $p<0.0001$; $t(27)=-0.97$, $p=0.34$).

DISCUSSION

Data sharing of MRIs is crucial to transparent and reproducible research, particularly in the era of predictive AI that requires ample volumes of representative data. Widely available pediatric imaging datasets are needed to accelerate discoveries in neuroscience, particularly in rare disease contexts. To this end, we aim to enable MRI data sharing through the development of an open-source de-identification tool for the automatic removal of identifiable facial features. A deep learning model for face masking was trained using a large, multi-institutional dataset of clinically acquired, multi-parametric MRIs (Children's Brain Tumor Network).

The trained model had strong performance removing the face (eyes, nose, mouth) in an unseen dataset, with adequate, though lower, performance on ear removal. This is potentially due to a lack of presence of ears in some images in the training dataset (limited field of view). Notably, although the model was trained on data from brain tumor patients, it could generalize to a separate dataset of clinically matched controls indicating its potential use across anatomically normal and disease-impacted cohorts. To enable wider usage by the community, the trained model is publicly provided as an open-source software package, and we encourage further model development to extend the model to additional disease and healthy populations (see potential clinical limitations in *Supp. Results*).

Critically, image alteration by defacing should not impact usage in intended research purposes. To ensure this, we compared the outputs of standard processing pipelines between defaced and original (non-defaced) images. Statistical trends for AI-estimated whole brain and tumor volumes (brain tumor group), in addition to derived brain region volumes, global brain metrics, and AI-generated temporalis muscle measurements (control group), were unaffected by defacing. Most estimated measures were equivalent between defaced and original images, and any resulting measurement differences did not impact overall patterns at a group-level. Thus, there was minimal impact of defacing on the utility of the structural images for downstream analysis with standard research pipelines.

Many existing defacing tools are limited to T1w sequences[13,22,39], and we sought to expand support to additional structural image types (T2w, FLAIR, T1w-CE), given their prevalence in clinical and research practices. That said, our tool is limited to four sequences, and further development could expand to additional types such as functional MRI and other advanced imaging (e.g., diffusion weighted imaging). Although consensus review was used to assess defacing performance, additional quantitative metrics such as face recognition rate may provide a more objective measure of de-identification performance. Another limitation of this study is that, while the training dataset included images across six institutions, a large portion of the dataset came from a single institution (CHOP). Future work should focus on expanding to larger studies to bolster model generalizability, and would benefit from direct comparison between deep learning and existing computer-vision methods.

CONCLUSIONS

In conclusion, we developed an AI-powered pediatric defacing tool with the goal of facilitating wider de-identification of structural MRIs for data sharing purposes. The tool is publicly available (https://github.com/d3b-center/pediatric-auto-defacer-public) and can be used on multiple image types. Future work can extend the model to additional populations and MR sequences to provide a universal method to facilitate data sharing and ultimately drive discoveries in neuroscience research.

ACKNOWLEDGMENTS

This project was supported in part by the NIH National Heart, Lung, and Blood Institute (grant number U2CHL156291 / 3U2CHL156291-02S1 to A.C.R.).

REFERENCES

1. Chen RJ, Wang JJ, Williamson DFK, et al. Algorithmic fairness in artificial intelligence for medicine and healthcare. *Nat Biomed Eng*. 2023;7(6):719-742. doi:10.1038/s41551-023-01056-8
2. Wilkinson MD, Dumontier M, Aalbersberg IjJ, et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data*. 2016;3(1):160018. doi:10.1038/sdata.2016.18
3. Mueller SG, Weiner MW, Thal LJ, et al. Ways toward an early diagnosis in Alzheimer's disease: the Alzheimer's Disease Neuroimaging Initiative (ADNI). *Alzheimers Dement*. 2005;1(1):55-66.
4. Prior FW, Clark K, Commean P, et al. TCIA: an information resource to enable open science. In: *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE; 2013:1282-1285.
5. National Lung Screening Trial Research Team, Aberle DR, Adams AM, et al. Reduced lung-cancer mortality with low-dose computed tomographic screening. *N Engl J Med*. 2011;365(5):395-409. doi:10.1056/NEJMoa1102873
6. Buda M, Saha A, Walsh R, et al. A data set and deep learning algorithm for the detection of masses and architectural distortions in digital breast tomosynthesis images. *JAMA Netw Open*. 2021;4(8):e2119100-e2119100.
7. Schwarz CG, Kremers WK, Therneau TM, et al. Identification of Anonymous MRI Research Participants with Face-Recognition Software. *N Engl J Med*. 2019;381(17):1684-1686. doi:10.1056/NEJMc1908881
8. 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(3):347-351. doi:10.1007/s10278-011-9429-3
9. Abramian D, Eklund A. Refacing: Reconstructing Anonymized Facial Features Using GANS. In: *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*. 2019:1104-1108. doi:10.1109/ISBI.2019.8759515

10. Bischoff-Grethe A, Ozyurt IB, Busa E, et al. A technique for the deidentification of structural brain MR images. *Hum Brain Mapp*. 2007;28(9):892-903. doi:10.1002/hbm.20312
11. Gulban OF, Nielson D, Poldrack R, Lee J, Gorgolewski KJ, Vanessasaurus Ghosh S. poldracklab/pydeface: v2. 0.0. *Zenodo https://doi org/105281/zenodo*. 2019;3524401.
12. Alfaro-Almagro F, Jenkinson M, Bangerter NK, et al. Image processing and Quality Control for the first 10,000 brain imaging datasets from UK Biobank. *NeuroImage*. 2018;166:400-424. doi:10.1016/j.neuroimage.2017.10.034
13. Khazane A, Hoachuck J, Gorgolewski KJ, Poldrack RA. DeepDefacer: Automatic Removal of Facial Features via U-Net Image Segmentation. Published online May 31, 2022. Accessed January 26, 2024. http://arxiv.org/abs/2205.15536
14. Milchenko M, Marcus D. Obscuring Surface Anatomy in Volumetric Imaging Data. *Neuroinform*. 2013;11(1):65-75. doi:10.1007/s12021-012-9160-3
15. de Sitter A, Visser M, Brouwer I, et al. Facing privacy in neuroimaging: removing facial features degrades performance of image analysis methods. *Eur Radiol*. 2020;30(2):1062-1074. doi:10.1007/s00330-019-06459-3
16. Rubbert C, Wolf L, Turowski B, et al. Impact of defacing on automated brain atrophy estimation. *Insights Imaging*. 2022;13(1):54. doi:10.1186/s13244-022-01195-7
17. Theyers AE, Zamyadi M, O'Reilly M, et al. Multisite Comparison of MRI Defacing Software Across Multiple Cohorts. *Front Psychiatry*. 2021;12. Accessed January 26, 2024. https://www.frontiersin.org/articles/10.3389/fpsyt.2021.617997
18. Buimer EEL, Schnack HG, Caspi Y, et al. De-identification procedures for magnetic resonance images and the impact on structural brain measures at different ages. *Human Brain Mapping*. 2021;42(11):3643-3655. doi:10.1002/hbm.25459
19. Familiar AM, Kazerooni AF, Anderson H, et al. A multi-institutional pediatric dataset of clinical radiology MRIs by the Children's Brain Tumor Network. Published online October 2, 2023. doi:10.48550/arXiv.2310.01413
20. Schabdach JM, Schmitt JE, Sotardi S, et al. Brain growth charts of "clinical controls" for quantitative analysis of clinically acquired brain MRI. *medRxiv*. Published online 2023:2023-01.
21. Lilly JV, Rokita JL, Mason JL, et al. The children's brain tumor network (CBTN) - Accelerating research in pediatric central nervous system tumors through collaboration and open science. *Neoplasia*. 2023;35:100846. doi:10.1016/j.neo.2022.100846
22. MiDeFace - Free Surfer Wiki. Accessed March 17, 2023. https://surfer.nmr.mgh.harvard.edu/fswiki/MiDeFace#Notes
23. Yushkevich PA, Pashchinskiy A, Oguz I, et al. User-Guided Segmentation of Multi-modality Medical Imaging Datasets with ITK-SNAP. *Neuroinform*. 2019;17(1):83-102. doi:10.1007/s12021-018-9385-x
24. Zapaishchykova A, Liu KX, Saraf A, et al. Automated temporalis muscle quantification and growth charts for children through adulthood. *Nat Commun*. 2023;14(1):6863. doi:10.1038/s41467-023-42501-1
25. Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat Methods*. 2021;18(2):203-211. doi:10.1038/s41592-020-01008-z
26. Rohlfing T, Zahr NM, Sullivan EV, Pfefferbaum A. The SRI24 multichannel atlas of normal adult human brain structure. *Hum Brain Mapp*. 2010;31:798-819. doi:10.1002/hbm.20906
27. Yushkevich PA, Pluta J, Wang H, Wisse LEM, Das S, Wolk D. Fast Automatic Segmentation of Hippocampal Subfields and Medial Temporal Lobe Subregions In 3 Tesla and 7 Tesla T2-Weighted MRI. *Alzheimers Dement*. 2016;12(7S_Part_2). doi:10.1016/j.jalz.2016.06.205
28. Pati S, Singh A, Rathore S, et al. The Cancer Imaging Phenomics Toolkit (CaPTk): Technical Overview. In: Crimi A, Bakas S, eds. *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*. Lecture Notes in Computer Science. Springer International Publishing; 2020:380-394. doi:10.1007/978-3-030-46643-5_38
29. Fathi Kazerooni A, Arif S, Madhogarhia R, et al. Automated tumor segmentation and brain tissue extraction from multiparametric MRI of pediatric brain tumors: A multi-institutional study. *Neuro-oncol adv*. 2023;5(1):vdad027. doi:10.1093/noajnl/vdad027
30. Vossough A, Khalili N, Familiar AM, et al. Training and Comparison of nnU-Net and DeepMedic Methods for Autosegmentation of Pediatric Brain Tumors. *American Journal of Neuroradiology*. Published online May 9, 2024. doi:10.3174/ajnr.A8293
31. Fischl B. FreeSurfer. *NeuroImage*. 2012;62(2):774-781.
32. Tejani AS, Klontzas ME, Gatti AA, et al. Checklist for Artificial Intelligence in Medical Imaging (CLAIM): 2024 Update. *Radiology: Artificial Intelligence*. 2024;6(4):e240300. doi:10.1148/ryai.240300
33. Mongan J, Moy L, Charles E Kahn J. Checklist for Artificial Intelligence in Medical Imaging (CLAIM): A Guide for Authors and Reviewers. *Radiology Artificial intelligence*. 2020;2(2):e200029. doi:10.1148/ryai.2020200029
34. Pham N, Hill V, Rauschecker A, et al. Critical Appraisal of Artificial Intelligence–Enabled Imaging Tools Using the Levels of Evidence System. *American Journal of Neuroradiology*. 2023;44(5):E21-E28. doi:10.3174/ajnr.A7850
35. Lee B, Bae YJ, Jeong WJ, Kim H, Choi BS, Kim JH. Temporalis muscle thickness as an indicator of sarcopenia predicts progression-free survival in head and neck squamous cell carcinoma. *Sci Rep*. 2021;11(1):19717.
36. Cho J, Park M, Moon WJ, Han SH, Moon Y. Sarcopenia in patients with dementia: correlation of temporalis muscle thickness with appendicular muscle mass. *Neurol Sci*. 2022;43(5):3089-3095. doi:10.1007/s10072-021-05728-8
37. Muglia R, Simonelli M, Pessina F, et al. Prognostic relevance of temporal muscle thickness as a marker of sarcopenia in patients with glioblastoma at diagnosis. *Eur Radiol*. 2021;31(6):4079-4086. doi:10.1007/s00330-020-07471-8
38. Nozoe M, Kubo H, Kanai M, et al. Reliability and validity of measuring temporal muscle thickness as the evaluation of sarcopenia risk and the relationship with functional outcome in older patients with acute stroke. *Clin Neurol Neurosurg*. 2021;201:106444.
39. Schwarz CG, Kremers WK, Wiste HJ, 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

SUPPLEMENTAL FILES

Table of Contents

Supplemental Methods

	Class	N subjects
Diagnosis		
	Low-grade glioma/astrocytoma (WHO grade I/II)	130
	Medulloblastoma	48
	High-grade glioma/astrocytoma (WHO grade III/IV)	16
	Brainstem glioma (diffuse intrinsic pontine glioma)	11
	Ganglioglioma	1
	Supratentorial PNET	1
	Not otherwise specified	1
General tumor location		
	Cerebellum/Posterior fossa	102
	Midline or brain stem	60
	Cortical	28
	Ventricles	10
	Optic pathway	7

Table S1. Diagnosis and tumor location information for CBTN training dataset.
Histopathologically-confirmed diagnosis and general tumor location (primary) for each subject in the model training cohort. "Midline or brain stem" location can include: pons, basal ganglia, thalamus, suprasellar/hypothalamic/pituitary, midbrain/tectum, and/or medulla.

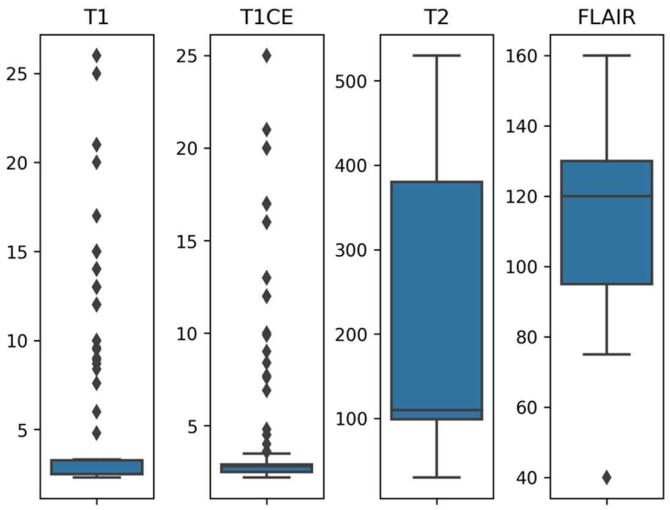

Figure S1. Distribution of TE in training dataset (CBTN; N subjects = 146; N images = 584) for T1w, T1w-CE, T2w, and FLAIR sequences.

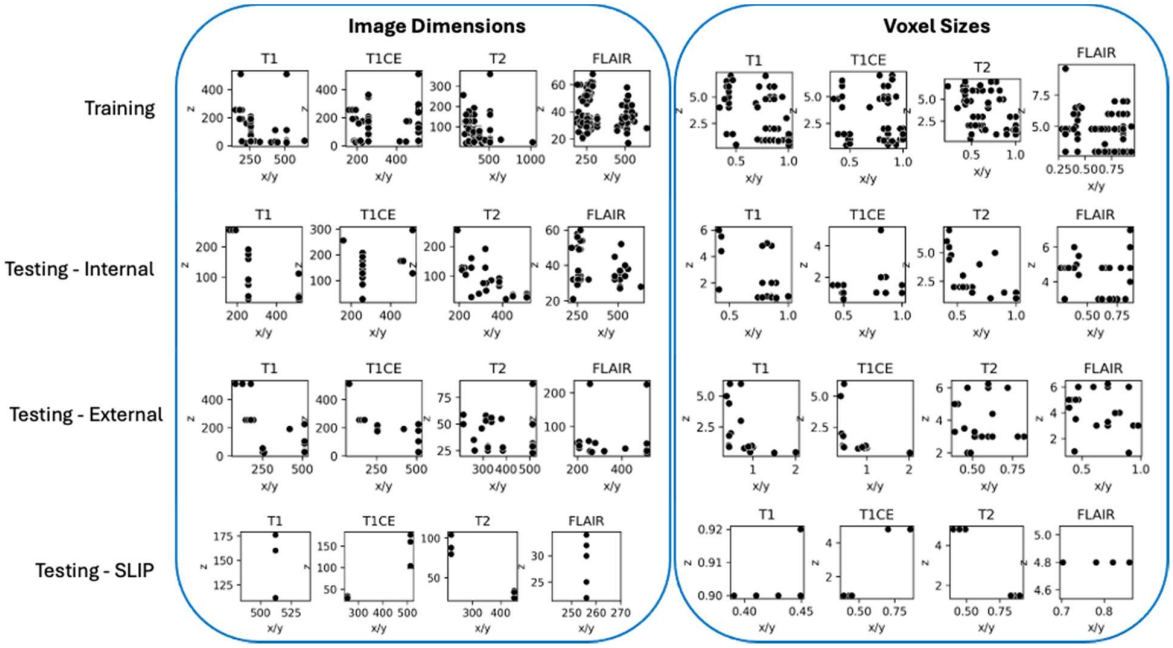

Figure S2. Distributions of image dimensions (left) and voxel sizes (right) for T1w, T1w-CE, T2w, and FLAIR sequences across each training and testing cohort.

nnU-Net

We employed the self-adaptive nnU-Net framework for model development (https://github.com/MIC-DKFZ/nnUNet/tree/nnunetv1) [1,2]. This deep learning method has shown strong performance across a variety of 2D and 3D image segmentation tasks (MRI, CT) and has outperformed other models in community benchmarking challenges, particularly in the context of generalization to new datasets with minimal over-fitting[3,4]. Because the model development is self-configured, manual tuning is not required to achieve strong predictive performance. The framework includes five-fold cross-validation using the training dataset to optimize the model's hyperparameters with an ensemble approach. Additionally, input images undergo standardized pre-processing and post-processing steps, allowing a complete pipeline that can receive various unprocessed images as input and produce uniform results (across parameters such as image size, resolution, intensity distributions). In the context of the present study, this allowed us to develop a robust tool that does not require the user to perform any image preparation and can be utilized across various imaging protocols at different scanners and institutions. For more details on the model architecture and nnU-Net methodology, please see original papers: Isensee et al., 2021a, 2021b.

Our training dataset consisted of 584 total images (146 T1w, 146 T1w-CE, 146 T2w, 146 FLAIR) from 146 CBTN patients. The model was trained to predict a single segmentation class.

1. Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat Methods*. 2021;18(2):203-211. doi:10.1038/s41592-020-01008-z

2. Isensee F, Jäger PF, Full PM, Vollmuth P, Maier-Hein KH. nnU-Net for Brain Tumor Segmentation. In: Crimi A, Bakas S, eds. *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*. Springer International Publishing; 2021:118-132. doi:10.1007/978-3-030-72087-2_11

3. Wang R, Lei T, Cui R, Zhang B, Meng H, Nandi AK. Medical image segmentation using deep learning: A survey. *IET Image Processing*. 2022;16(5):1243-1267. doi:10.1049/ipr2.12419

4. Jiang H, Diao Z, Shi T, et al. A review of deep learning-based multiple-lesion recognition from medical images: classification, detection and segmentation. *Computers in Biology and Medicine*. 2023;157:106726. doi:10.1016/j.compbiomed.2023.106726

Training procedures for generation of face masks

Two authors served as annotators (C.S., E.G.), each with neuroanatomy education and experience with medical imaging data (MRIs). In an initial training session 10 cases were reviewed, and several additional review sessions were conducted throughout the annotation process. Annotators were instructed to open a given image and initial face mask in ITK-Snap and scroll through the 3D image volumes (across axial, sagittal and coronal views) to assess coverage of each facial feature and make modifications as needed using the built-in editing tools. Generally, the mask was required to cover the superficial layer of skin, from forehead to jaw and all area between the outer cheeks and extending into the air in front of the face. With coverage of:

Feature	**Description**
Eyes	Orbital regions between left to right temple including upper nasal bone between the eyes.
Mouth	Lips and perioral region from jaw to nose and extending into either cheek regions up to nasolabial fold.
Nose	Full nasal region
Ear	All outer ear structures

Annotators were also instructed to ensure that the mask did not impact brain tissue voxels.

Statistical comparisons

Defacing Accuracy

Dice scores: Dice scores were calculated based on the spatial overlap of model-predicted versus ground truth face masks. Several statistical tests were used to assess the influence of different variables on dice scores, for each testing cohort separately (internal CBTN, external CBTN, SLIP). A repeated measures analysis of variance (ANOVA) was used to test for the effect of image sequence type (4 levels; within-subject variable) on dice scores (1 level; dependent variable). After averaging dice scores across image type, Pearson correlations were used to measure the linear relationship between subject age and dice score. One-way ANOVAs were used to measure the effect of subject sex (2 levels) or race (4) on dice scores. Additional ANOVAs tested the effect of diagnosis (internal: 5; external: 2) and general tumor location (internal: 5; external: 4) on dice scores of the CBTN test datasets. Paired t-tests were used to assess for differences in group mean dice score (defaced vs. original).

Manual ratings: A repeated measures ANOVA (4 x 1) was used to test the effect of image sequence type on composite ratings (average performance rating across eyes, mouth, & nose) for each testing cohort separately.

Assessing impact of defacing on downstream analytics

AI-predicted brain volume (internal & external CBTN cohorts): A Spearman correlation was used to measure the rank-order relationship between brain volume generated from defaced images, versus brain volume generated from original images. Paired t-tests assessed differences in group means (defaced vs. original) of predicted brain volume.

AI-predicted tumor segmentations (internal & external CBTN cohorts): Spearman correlations were used to measure the rank-order relationships between tumor subregion volumes (enhancing, non-enhancing, cystic, edema) generated from defaced images, versus tumor subregion volumes generated from original images. Paired t-tests assessed differences in group means (defaced vs. original) of whole tumor segmentation volume.

Freesurfer generated measures (SLIP cohort): Spearman correlations were used to assess the rank-order relationship of global and regional (cortical and subcortical) volumetric measurements between defaced and original T1w images. Paired t-tests were also used to assess for differences in group means (defaced vs. original) for each measurement separately.

AI-predicted temporalis muscle thickness (TMT), cross-sectional area (CSA), & centile scores (SLIP cohort): Spearman correlations were used to assess the rank-order relationship of TMT, CSA, and centile scores between defaced and original T1w images. Paired t-tests were also used to assess differences in group means (defaced vs. original) for each measure separately.

Supplemental Results

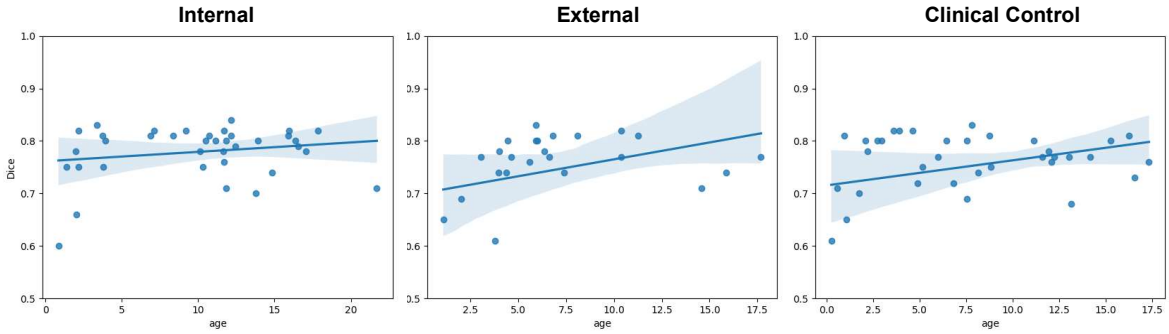

Figure S3. Correlations between subject age (x-axis) and dice scores across image types (y-axis) indicated no effect of age on model performance in any of the testing datasets. Each point represents one subject, line indicates linear regression fit to illustrate linear trend.

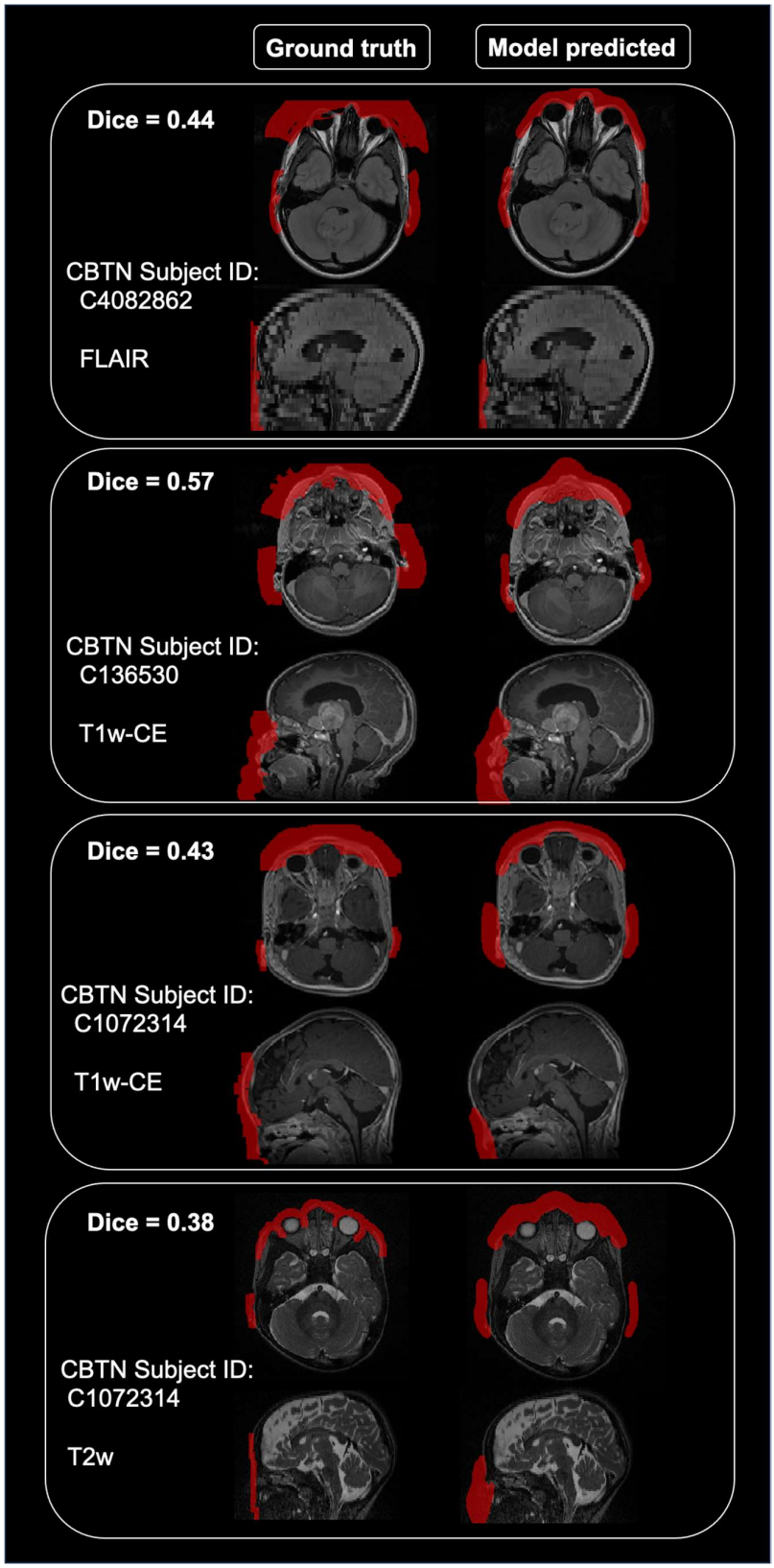

Figure S4. Additional examples of images with proper model-generated face masks, but low dice scores due to inconsistencies in the manually generated ground truth masks (e.g., variable extension into air).

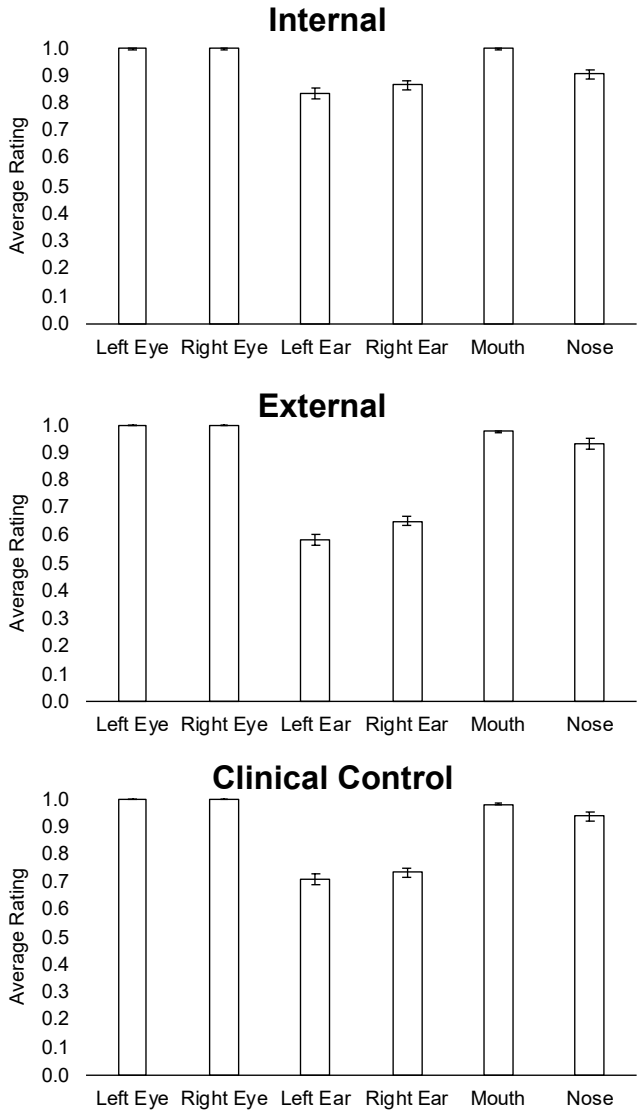

Figure S5. Average consensus ratings across image types, for each body part separately, in the internal (CBTN), external (CBTN), and clinical control (SLIP) groups. Error bars indicate +/- Standard Error of the Mean.

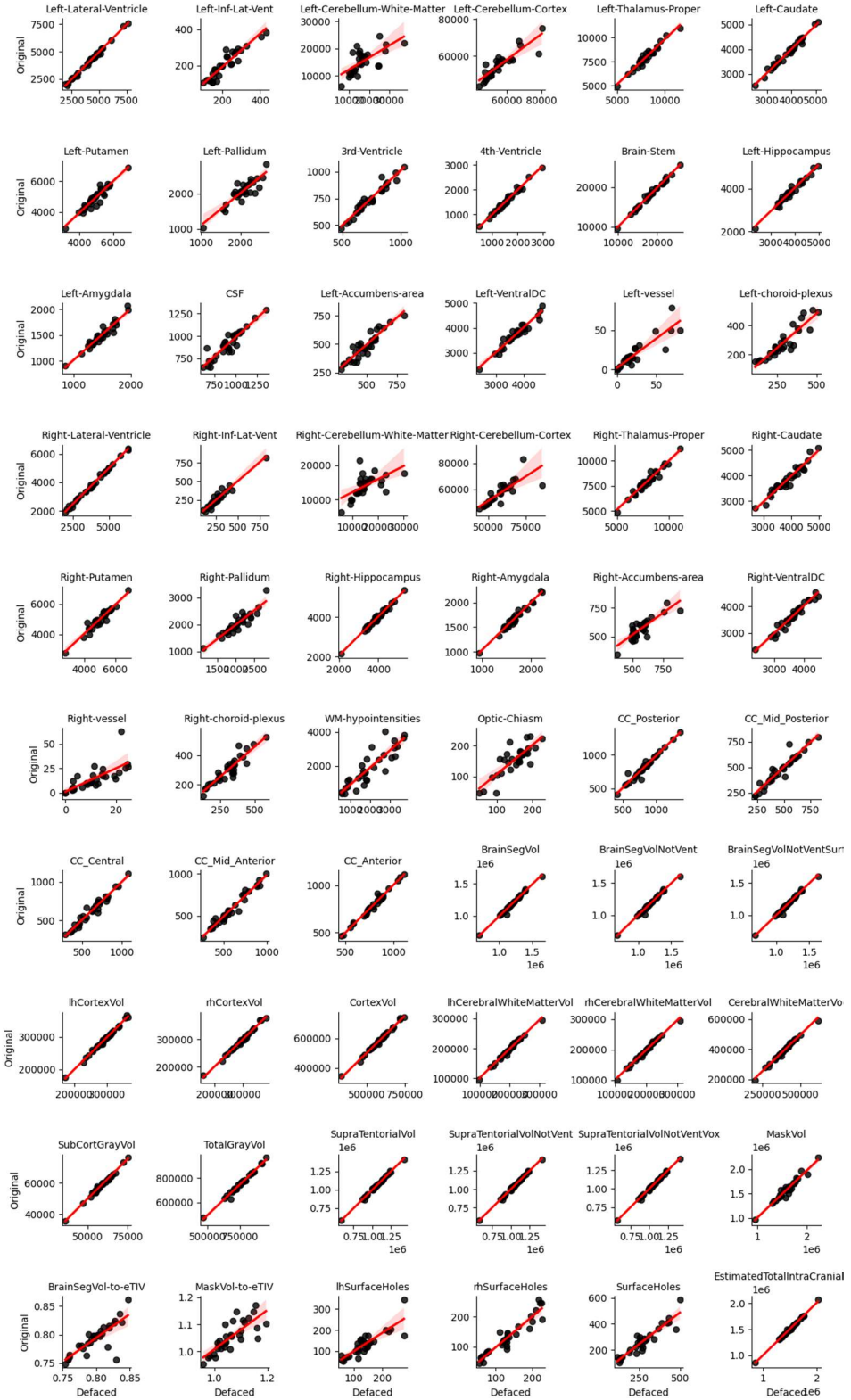

Figure S6. Scatter plots showing estimated volumes of cortical and subcortical regions derived from original (face intact; y-axis) and defaced (x-axis) images for each subject in the SLIP cohort (N=31; each dot represents one subject). Red line indicates linear regression fit to illustrate linear trend.

	Validation	**Internal Testing** *CBTN*	**External Testing** *CBTN*	**Clinical Control Testing** *SLIP*
Dice Similarity Mean / Median / SEM	0.78 / 0.80 / 0.01	**0.78 / 0.80 / 0.01**	**0.75 / 0.78 / 0.02**	**0.75 / 0.77 / 0.01**
T1w		0.78 / 0.80 / 0.01	0.76 / 0.78 / 0.02	0.73 / 0.78 / 0.02
T1w-CE		0.78 / 0.79 / 0.01	0.76 / 0.79 / 0.02	0.72 / 0.75 / 0.02
T2w		0.79 / 0.81 / 0.01	0.72 / 0.78 / 0.03	0.79 / 0.81 / 0.01
FLAIR		0.77 / 0.79 / 0.01	0.74 / 0.78 / 0.02	0.76 / 0.79 / 0.02
Sensitivity Mean / Median / SEM	0.80 / 0.83 / 0.01	**0.79 / 0.83 / 0.01**	**0.76 / 0.80 / 0.03**	**0.78 / 0.83 / 0.02**
T1w		0.79 / 0.82 / 0.02	0.76 / 0.79 / 0.03	0.77 / 0.82 / 0.03
T1w-CE		0.80 / 0.84 / 0.02	0.76 / 0.79 / 0.03	0.75 / 0.81 / 0.02
T2w		0.81 / 0.83 / 0.01	0.74 / 0.81 / 0.04	0.84 / 0.85 / 0.01
FLAIR		0.78 / 0.82 / 0.02	0.76 / 0.78 / 0.03	0.78 / 0.82 / 0.02
95% Hausdorff Distance Mean / Median / SEM	7.71 / 5.39 / 0.60	**7.10 / 5.46 / 0.59**	**8.45 / 6.26 / 1.78**	**10.09 / 7.38 / 1.02**
T1w		7.59 / 5.83 / 1.01	8.77 / 7.14 / 1.45	14.93 / 10.30 / 2.30
T1w-CE		7.88 / 5.39 / 1.08	8.99 / 7.07 / 1.24	15.33 / 10.30 / 1.78
T2w		6.09 / 5.10 / 0.55	8.95 / 4.24 / 3.44	5.65 / 4.79 / 0.47
FLAIR		6.83 / 5.39 / 1.01	7.09 / 5.00 / 1.56	4.43 / 3.39 / 0.60
Composite face score (eyes + mouth + nose)		**0.97**	**0.98**	**0.98**
Eyes		1	1	1
Mouth		1	0.97	0.98
Nose		0.90	0.93	0.94
Ears		0.85	0.62	0.72
Percentage of images with brain voxels impacted		**0%**	**0%**	**0%**

Table S2. Summary statistics of defacing model performance. (Upper) Standard metrics (dice, sensitivity, Hausdorff distance) of model performance for predicted (compared to ground truth) face masks across each testing cohort. (Lower) Rater-determined accuracy of model-generated defacing performance. Scores represent the average percent coverage of facial features (scale of 0-1). SEM = Standard error of the mean.

	Internal Testing *CBTN*	**External Testing** *CBTN*
Whole brain volume	$t(35) = -0.58$, $p = 0.566$	$t(24) = 1.45$, $p = 0.16$
Tumor segmentation volume	$t(35) = -1.67$, $p = 0.104$	$t(24) = -0.54$, $p = 0.595$

Table S3. Paired t-tests comparing AI-generated volumes between defaced and original images. Results show no significant difference in model-predicted whole brain or tumor volumes.

	Region	$r_S(29)$	p	$t(30)$	p
Global measures (volume)					
	Cortex	>0.99	**3.62E-34**	0.15	0.88
	Cerebral White Matter	>0.99	**2.89E-27**	0.93	0.359
	Brain Seg	>0.99	**5.68E-29**	1.41	0.17
	Brain Seg (NotVent)	>0.99	**1.43E-27**	1.39	0.176
	Brain Seg (NotVent) Surface	>0.99	**1.36E-28**	1.38	0.177
	Estimated Total Intracranial (eTIV)	0.99	**6.32E-26**	-1.65	0.11
	Total Gray Matter	>0.99	**6.81E-28**	0.64	0.526
	Subcortical Gray Matter	0.98	**2.35E-22**	-1.03	0.309
	CSF	0.86	**6.34E-10**	-0.27	0.788
Regional measures (volume)					
	Supratentorial (NotVent)	>0.99	**3.62E-34**	0.97	0.341
	LH Cortex	>0.99	**1.38E-32**	0.11	0.914
	Right Lateral Ventricle	>0.99	**6.33E-32**	0.31	0.759
	Supratentorial	>0.99	**8.86E-31**	1.01	0.323
	Brain Stem	>0.99	**2.82E-30**	0.95	0.351
	RH Cerebral White Matter	>0.99	**1.43E-27**	0.6	0.554
	Left Lateral Ventricle	>0.99	**2.89E-27**	1.56	0.13
	RH Cortex	>0.99	**1.08E-26**	0.17	0.866
	4th Ventricle	0.99	**6.32E-26**	0.72	0.476
	LH Cerebral White Matter	0.99	**7.99E-25**	1.22	0.234
	Left Caudate	0.98	**1.50E-23**	-1.15	0.258
	Right Ventral DC	0.98	**7.80E-22**	-0.86	0.395
	CC Posterior	0.98	**1.35E-20**	0.06	0.955
	CC Mid Anterior	0.97	**1.03E-19**	0.65	0.518
	CC Central	0.97	**4.58E-19**	-1.19	0.243
	Right Hippocampus	0.97	**4.58E-19**	-2.36	**0.025**
	Left Hippocampus	0.96	**2.57E-18**	-1.23	0.227
	CC Mid Posterior	0.96	**9.22E-18**	-0.8	0.43
	Right Thalamus Proper	0.96	**1.59E-17**	-0.92	0.366
	Left Amygdala	0.96	**2.78E-17**	-1.17	0.253
	Right Caudate	0.96	**3.18E-17**	0.62	0.54
	3rd Ventricle	0.96	**3.64E-17**	-0.93	0.362
	CC Anterior	0.95	**7.72E-17**	-1.07	0.291
	Left Ventral DC	0.95	**1.31E-16**	-1.26	0.219
	Right Amygdala	0.95	**1.48E-16**	-1.78	0.085
	Right Inferior Lateral Ventricle	0.95	**1.88E-16**	0.55	0.583
	Left Thalamus Proper	0.95	**3.78E-16**	1.08	0.29
	Left Choroid Plexus	0.93	**1.65E-14**	0.41	0.686
	Right Cerebellum Cortex	0.93	**2.08E-14**	0.54	0.595

	Region	$r_S(29)$	*p*	*t*(30)	*p*
	Right Putamen	0.93	**3.71E-14**	-0.28	0.779
	Left Cerebellum Cortex	0.92	**4.28E-13**	0.8	0.427
	Left Inferior Lateral Ventricle	0.92	**5.25E-13**	1.63	0.113
	Left Putamen	0.91	**1.00E-12**	0.41	0.688
	Right Choroid Plexus	0.91	**2.68E-12**	0.85	0.401
	Left Accumbens area	0.90	**2.84E-12**	0.17	0.863
	Right Pallidum	0.90	**3.86E-12**	0.19	0.85
	Left Vessel	0.90	**9.42E-12**	1.87	0.072
	Right Vessel	0.84	**2.27E-09**	-2.32	**0.027**
	Left Pallidum	0.78	**2.94E-07**	-0.45	0.657
	Optic Chiasm	0.76	**5.58E-07**	-0.84	0.405
	Right Accumbens area	0.72	**5.53E-06**	-0.33	0.745
	Left Cerebellum White Matter	0.71	**7.53E-06**	1.49	0.147
	Right Cerebellum White Matter	0.69	**1.81E-05**	1.27	0.212
Other Global Measures					
	Surface Holes	0.89	**1.51E-11**	0.91	0.371
	LH Surface Holes	0.84	**3.93E-09**	0.98	0.337
	RH Surface Holes	0.92	**1.43E-13**	0.36	0.725
	MaskVol-to-eTIV	0.84	**3.46E-09**	1.48	0.148
	BrainSegVol-to-eTIV	0.78	**2.44E-07**	1.93	0.063
	WM hypointensities	0.90	**4.77E-12**	1.02	0.314

Table S4. Comparison of Freesurfer brain measures between original (non-defaced) and defaced T1w images of the clinical control group. Group-level spearman correlations (r_S) and paired t-tests (t) indicate high agreement in estimated measures across regions (*N*=31). Statistically significant comparisons are indicated with bold text. Abbreviations - NotVent: excluding ventricles or CSF; CC: corpus callosum; DC: diencephalon.

Potential limitations in application to clinical populations

Please note that craniofacial, orbital, and ear pathologies can be obscured when using this algorithm. In addition, it is important to note that our training dataset exclusively included images acquired from brain/CNS tumor patients (Table S1). Our results therefore do not determine how the model would perform in clinical populations with structural malformations and anomalies particularly in craniofacial regions. Some examples are: cleft lip and palate, craniosynostosis, hemifacial microsomia, hemangiomas, temporal bone and ear pathologies. Because such samples are not included in the training dataset, it is possible that the model will not properly deface the brain images from these patients.

If it is desired to utilize this tool on populations with potential craniofacial deformities, the resulting files (face masks and/or defaced images) output by the model should be visually reviewed to ensure sufficient facial masking coverage for de-identification purposes. Manual refinement may be necessary.

Checklist for Artificial Intelligence in Medical Imaging (CLAIM; 2024)

Section / Topic	No.	Item	Page / Line	No	NA
TITLE / ABSTRACT					
	1	Identification as a study of AI methodology, specifying the category of technology used (e.g., deep learning)	Title		
ABSTRACT					
	2	Summary of study design, methods, results, and conclusions	Sections “Materials and Methods”, “Results”, “Conclusions”		
INTRODUCTION					
	3	Scientific and/or clinical background, including the intended use and role of the AI approach	Paragraphs 1-3		
	4	Study aims, objectives, and hypotheses	Paragraph 3		
METHODS					
Study Design	**5**	Prospective or retrospective study	Section “Patient cohorts” sentence 1		
	6	Study goal	Section “Patient cohorts”		
Data	**7**	Data sources	Section “Patient cohorts”		
	8	Inclusion and exclusion criteria	Section “Patient cohorts”		
	9	Data pre-processing	Section “Ground truth creation with semi-automated face mask segmentation”		
	10	Selection of data subsets	Section “AI deep learning model development”		
	11	De-identification methods			**X**
	12	How missing data were handled			**X**
	13	Image acquisition protocol			**X**
Reference Standard	**14**	Definition of method(s) used to obtain reference standard	Section “Ground truth creation with semi-automated face mask segmentation” Supp. Methods section “Training procedures for generation of face masks”		

	15	Rationale for choosing the reference standard	Section “Ground truth creation with semi-automated face mask segmentation” Supp. Methods section “Training procedures for generation of face masks”		
	16	Source of reference standard annotations	Section “Ground truth creation with semi-automated face mask segmentation” Supp. Methods section “Training procedures for generation of face masks”		
	17	Annotation of test set	Section “Ground truth creation with semi-automated face mask segmentation” Supp. Methods section “Training procedures for generation of face masks”		
	18	Measures of inter- and intra-rater variability of features described by the annotators	Section “Ground truth creation with semi-automated face mask segmentation” Supp. Methods section “Training procedures for generation of face masks”		
Data Partitions	**19**	How data were assigned to partitions	Section “AI deep learning model development”		
	20	Level at which partitions are disjoint	Section “AI deep learning model development”, sentences 3-4		
Testing Data	**21**	Intended sample size			**X**
Model	**22**	Detailed description of model	Section “AI deep learning model development”, sentence 2 Supp. Methods section “nnU-Net”		
	23	Software libraries, frameworks, and packages	Section “AI deep learning model development”, sentence 2 Supp. Methods section “nnU-Net”		

	24	Initialization of model parameters	Section “AI deep learning model development”, sentence 2 Supp. Methods section “nnU-Net”		
Training	**25**	Details of training approach	Section “AI deep learning model development” Supp. Methods section “nnU-Net”		
	26	Method of selecting the final model	Supp. Methods section “nnU-Net”		
	27	Ensembling techniques	Supp. Methods section “nnU-Net”		
Evaluation	**28**	Metrics of model performance	Section “Defacing accuracy”		
	29	Statistical measures of significance and uncertainty	Supp. Methods, section “Statistical comparisons”		
	30	Robustness or sensitivity analysis	Section “Defacing accuracy”		
	31	Methods for explainability or interpretability	Section “Defacing accuracy”		
	32	Evaluation on internal data	Section “AI deep learning model development” Table 1		
	33	Testing on external data	Section “Patient cohorts” second paragraph Section “AI deep learning model development” Table 1		
	34	Clinical trial registration			**X**
RESULTS					
Data	**35**	Numbers of patients or examinations included and excluded	Methods section “Patient cohorts” Fig. 1, Table 1		
	36	Demographic and clinical characteristics of cases in each partition	Tables 1 & S1		
Model performance	**37**	Performance metrics and measures of statistical uncertainty	Section “Defacing Accuracy”		
	38	Estimates of diagnostic performance and their precision	Section “Defacing Accuracy”		
	39	Failure analysis of incorrect results	Discussion, second paragraph, sentences 1-2		
DISCUSSION					

	40	Study limitations	Paragraphs 2 & 4 Supp. Results section "Potential limitations in application to clinical populations"		
	41	Implications for practice, including intended use and/or clinical role	Paragraphs 1 & 2 Conclusions		
OTHER INFORMATION					
	42	Provide a reference to the full study protocol or to additional technical details	Supp. Methods Fig S1 & S2		
	43	Statement about the availability of software, trained model, and/or data	Abstract Conclusions Methods section "Patient cohorts" Figure 1		
	44	Sources of funding and other support; role of funders	Acknowledgements		

Tejani AS, Klontzas ME, Gatti AA, et al. Checklist for Artificial Intelligence in Medical Imaging (CLAIM):2024 Update. Radiol Artif Intell 2024;6(4):e240300. https://doi.org/10.1148/ryai.240300