AJNR

Empowering Data Sharing in Neuroscience: A Deep Learning Deidentification Method for Pediatric Brain MRIs

Ariana M. Familiar, Neda Khalili, Nastaran Khalili, Cassidy Schuman, Evan Grove, Karthik Viswanathan, Jakob Seidlitz, Aaron Alexander-Bloch, Anna Zapaishchykova, Benjamin H. Kann, Arastoo Vossough, Phillip B. Storm, Adam C. Resnick, Anahita Fathi Kazerooni and Ali Nabavizadeh

AJNR Am J Neuroradiol 2025, 46 (5) 964-972
doi: https://doi.org/10.3174/ajnr.A8581
http://www.ajnr.org/content/46/5/964

Empowering Data Sharing in Neuroscience: A Deep Learning Deidentification Method for Pediatric Brain MRIs

Ariana M. Familiar, Neda Khalili, Nastaran Khalili, Cassidy Schuman, Evan Grove, Karthik Viswanathan, Jakob Seidlitz, Aaron Alexander-Bloch, Anna Zapaishchykova, Benjamin H. Kann, Arastoo Vossough, Phillip B. Storm, Adam C. Resnick, Anahita Fathi Kazerooni, and Ali Nabavizadeh

ABSTRACT

BACKGROUND AND PURPOSE: Privacy concerns, such as identifiable facial features within brain scans, have hindered the availability of pediatric neuroimaging data sets 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 National Institutes of Health data sharing mandates, novel solutions are a critical need.

MATERIALS AND METHODS: To develop an artificial intelligence (AI)-powered tool for automatic defacing of pediatric brain MRIs, deep learning methodologies (nnU-Net) were used by using a large, diverse multi-institutional data set of clinical radiology images. This included multiparametric MRIs (T1-weighted [T1W], T1W-contrast-enhanced, T2-weighted [T2W], T2W-FLAIR) with 976 total images from 208 patients with brain tumor (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 were 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 (nondefaced) 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 $P < .0001$). 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 $r_S > 0.9$, $P < .0001$).

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; CE = contrast-enhanced; CHOP = Children's Hospital of Philadelphia; CSA = cross-sectional area; LH = left hemisphere; NIH = National Institutes of Health; RH = right hemisphere; SEM = standard error of the mean; SLIP = Scans with Limited Imaging Pathology; T1W = T1-weighted; T2W = T2-weighted; TMT = temporalis muscle thickness

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

Received July 24, 2024; accepted after revision November 7.

From the Center for Data-Driven Discovery in Biomedicine (D^{3}b) (A.M.F., Neda K., Nastaran K., K.V., A.V., P.B.S., A.C.R., A.F.K., A.N.), Departments of Neurosurgery (A.M.F., Neda K., Nastaran K., K.V., P.B.S., A.C.R., A.F.K), and Child and Adolescent Psychiatry and Behavioral Science (J.S., A.A.-B.), and Division of Radiology (A.V.), The Children's Hospital of Philadelphia, Philadelphia, Pennsylvania; School of Engineering and Applied Science (C.S., E.G.), Departments of Psychiatry (J.S., A.A.-B.) and Radiology, Perelman School of Medicine (A.V., A.N.), and AI2D Center for AI and Data Science for Integrated Diagnostics (A.F.K.), University of Pennsylvania, Philadelphia, Pennsylvania; Lifespan Brain Institute at the Children's Hospital of Philadelphia and University of Pennsylvania (A.A.-B.), Philadelphia, Pennsylvania; Artificial Intelligence in Medicine (AIM) Program (A.Z., B.H.K.), Mass General Brigham, Harvard Medical School, Boston, Massachusetts; and Department of Radiation Oncology (A.Z., B.H.K.), Dana-Farber Cancer Institute and Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts.

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

Please address correspondence to Ali Nabavizadeh, MD, Associate Professor of Radiology, Department of Radiology, 1 Silverstein Building, Hospital of the University of Pennsylvania, 3400 Spruce St, Philadelphia, PA 19104; e-mail: Ali.Nabavizadeh@pennmedicine.Upenn.edu; @Ali_Nabavizadeh; @PennMedicine; @FamiliarAriana; @ChildrensPhila

Indicates article with supplemental data.

Indicates article that contains code.

http://dx.doi.org/10.3174/ajnr.A8581

SUMMARY

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 data sets; however, pediatric data sets have lagged behind their adult counterparts, and neuroimaging data are particularly challenging to share due to privacy concerns, because 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 multiparametric 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 (eg, image preprocessing or measured group-level trends).

KNOWLEDGE ADVANCEMENT: To facilitate broad sharing of pediatric neuroimaging data sets, a robust, automatic deidentification tool is provided to ease the burden on research teams to prepare and release imaging data while protecting patient privacy.

disciplines, data sharing plans have recently become a mandate for National Institutes of Health (NIH)-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 artificial intelligence (AI) in health care.[1]

In alignment with FAIR[2] principles, several imaging data repositories have been established such as the Alzheimer Disease Neuroimaging Initiative[3] and the National Cancer Institute's The Cancer Imaging Archive[4] and Imaging Data Commons, which provide effective data discovery and accessibility. While several large-scale, multi-institutional imaging data sets exist, such as the National Lung Screening Trial (NLST) for lung cancer (chest CTs from more than 26,000 patients)[5] and the Breast Cancer Screening Digital Breast Tomosynthesis (breast mammograms from 5060 patients),[6] comparable radiology data sets in neuroscience fields have lagged behind 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 (eg, 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, 1 study found that FSL's defacing removed brain tissue in most children (ages 8–11) and in some young adult (ages 19–31) cases, and had worse performance for eyes and mouth removal compared with 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 brainstem 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 (eg, 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 data sets. 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 multiparametric MRIs, we use a large, multi-institutional clinical radiology data set (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 4 image sequences included per subject (T1W, T1W contrast-enhanced [T1W-CE], T2W, and T2W-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 were collected from the CBTN,[19] a large-scale, multi-institutional repository of longitudinal clinical, imaging, genomic, and other paired data.[21] Two hundred eight subjects were selected based on imaging availability and inclusion of a range of ages at the time of imaging (median age 8; minimum = 0.35, maximum = 21.71 years) and cancer histologies (Fig 1, Table, Supplemental Data). MRI scans were unprocessed images from treatment-naïve clinical examinations (T1W, T1W-CE, T2W, and T2W-FLAIR). All subjects had histologically confirmed pediatric brain tumors.

To test generalizability to nonbrain tumor patients (clinical control group), a cohort of 40 subjects with available images from

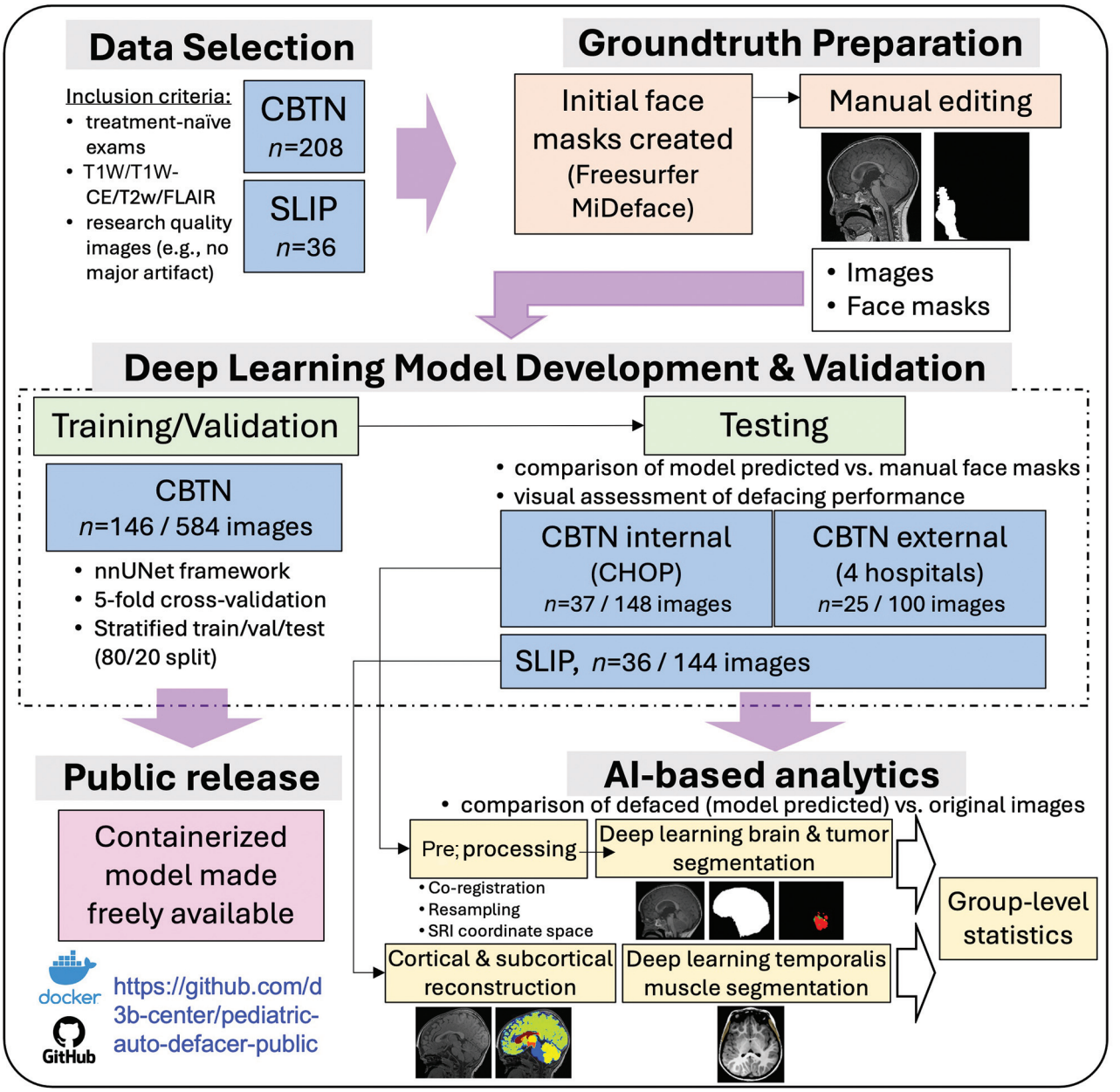

FIG 1. Diagram of overall study workflow. Data cohorts included brain tumor (CBTN) and nonbrain tumor control (SLIP). Initial ground truth face masks were created with MiDeface and manually edited. A 3D deep learning model was trained with the nnUNet framework, by using a single image as input, and tested 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.

the SLIP[20] data set were selected to match the general distributions of age and sex of the CBTN cohort. Thirty-six subjects had sufficient images and were included in the main analyses.

Ground Truth Creation with Semiautomated Face Mask Segmentation

Preliminary face masks were generated for each image by using the MiDeface[22] algorithm and then were manually edited. Of the 976 images, 507 (52%) were found to be inaccurately defaced and were manually revised by using the ITK-SNAP[23] software (by authors C.S., E.G.; Supplemental Data). 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.

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). nnUNet[25] v1 (https://github.com/MIC-DKFZ/nnUNet/tree/nnunetv1; 3D full resolution; Supplemental Data) was used with 5-fold cross-validation, initial learning rate 0.01, stochastic gradient descent with Nesterov momentum ($\mu = 0.99$), and number of epochs = 1000 × 250 minibatches.

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 1 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/Diffuse intrinsic pontine glioma	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 Healthcare	10 (7%)	4 (11%)	9 (36%)	
Philips	1 (1%)			
Toshiba	1 (1%)			

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 (ie, 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 Children's Hospital of Philadelphia (CHOP), we additionally split the testing cohort into "internal" (CHOP) and "external" (4 separate institutions) testing data sets.

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, 2 raters (authors Neda K. and Nastaran 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 by using standard image processing steps, in both the brain tumor (CBTN) and nonbrain tumor (SLIP) groups separately.

Preprocessing and Application of Pretrained AI Models. For each subject in the CBTN testing cohorts, T1W, T2W, and FLAIR sequence images were coregistered with their corresponding T1W-CE sequence and resampled to an isotropic resolution of 1 mm^3 based on the anatomic SRI24 atlas[26] by 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 coregistration was confirmed by visual assessment of the 4 images.

Preprocessed data for each subject were then input into existing pretrained AI models for automatic brain tissue extraction

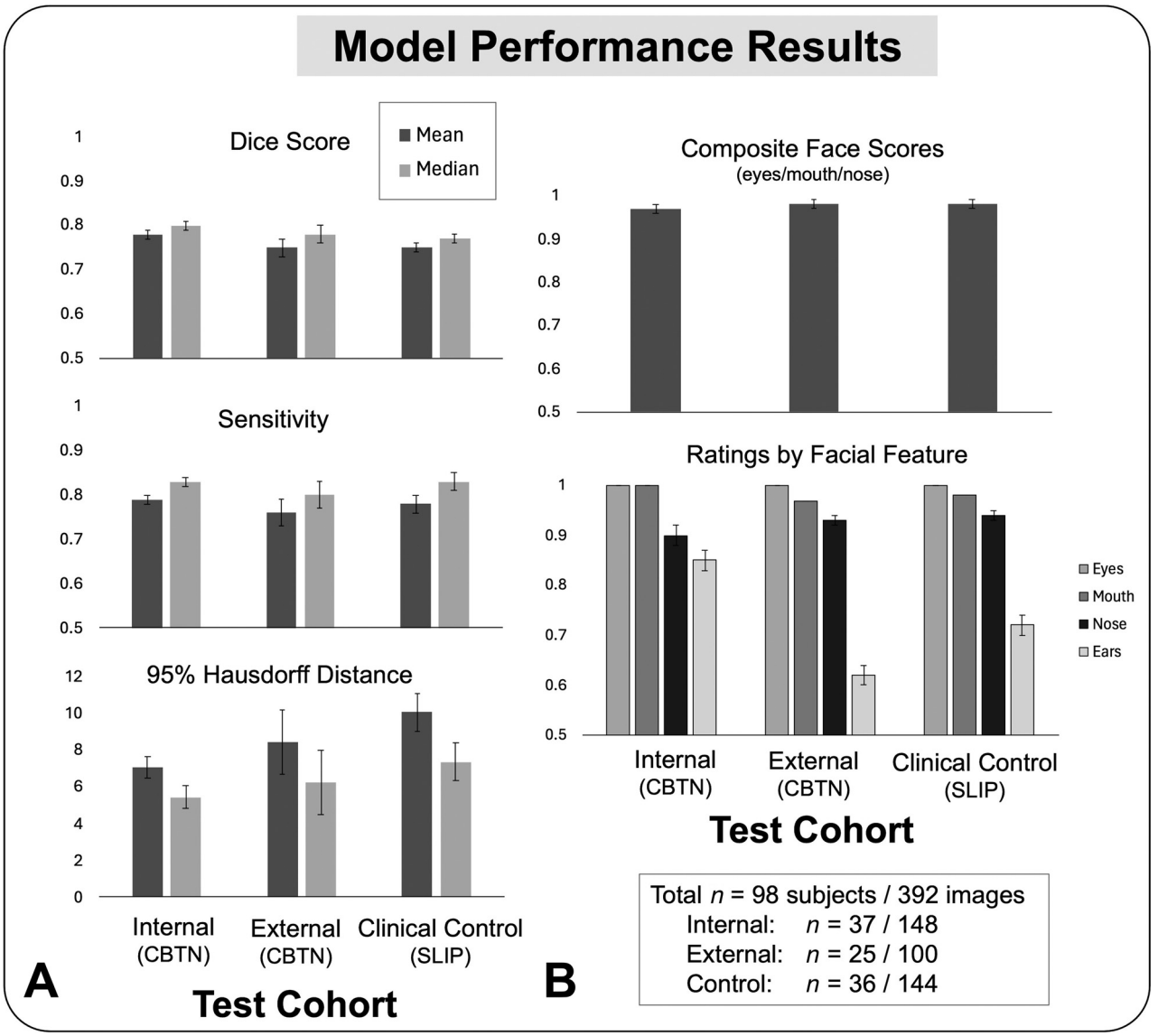

FIG 2. Model performance results. Plots show aggregate metrics across image types for each testing cohort (see Supplemental Data for results for image type separately); error bars represent SEM. *A*, Standard metrics for segmentation evaluation including Dice similarity, sensitivity, and 95% Hausdorff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

and tumor subregion segmentation (https://github.com/d3b-center/peds-brain-seg-pipeline-public).[29,30] This was performed once by using the original images (nondefaced), and once by 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 (temporalis muscle 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 Data for a description of all statistical comparisons and a CLAIM checklist to indicate alignment with the proposed methodologic guidelines recommended for AI in medical imaging.[32–34]

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 =

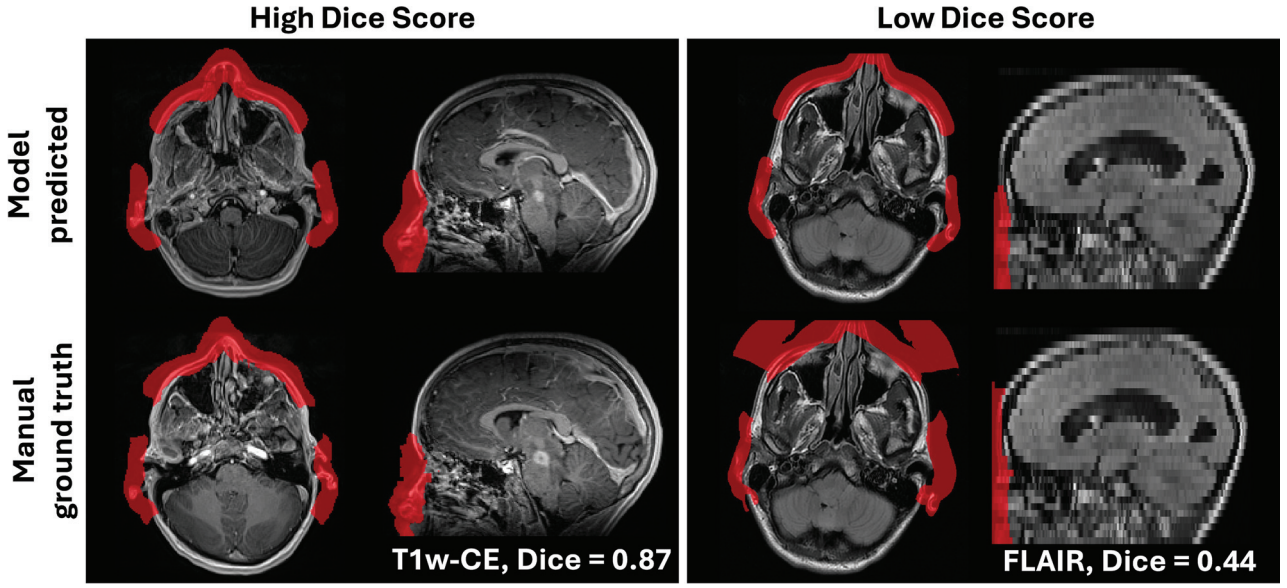

FIG 3. Representative example images of model predicted versus manual ground truth segmentation masks. Subjects shown with high (*left box*; T1W-CE sequence) and low (*right box*; FLAIR sequence) Dice similarity scores between the model predicted (*upper row*) and manual ground truth (*lower row*) face 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).

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 ($F(3,108) = 0.38$, $P = .77$) and external ($F(3,72) = 1.8$, $P = .16$) cohorts, however there was a significant effect in the clinical control group ($F(3,105) = 6.14$, $P = .007$) with better model performance for T2W and FLAIR compared with T1W and T1W-CE (Supplemental Data). Pearson correlations showed no effect of age on Dice scores averaged across image types (internal: $r(35) = 0.19$, $P = .25$; external: $r(23) = 0.29$, $P = .17$; control: $r(34) = 0.28$, $P = .095$; Supplemental Data). One-way ANOVAs indicated no effect of sex (internal: $F(1,35) = 2.0$, $P = .17$; external: $F(1,23) = 0.28$, $P = .6$; control: $F(1,34) = 3.17$, $P = .08$) or race (internal: $F(3,33) = 0.18$, $P = .911$; external: $F(2,22) = 0.61$, $P = .551$; control: $F(2,32) = 1.07$, $P = .356$) on Dice scores, and no effect of histopathologic diagnosis (internal: $F(4, 32) = 0.442$, $P = .777$; external: $F(1, 23) = 0.377$, $P = .545$) or general tumor location (internal: $F(4,32) = 0.837$, $P = .512$; external: $F(3,21) = 0.1$, $P = .959$) in the CBTN testing cohorts.

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 (Fig 3, Supplemental Data). To more accurately assess model performance, 2 raters (Neda K., Nastaran 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, Supplemental Data; 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 ($F(3,75) = 10.8$, $P < .0001$), with higher average ratings for T1W ($M = 0.91$) and T1W-CE ($M = 0.91$) compared with T2W ($M = 0.89$) and FLAIR ($M = 0.86$); but no effect of image type in the internal ($F(3,108) = 1.17$, $P = .33$) or external ($F(3,72) = 0.32$, $P = .81$) groups. Average rating across subjects and image types for each feature is displayed in the Supplemental Data.

Assessing Impact of Defacing on Downstream Analytics

Preprocessing and Application of Pretrained AI Models. Defaced and original (nondefaced) images underwent preprocessing and were input to pretrained AI tools to assess any impact of defacing on standard downstream analysis by using all 4 image sequences (T1W/T1W-CE/T2W/FLAIR). Visual inspection showed equivalent coregistration performance between defaced and original images. For the pediatric brain tumor test data sets, the volumes of AI-generated brain masks were equivalent between defaced and nondefaced images (internal: $r_S(35) > 0.99$, $P < .0001$; external: $r_S(23) > 0.99$, $P < .0001$; Fig 4, upper and middle). AI-generated tumor segmentations were also unaffected by defacing, indicated by equivalent volumes of contrast-enhancing tumor, nonenhancing tumor, cystic, and edema subregions (internal: all subregions $r_S(35) > 0.99$, $P < .0001$; external: all subregions $r_S(23) > 0.99$, $P < .0001$; Fig 4, Supplemental Data).

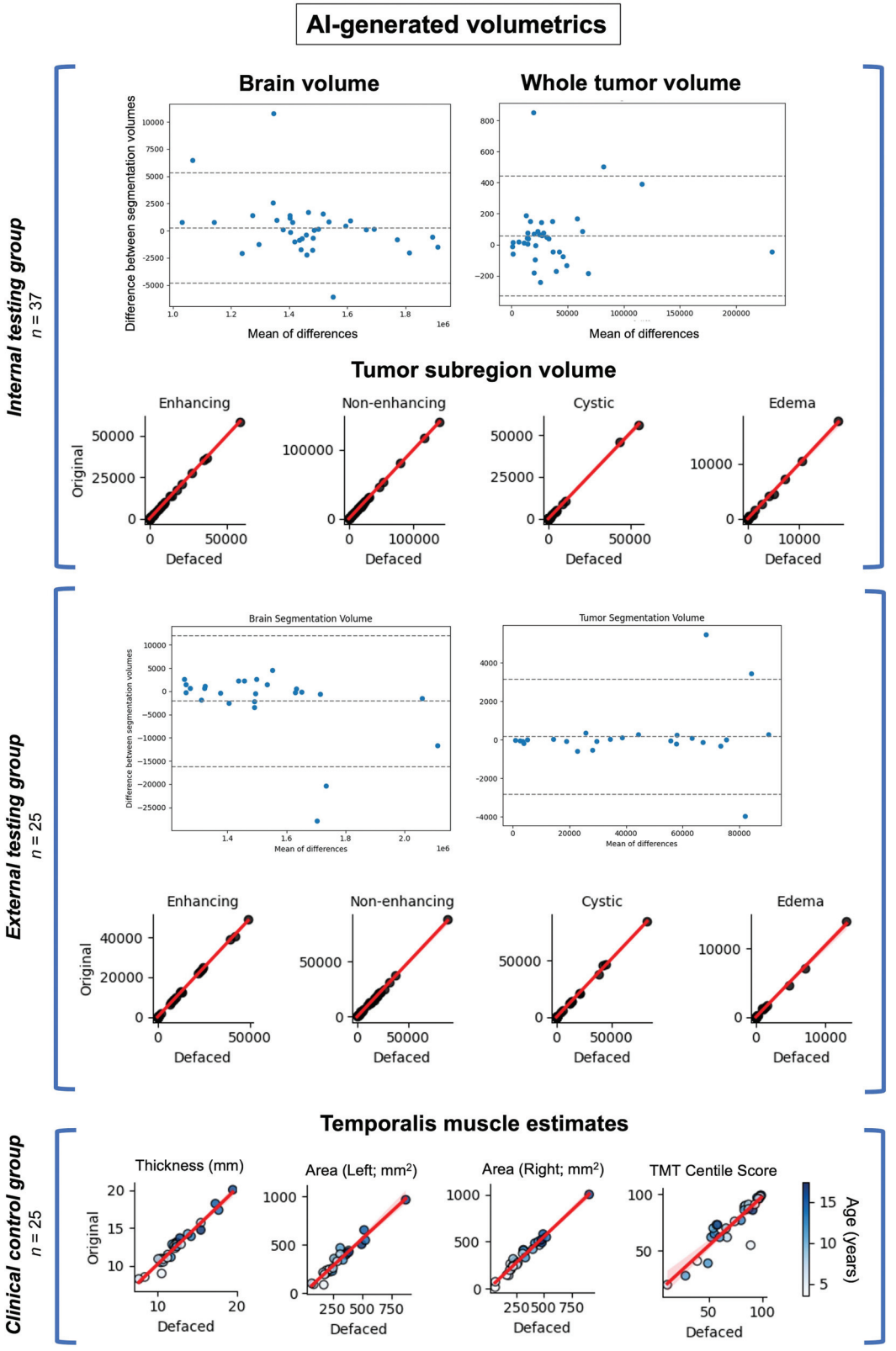

FIG 4. Testing the impact of defacing on AI-generated volumetrics. Each point represents 1 subject; the red line indicates a linear trend. *Upper/middle*: Comparison of tumor subregion volumes between defaced (x-axis) and original (y-axis) images in pediatric brain tumor subjects. There was very high agreement between brain and tumor segmentation volumes. *Lower*: Comparison of estimated 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 TMT, CSA, 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 by using a standard anatomic 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 $P < .0001$; Supplemental Data). 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: $r_S(29) = 0.71$, $P < .0001$; right: $r_S(29) = 0.69$, $P < .0001$). Nine global measurements (cortex, cerebral white matter, subcortical gray matter, total gray matter, total brain [including cerebellum], total brain excluding ventricles [surface], total brain excluding ventricles [volume], 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 (Supplemental Data), with the exception of the right vessel (original $= 11.3$, $SEM = 1.38$; defaced $M = 14.7$, $SEM = 2.19$; $t(30) = -2.32$, $P = .03$) and the right hippocampus (original $M = 3940.8$, $SEM = 101$; defaced $M = 3972.8$, $SEM = 101$; $t(30) = -2.36$, $P = .03$), which were estimated to be slightly larger on average in the defaced compared with original images. Overall, these results indicate defacing had minimal impact on cortical and subcortical volumetric assessments by 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 TMT (mm) and CSA measurements (SLIP cohort ages >3 years; n=28) by using an existing AI-powered pipeline[24] with T1W images. Notably, TMT scores have been implicated as a predictive marker for sarcopenia across patient populations.[35–38] Spearman correlations showed high agreement of estimated TMT ($r_S26) = 0.96$, all $P < .0001$) and CSA (left hemisphere [LH]: $r_S26) = 0.96$, $P < .0001$; right hemisphere [RH]: $r_S26) = 0.97$, $P < .0001$; Fig 4, lower) between defaced and original images. Paired t tests indicated no difference in TMT volumes between original and defaced images ($t(27) = -1.8$, $P = .08$), but a significant difference in CSA (LH: $t(27) = -3.74$, $P < .0001$; RH: $t(27) = -4.79$, $P = .0009$) with lower surface area estimates for the defaced (LH: $M = 306.2$, $SEM = 30$; RH: $M = 314.7$, SEM = 33) compared with 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 data sets[24]) were not significantly affected by defacing ($r_S(26) = 0.9$, $P < .0001$; $t(27) = -0.97$, $P = .34$).

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 data sets 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 by using a large, multi-institutional data set of clinically acquired, multiparametric MRIs (CBTN).

The trained model had strong performance removing the face (eyes, nose, mouth) in an unseen data set, 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 data set (limited field of view). Notably, although the model was trained on data from patients with brain tumor, it could generalize to a separate data set 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 the Supplemental Data).

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 (nondefaced) 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 4 sequences, and further development could expand to additional types such as functional MRI and other advanced imaging (eg, 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 data set included images across 6 institutions, a large portion of the data set 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

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.

Disclosure forms provided by the authors are available with the full text and PDF of this article at www.ajnr.org.

REFERENCES

1. Chen RJ, Wang JJ, Williamson DFK, et al. **Algorithmic fairness in artificial intelligence for medicine and healthcare.** *Nat Biomed Eng* 2023;7:719–42 CrossRef Medline
2. Wilkinson MD, Dumontier M, Aalbersberg IJ, et al. **The FAIR Guiding Principles for scientific data management and stewardship.** *Sci Data* 2016;3:160018 CrossRef Medline
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;1:55–66 CrossRef Medline
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; 2013:1282–85 CrossRef Medline
5. Aberle DR, Adams AM, Berg CD, et al; National Lung Screening Trial Research Team. **Reduced lung-cancer mortality with low-dose computed tomographic screening.** *N Engl J Med* 2011;365:395–409 CrossRef Medline
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:e2119100 CrossRef Medline
7. Schwarz CG, Kremers WK, Therneau TM, et al. **Identification of anonymous MRI research participants with face-recognition software.** *N Engl J Med* 2019;381:1684–86 CrossRef Medline
8. Mazura JC, Juluru K, Chen JJ, 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* 2012;25:347–51 CrossRef Medline
9. Abramian D, Eklund A. **Refacing: Reconstructing Anonymized Facial Features Using GANS.** In: *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*. IEEE; 2019:1104–08 CrossRef
10. Bischoff-Grethe A, Ozyurt IB, Busa E, et al. **A technique for the deidentification of structural brain MR images**. *Hum Brain Mapp* 2007;28:892–903 CrossRef Medline
11. Gulban OF, Nielson D, Poldrack R, et al. **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;166:400–24 CrossRef Medline
13. Khazane A, Hoachuck J, Gorgolewski KJ, et al. **DeepDefacer: automatic removal of facial features via U-Net image segmentation.** *arXiv.org* 2022. http://arxiv.org/abs/2205.15536. Accessed January 26, 2024.
14. Milchenko M, Marcus D. **Obscuring surface anatomy in volumetric imaging data.** *Neuroinform* 2013;11:65–75 CrossRef Medline
15. de Sitter A, Visser M, Brouwer I, et al; MAGNIMS Study Group and Alzheimer's Disease Neuroimaging Initiative. **Facing privacy in neuroimaging: removing facial features degrades performance of image analysis methods**. *Eur Radiol* 2020;30:1062–74 CrossRef Medline
16. Rubbert C, Wolf L, Turowski B, et al; Alzheimer's Disease Neuroimaging Initiative. **Impact of defacing on automated brain atrophy estimation**. *Insights Imaging* 2022;13:54 CrossRef Medline
17. Theyers AE, Zamyadi M, O'Reilly M, et al. **Multisite comparison of MRI defacing software across multiple cohorts**. *Front Psychiatry* 2021;12:617997 CrossRef Medline
18. Buimer EEL, Schnack HG, Caspi Y, et al; Alzheimer's Disease Neuroimaging Initiative. **De-identification procedures for magnetic resonance images and the impact on structural brain measures at different ages**. *Hum Brain Mapp* 2021;42:3643–55 CrossRef Medline
19. Familiar AM, Kazerooni AF, Anderson H, et al. **A multi-institutional pediatric data set of clinical radiology MRIs by the Children's Brain Tumor Network**. *arXiv.org* 2023. https://arxiv.org/abs/10.48550/arXiv.2310.01413. October 15, 2024
20. Schabdach JM, Schmitt JE, Sotardi S, et al. **Brain growth charts of "clinical controls" for quantitative analysis of clinically acquired brain MRI**. *Radiology* 2023;309:e230096
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 CrossRef Medline
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;17:83–102 CrossRef Medline
24. Zapaishchykova A, Liu KX, Saraf A, et al. **Automated temporalis muscle quantification and growth charts for children through adulthood.** *Nat Commun* 2023;14:6863 CrossRef Medline
25. Isensee F, Jaeger PF, Kohl SAA, et al. **nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation.** *Nat Methods* 2021;18:203–11 CrossRef Medline
26. Rohlfing T, Zahr NM, Sullivan EV, et al. **The SRI24 multichannel atlas of normal adult human brain structure.** *Hum Brain Mapp* 2010;31:798–819 CrossRef Medline
27. Yushkevich PA, Pluta J, Wang H, et al. **Fast automatic segmentation of hippocampal subfields and medial temporal lobe subregions in 3 Tesla and 7 Tesla T2-weighted MRI**. *Alzheimers Dement* 2016;12:P126–27
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-Verlag International Publishing; 2020:380–94 CrossRef Medline
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.** *Neurooncol Adv* 2023;5:vdad027 CrossRef Medline
30. Vossough A, Khalili N, Familiar AM, et al. **Training and comparison of nnU-Net and DeepMedic methods for autosegmentation of pediatric brain tumors.** *AJNR Am J Neuroradiol* 2024;45:1081–89 CrossRef Medline
31. Fischl B. **FreeSurfer.** *Neuroimage* 2012;62:774–81 CrossRef Medline
32. Tejani AS, Klontzas ME, Gatti AA, et al; CLAIM 2024 Update Panel. **Checklist for Artificial Intelligence in Medical Imaging (CLAIM): 2024 update.** *Radiol Artif Intell* 2024;6:e240300 CrossRef Medline
33. Mongan J, Moy L, Charles E Kahn J. **Checklist for Artificial Intelligence in Medical Imaging (CLAIM): a Guide for Authors and Reviewers.** *Radiol Artif Intell* 2020;2:e200029 CrossRef Medline
34. Pham N, Hill V, Rauschecker A, et al. **Critical appraisal of artificial intelligence–enabled imaging tools using the levels of evidence system**. *AJNR Am J Neuroradiol* 2023;44:E21–28 CrossRef Medline
35. Lee B, Bae YJ, Jeong WJ, et al. **Temporalis muscle thickness as an indicator of sarcopenia predicts progression-free survival in head and neck squamous cell carcinoma.** *Sci Rep* 2021;11:19717 CrossRef Medline
36. Cho J, Park M, Moon WJ, et al. **Sarcopenia in patients with dementia: correlation of temporalis muscle thickness with appendicular muscle mass.** *Neurol Sci* 2022;43:3089–95 CrossRef Medline
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;31:4079–86 CrossRef Medline
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 CrossRef Medline
39. Schwarz CG, Kremers WK, Wiste HJ, et al; Alzheimer's Disease Neuroimaging Initiative. **Changing the face of neuroimaging research: comparing a new MRI de-facing technique with popular alternatives**. *NeuroImage* 2021;231:117845 CrossRef Medline