



ORIGINAL ARTICLE

Open Access



# Impact of defacing on automated brain atrophy estimation

Christian Rubbert<sup>1\*</sup> , Luisa Wolf<sup>1</sup>, Bernd Turowski<sup>1</sup>, Dennis M. Hedderich<sup>2</sup>, Christian Gaser<sup>3</sup>, Robert Dahnke<sup>3,4,5</sup> , Julian Caspers<sup>1</sup> and for the Alzheimer's Disease Neuroimaging Initiative

## Abstract

**Background:** Defacing has become mandatory for anonymization of brain MRI scans; however, concerns regarding data integrity were raised. Thus, we systematically evaluated the effect of different defacing procedures on automated brain atrophy estimation.

**Methods:** In total, 268 Alzheimer's disease patients were included from ADNI, which included unaccelerated ( $n = 154$ ), within-session unaccelerated repeat ( $n = 67$ ) and accelerated 3D T1 imaging ( $n = 114$ ).

Atrophy maps were computed using the open-source software *veganbagel* for every original, unmodified scan and after defacing using *afni\_refacer*, *fsl\_deface*, *mri\_deface*, *mri\_reface*, *PyDeface* or *spm\_deface*, and the root-mean-square error (RMSE) between z-scores was calculated.

RMSE values derived from unaccelerated and unaccelerated repeat imaging served as a benchmark. Outliers were defined as  $RMSE > 75$ th percentile and by using Grubbs's test.

**Results:** Benchmark RMSE was  $0.28 \pm 0.1$  (range 0.12–0.58, 75th percentile 0.33).

Outliers were found for unaccelerated and accelerated T1 imaging using the 75th percentile cutoff: *afni\_refacer* (unaccelerated: 18, accelerated: 16), *fsl\_deface* (unaccelerated: 4, accelerated: 18), *mri\_deface* (unaccelerated: 0, accelerated: 15), *mri\_reface* (unaccelerated: 0, accelerated: 2) and *spm\_deface* (unaccelerated: 0, accelerated: 7). *PyDeface* performed best with no outliers (unaccelerated mean RMSE  $0.08 \pm 0.05$ , accelerated mean RMSE  $0.07 \pm 0.05$ ).

The following outliers were found according to Grubbs's test: *afni\_refacer* (unaccelerated: 16, accelerated: 13), *fsl\_deface* (unaccelerated: 10, accelerated: 21), *mri\_deface* (unaccelerated: 7, accelerated: 20), *mri\_reface* (unaccelerated: 7, accelerated: 6), *PyDeface* (unaccelerated: 5, accelerated: 8) and *spm\_deface* (unaccelerated: 10, accelerated: 12).

**Conclusion:** Most defacing approaches have an impact on atrophy estimation, especially in accelerated 3D T1 imaging. Only *PyDeface* showed good results with negligible impact on atrophy estimation.

**Keywords:** Magnetic resonance imaging, Brain, Atrophy, De-identification, Privacy

## Key points

- **Background:** defacing MRI examinations of the brain is important to preserve privacy.
- Defacing procedures may interfere with software-based brain atrophy estimation (e.g., *veganbagel*).
- Most defacing procedures lead to systematic bias concerning atrophy estimation.

\*Correspondence: christian.rubbert@med.uni-duesseldorf.de

<sup>1</sup> University Dusseldorf, Medical Faculty, Department of Diagnostic and Interventional Radiology, D-40225 Dusseldorf, Germany  
Full list of author information is available at the end of the article  
Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: [http://adni.loni.usc.edu/wp-content/uploads/how\\_to\\_apply/ADNI\\_Acknowledgement\\_List.pdf](http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf).

- Substantial z-score deviations were found in 0–17.9% compared to a benchmark.
- Only defacing with PyDeface had a negligible impact on atrophy estimation.

**Background**

Magnetic resonance imaging (MRI) studies of the brain usually include facial features of a patient. For example, brain imaging in patients suffering from a neurocognitive decline or neurodegenerative diseases usually include a 3D T1-weighted anatomical dataset, which commonly depicts the ears and face. The threat of identifying patients or research subjects by applying face recognition techniques to such MRIs has been increasingly recognized in recent years. While only 40% of human volunteers were able to match face reconstructions based on volumetric renderings of 3D MRI data to the respective participants photographs with a success rate of “greater than chance” in a study from 2009 [1], studies employing machine learning approaches were successful in 83% of the cases in 2019 [2] and in 97% of the cases in 2021 [3].

Therefore, it is highly desirable—and frankly necessary—to remove identifying features, such as the face or ears [1–4], from MRI examinations depicting these in detail prior to a public data release or even before submitting MRI scans to an off-site service, e.g., to a cloud-based service for brain atrophy estimation. Several defacing approaches have been developed, some of which were applied to large-scale cohort studies, such as the Human Connectome Project (HCP [5]) and the Nathan Kline Institute—Rockland Sample (Rockland [6]). The most popular defacing approaches include afni\_refacer (based on AFNI [7]), mask\_face (released by the Neuroinformatics Research Group [8]), mri\_deface (based on FreeSurfer [9]), fsl\_deface (based on FMRIB’s Software Library (FSL) [10]), PyDeface (released by the Poldrack Lab [11]) and spm\_deface (included in the Statistical Parametric Mapping (SPM) package [12]). However, concerns have been raised with respect to data alteration after defacing, with some studies reporting significant

deviations in brain volume assessments [13, 14], while other studies have shown almost no effects of defacing [3, 15].

Brain atrophy is a feature of many neurodegenerative diseases, and characteristic brain volume changes may be decisive for diagnosis, for example in Alzheimer’s disease (AD [16]) or frontotemporal dementia [17], among others [18–20]. Brain volume changes are furthermore increasingly used in treatment monitoring, for example in the early stages of multiple sclerosis. [21] Several software packages for evaluation of regional brain volume alterations have been made available in the recent years. Software-augmented reading has been shown to help detect subtle volume losses in the early course of a disease and to decrease high inter-reader variation in reporting of regional brain atrophy [22, 23].

While technical accuracy is a hallmark of volumetric brain atrophy estimation, the impact of defacing procedures on the result has not been systematically studied. The current study evaluates the impact of commonly used defacing procedures on brain atrophy z-score maps in a large sample from the ADNI cohort using veganbagger, an open-source software for automatic brain atrophy estimation built around CAT12 for SPM12.

**Methods**

AD patients from the Alzheimer’s Disease Neuroimaging Initiative (ADNI [24]) database were retrospectively included in the analysis. The ADNI was launched in 2003 as a public–private partnership, led by Principal Investigator Michael W. Weiner, MD (<http://www.adni-info.org/>). AD patients were included, when (1) there was a 3D T1-weighted MRI series acquired at the Screening visit with a slice thickness of ≤ 1.5 mm, and (2) patients were aged younger than 75 years at the time of the MRI. Any MRI acquisition failing ADNI’s quality control (QC) were excluded. Unaccelerated repeats passing QC were excluded, when the initial imaging failed QC. A total of 268 AD patients were included (Table 1). The study was approved by the local ethics committee. Only publicly

**Table 1** Demographics of all included patients from the Alzheimer’s Disease Neuroimaging Initiative (ADNI) as well as the analyzed subgroups

	n	Females	Age	Study phase (ADNI 1/2/3)	Scanners <sup>a</sup>	GE/Philips/Siemens (1.5 T/3 T)
All AD patients	268	141 (52.6%)	67.7 ± 5.3 (55–74)	29%/56%/15%	95	38%/19%/43% (29%/71%)
Unaccelerated imaging	154	82 (53.2%)	67.8 ± 5.1 (55–74)	51%/49%/0%	76	45%/16%/38% (51%/49%)
Unaccelerated repeat imaging	67	38 (56.7%)	68.1 ± 5.0 (56–74)	100%/0%/0%	38	58%/6%/36% (100%/0%)
Accelerated imaging	114	59 (51.8%)	67.6 ± 5.5 (55–74)	0%/65%/35%	56	28%/24%/48% (0%/100%)

GE, general electric (Boston, MA, USA); Philips, Koninklijke Philips (Amsterdam, the Netherlands); Siemens, Siemens Healthineers (Erlangen, Germany); T, Tesla

<sup>a</sup> The number of scanners is estimated from the scanner device serial number included in the DICOM headers

available data were used. Statistical analysis was carried out using R [25].

In the ADNI, 3D T1 images were acquired (1) as unaccelerated imaging, (2) as unaccelerated repeat imaging (repetition of an unaccelerated 3D T1 within the same session) or (3) as accelerated imaging (i.e., parallel imaging using GRAPPA [26], SENSE [27] or the closely related ASSET). All patients with unaccelerated repeat imaging were also part of the group with unaccelerated imaging. Studies with different approaches to atrophy estimation have shown varying degrees of differences in atrophy assessments when using accelerated vs. unaccelerated imaging [28–31]. Since potentially different effects of defacing on accelerated vs. unaccelerated imaging have so far not been systematically studied, we show results for unaccelerated and accelerated 3D T1 imaging separately.

The previously published open-source software for volumetric estimation of gross atrophy and brain age longitudinally (veganbagel), an automatic workflow for generation of atrophy maps relative to age- and sex-specific normal templates [32], was adapted for the analysis. The latest available Docker-based version of veganbagel was used, containing the standalone version of CAT12.7 without the need for a separate MATLAB-license (<https://github.com/BrainImAccs/veganbagel>, commit 6a2ac5f). veganbagel was chosen for the analysis, since it is based on the established CAT12 for SPM12 software package and, to our knowledge, is the only open-source software readily allowing for single time point atrophy estimations for an individual brain scan.

In the workflow, standardized preprocessing of structural T1-weighted imaging is performed, comprising gray matter segmentation, normalization, modulation and spatial smoothing using CAT12 for SPM12. After preprocessing of healthy reference subjects from a normal cohort (Rockland [6], “Baseline1” visits,  $n = 949$  (65% female), mean age =  $46.3 \pm 17.1$  years (range 18–77)), mean and standard deviation (SD) templates are generated for each sex and age (containing the actual age  $\pm 2$  years).

Z-score maps (=“atrophy maps”) were then calculated for all AD patients using the equally preprocessed, unmodified, full face 3D T1 series (=“full face”) as well as algorithmically defaced 3D T1 series using the aforementioned age- and sex-specific templates. Six defacing approaches were applied separately: afni\_refacer (AFNI v21.0.21 [7]), fsl\_deface (FSL v6.0.3 [10]), mri\_deface (FreeSurfer v7.1.1 [9]), mri\_reface v0.2 [3], PyDeface v2.0.0 from the Poldrack Lab [11] and spm\_deface from SPM12 r7771 [12]. The previously mentioned mask\_face (released by the Neuroinformatics Research Group [8]) was excluded from the analysis, since the defacing has been shown to be reversible using Cycle-Consistent

Adversarial Networks [33]. mri\_deface, fsl\_deface and PyDeface each apply a linear registration, atlas and mask-based approach to identify the face and remove it. fsl\_deface also removes the ears. afni\_refacer and mri\_reface replace the ears and face with a population average. mri\_reface furthermore replaces some regions of air, which may include identifiable features due to wraparound artifacts. All defacing approaches were run with their respective default settings. The automatic generation of QC images, offered by afni\_refacer and mri\_reface, was disabled.

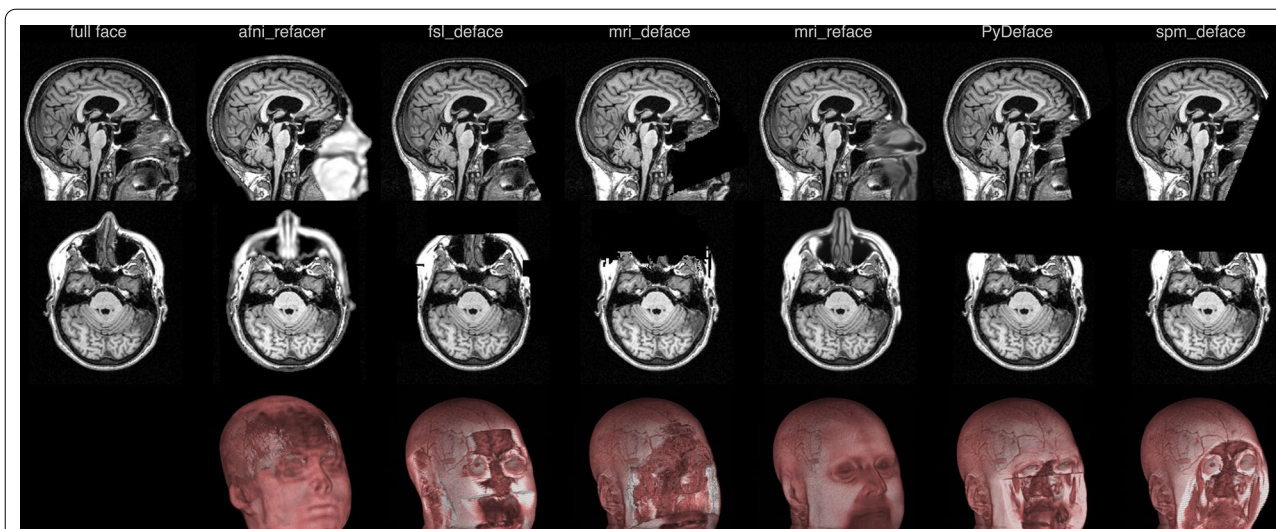
To analyze the impact of different defacing approaches on veganbagel’s atrophy estimation, gray matter atrophy z-scores after defacing were compared to the respective z-scores derived from the unmodified full face data. Specifically, for each voxel within the gray matter mask used in the veganbagel workflow, the root-mean-square error (RMSE) was calculated for the difference of the z-score after defacing minus the respective full face z-score.

Grubbs’s test, also known as the extreme studentized deviate test or maximum normalized residual test, was performed to identify outliers within the RMSE values of each defacing approach. Grubbs’s test by design only detects a single outlier. Therefore, the test was performed iteratively, i.e., the detected outlier was removed before rerunning the test until no more outliers are detected. In brief, Grubbs’s test is based on the difference of the mean and the minimum or maximum value of the data as determined by the standard deviation [34]. Grubbs’s test may produce false positives in distributions with a very large or very small standard deviation [35]. Therefore, in order to provide further context for the number of outliers detected by Grubbs’s test and to serve as a benchmark, z-score maps based on unmodified full face, unaccelerated 3D T1 imaging were compared to unmodified full face, unaccelerated within-session repeat 3D T1 imaging of the same subject, if available from the ADNI database (see Table 1, no repeats of accelerated imaging were acquired during ADNI). For each defacing approach the number of outliers with respect to the 75th percentile of the RMSE values of the benchmark results are reported.

Lastly, to visualize the regions most affected by each defacing approach, the absolute mean differences of the z-scores were plotted as a heat map onto representative axial slices of the SPM152 standard template, masked by the same, using MRICroGL [36].

## Results

In a total of 1877 of 1943 attempts (96.6%), the combination of defacing and the veganbagel workflow completed successfully. Examples are shown in Fig. 1.



**Fig. 1** Example of successful defacing approaches on an Alzheimer’s disease patient (female, 69 years of age). Top row shows the sagittal reformations of the defaced image volume, second row shows the axial reformations and the bottom row shows volume renderings. The volume rendering of the full face image (lower left) has been omitted for privacy reasons

afni\_refacer failed in two unaccelerated 3D T1 imaging, not yielding imaging volumes usable for further processing. mri\_deface crashed while processing 32 unaccelerated and 28 accelerated 3D T1 scans. In two accelerated 3D T1 imaging acquisitions, mri\_deface completed, but the adaptive maximum a posteriori (AMAP)-based segmentation step in CAT12 detected untypical tissue peaks and stopped further processing. The latter also occurred in two accelerated imaging acquisitions when using spm\_deface. In all other approaches, including full face and unaccelerated repeat imaging, no failed processing was noted.

The RMSE for the benchmark, comparing the gray matter z-scores of the full face unaccelerated 3D T1 imaging with the respective unaccelerated repeat imaging, is shown in the left column of Fig. 2. The mean benchmark RMSE was  $0.28 \pm 0.1$  (minimum: 0.12, 75th percentile: 0.33 and maximum: 0.58). No outliers were detected in the benchmark RMSE values using Grubbs’s test.

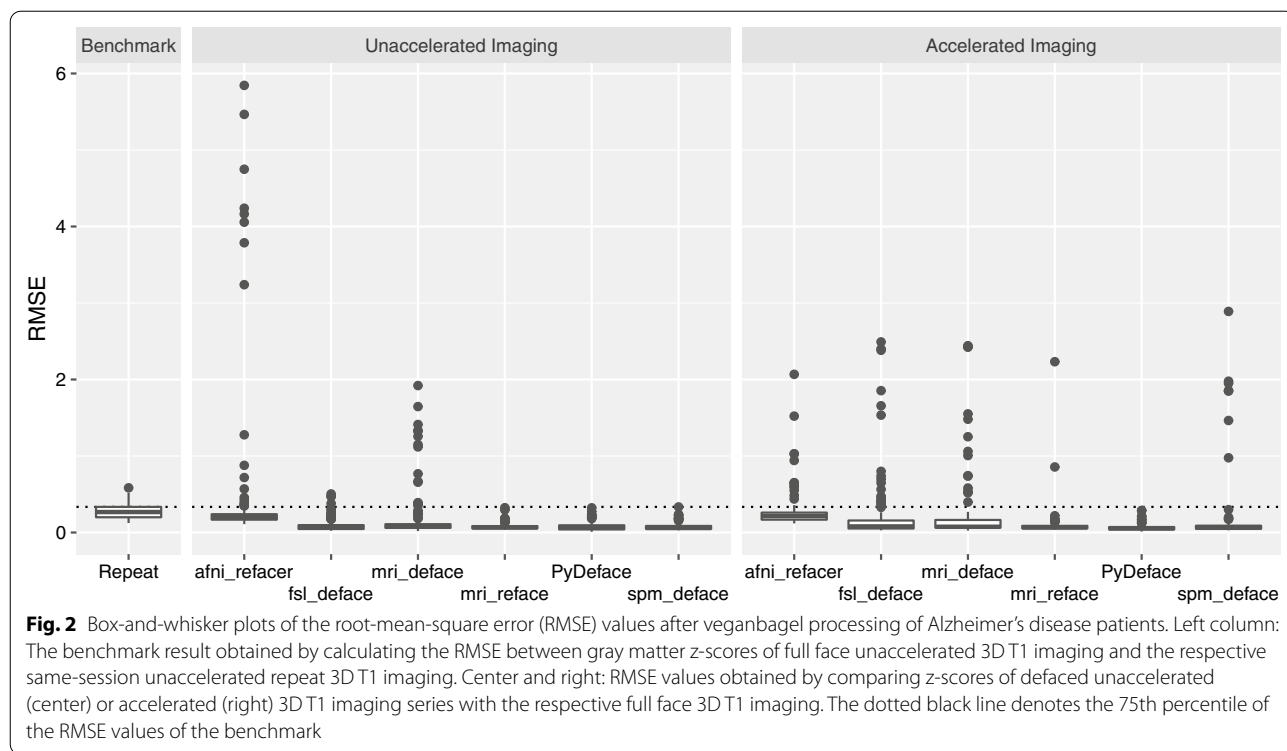
The RMSE comparing the gray matter z-scores based on the defaced 3D T1 imaging with the full face 3D T1 imaging series are shown in the center (unaccelerated imaging) and right (accelerated imaging) column of Fig. 2, respectively. RMSE values and outliers are furthermore reported in Table 2.

For defacing unaccelerated 3D T1 imaging, excellent results with a very small RMSE were found when applying fsl\_deface, mri\_reface, PyDeface and spm\_deface. However, Grubbs’s test detected 10, 7, 5 and 10 outliers, respectively, and for fsl\_deface, 4 outliers were detected with respect to the 75<sup>th</sup> percentile of the RMSE values

of the benchmark. Using afni\_refacer leads to a higher mean RMSE in comparison with the other approaches and 16 outliers according to Grubbs’s test and 18 outliers with respect to the benchmark were noted, respectively. mri\_deface, in comparison, leads to overall smaller RMSE values, but also results in several outliers (25 and 15, respectively) and, as noted above, crashed in a substantial amount of cases.

For the accelerated 3D T1 imaging, the smallest mean RMSE was obtained using PyDeface with 8 outliers in Grubbs’s test, which were found within a small range of the RMSE values from 0.01 to 0.29, and no outliers in comparison with the 75th percentile of the RMSE values of the benchmark. mri\_reface showed very good results overall with a small mean RMSE and a very small IQR. Six outliers were found according to Grubbs’s test, but also two outliers were noted in comparison with the benchmark with a RMSE of 0.86 and a relatively high RMSE of 2.23. Reviewing the defaced imaging volumes, no obvious errors in the defacing process were noted. spm\_deface and fsl\_deface performed worse in comparison with the unaccelerated 3D T1 imaging, with a higher mean RMSE and several more outliers. mri\_deface shows an overall similar performance when compared to the results from unaccelerated 3D T1 imaging with a slightly higher mean RMSE and wider range, as well as a comparable number of crashes. afni\_refacer results were better when compared to the unaccelerated imaging results, but still a comparable number of outliers were noted.

In order to visualize the regions most affected by each defacing approach, heat maps for the absolute mean



**Fig. 2** Box-and-whisker plots of the root-mean-square error (RMSE) values after veganbagel processing of Alzheimer’s disease patients. Left column: The benchmark result obtained by calculating the RMSE between gray matter z-scores of full face unaccelerated 3D T1 imaging and the respective same-session unaccelerated repeat 3D T1 imaging. Center and right: RMSE values obtained by comparing z-scores of defaced unaccelerated (center) or accelerated (right) 3D T1 imaging series with the respective full face 3D T1 imaging. The dotted black line denotes the 75th percentile of the RMSE values of the benchmark

differences of the z-scores between defaced and unmodified full face z-score maps are shown in Fig. 3 for unaccelerated and in Fig. 4 for accelerated imaging.

In unaccelerated imaging, afni\_refacer shows marked global deviations in z-scores, predominately over the left hemisphere. mri\_deface shows marked differences in the frontobasal brain and temporal poles, including changes in the basal ganglia, especially the caudate nuclei. fsl\_deface shows small focal differences in the frontal and frontobasal brain. The differences after mri\_reface, PyDeface and spm\_deface are less pronounced, with very small deviations, e.g., in the anterior frontobasal cortex, the thalami and occipital cortex after mri\_reface.

In accelerated imaging, in line with the previous analyses of the RMSE, the mean differences are globally higher after fsl\_deface, mri\_deface and spm\_deface. fsl\_deface and mri\_deface furthermore follow the same pattern of deviations with more pronounced deviations in the frontal (fsl\_deface) and frontobasal brain (fsl\_deface, mri\_deface). spm\_deface shows globally increased mean differences, especially including the basal ganglia. mri\_reface also shows a very slight global increase in mean differences, in accelerated imaging with a slight emphasis on the putamen, and less on the thalami. After afni\_refacer, less absolute mean differences overall are noted in accelerated imaging in comparison with unaccelerated imaging, showing a predominance over the right

hemisphere and focal differences along the occipital cortex and superficial cerebellum.

PyDeface shows no discernible patterns for both unaccelerated and accelerated imaging.

### Discussion

In the present study, we evaluated the impact of different defacing approaches on veganbagel, an open-source software built around CAT12 for SPM12. veganbagel allows for automatic brain atrophy estimation by comparing a subject’s structural brain scan to a normal cohort. Our results indicate that most defacing procedures are robust, with the exception of mri\_deface. Most of the defacing approaches introduced pronounced z-score deviations in the context of automatic brain atrophy estimation. The smallest bias with no notable outliers in comparison with the benchmark results was found when using PyDeface.

In recent years, ever stricter laws and regulations on data collection and processing have been imposed, introducing severe penalties for mishandling of information and data breaches. Recently, machine learning-based face recognition approaches have been shown to be alarmingly successful in matching photographs of participants to their respective MRI scans, with a success rate of up to 97% [2, 3]. On the other hand, concerns have been raised on the data integrity after defacing, especially with regard to common volumetric analysis [13, 14], while other

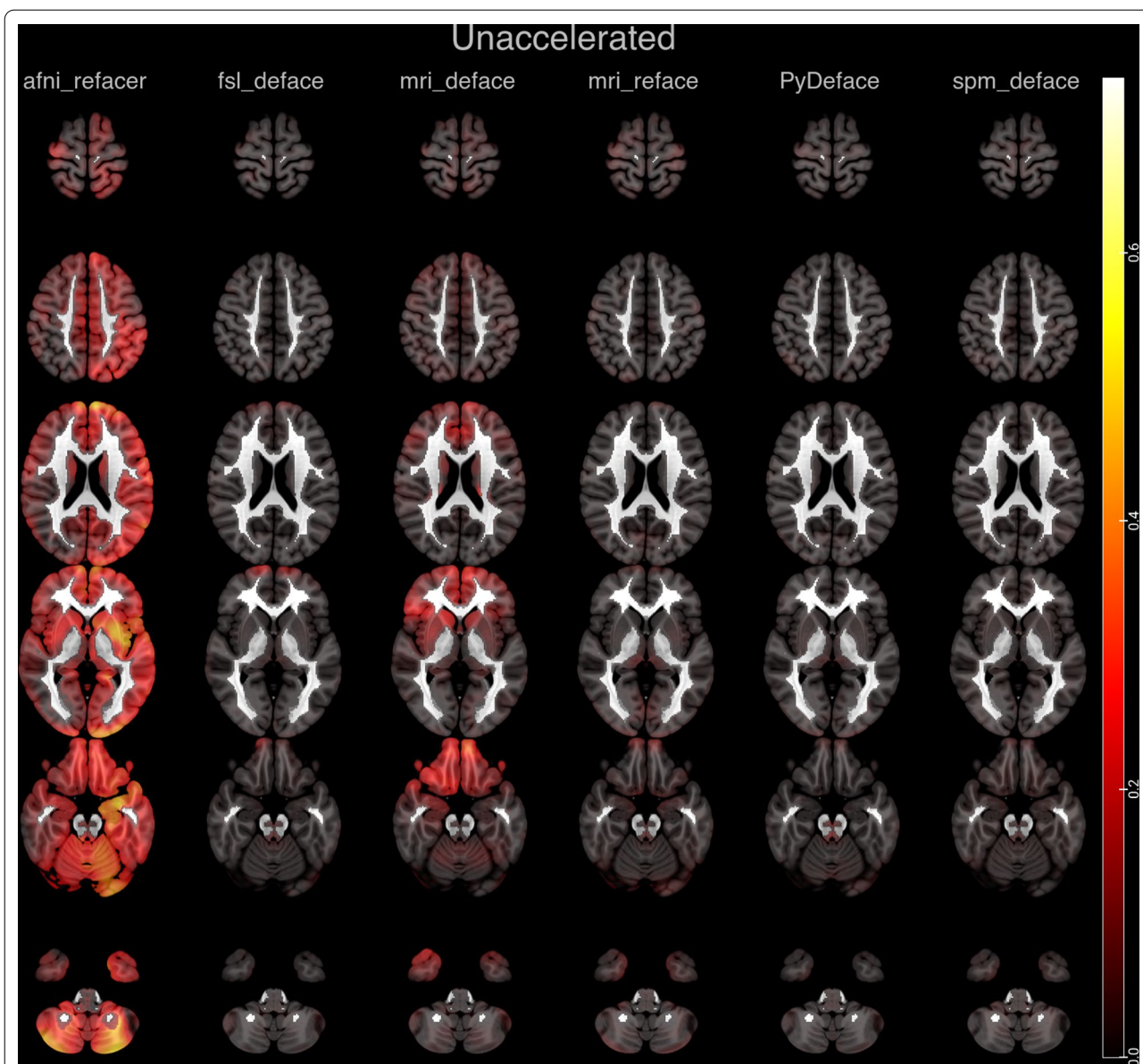
**Table 2** Voxel-wise z-score root-mean-square error (RMSE) values after veganbagel processing of Alzheimer’s disease patients and comparing the results after defacing with the respective z-scores derived from the respective full face 3D T1 imaging

	Unaccelerated imaging				Accelerated imaging			
	Failed processing <sup>a</sup>	Mean RMSE ± SD	Range (IQR)	Outliers Grubbs’s test Benchmark <sup>b</sup>	Failed processing <sup>a</sup>	Mean RMSE ± SD	Range (IQR)	Outliers Grubbs’s test Benchmark <sup>b</sup>
afni_refacer	2/154	0.45 ± 0.97	0.11–5.84 (0.24–2.16)	16 (10.5%) 18 (11.8%)	0/114	0.28 ± 0.26	0.12–2.07 (0.26–0.65)	13 (11.4%) 16 (14%)
fsi_deface	0/154	0.09 ± 0.08	0.02–0.50 (0.10–0.24)	10 (6.5%) 4 (2.6%)	0/114	0.23 ± 0.47	0.03–2.49 (0.16–1.06)	21 (18.4%) 18 (15.8%)
mri_deface	32/154	0.20 ± 0.35	0.02–1.92 (0.11–1.15)	25 (20.5%) 15 (12.3%)	30/114	0.30 ± 0.57	0.03–2.44 (0.16–1.54)	20 (23.8%) 15 (17.9%)
mri_reface	0/154	0.08 ± 0.04	0.03–0.32 (0.08–0.13)	7 (4.5%) 0 (0%)	0/114	0.10 ± 0.22	0.04–2.23 (0.09–0.14)	6 (5.3%) 2 (1.8%)
PyDeface	0/154	0.08 ± 0.05	0.01–0.32 (0.09–0.19)	5 (3.2%) 0 (0%)	0/114	0.07 ± 0.05	0.01–0.29 (0.07–0.17)	8 (7%) 0 (0%)
spm_deface	0/154	0.07 ± 0.05	0.03–0.33 (0.09–0.18)	10 (6.5%) 0 (0%)	2/114	0.18 ± 0.45	0.03–2.89 (0.09–1.20)	12 (10.5%) 7 (6.1%)

SD, standard deviation; IQR, interquartile range

<sup>a</sup> Failed processing denotes defacing crashing (the majority of cases) or veganbagel (i.e., CAT12 for SPM12) processing failing on the defaced image volume (n = 2 for mri\_deface and n = 2 for spm\_deface). Outliers are reported using Grubbs’s test for each individual approach

<sup>b</sup> By counting any RMSE values which were higher than the 75th percentile of the RMSE values of the benchmark result (0.33)

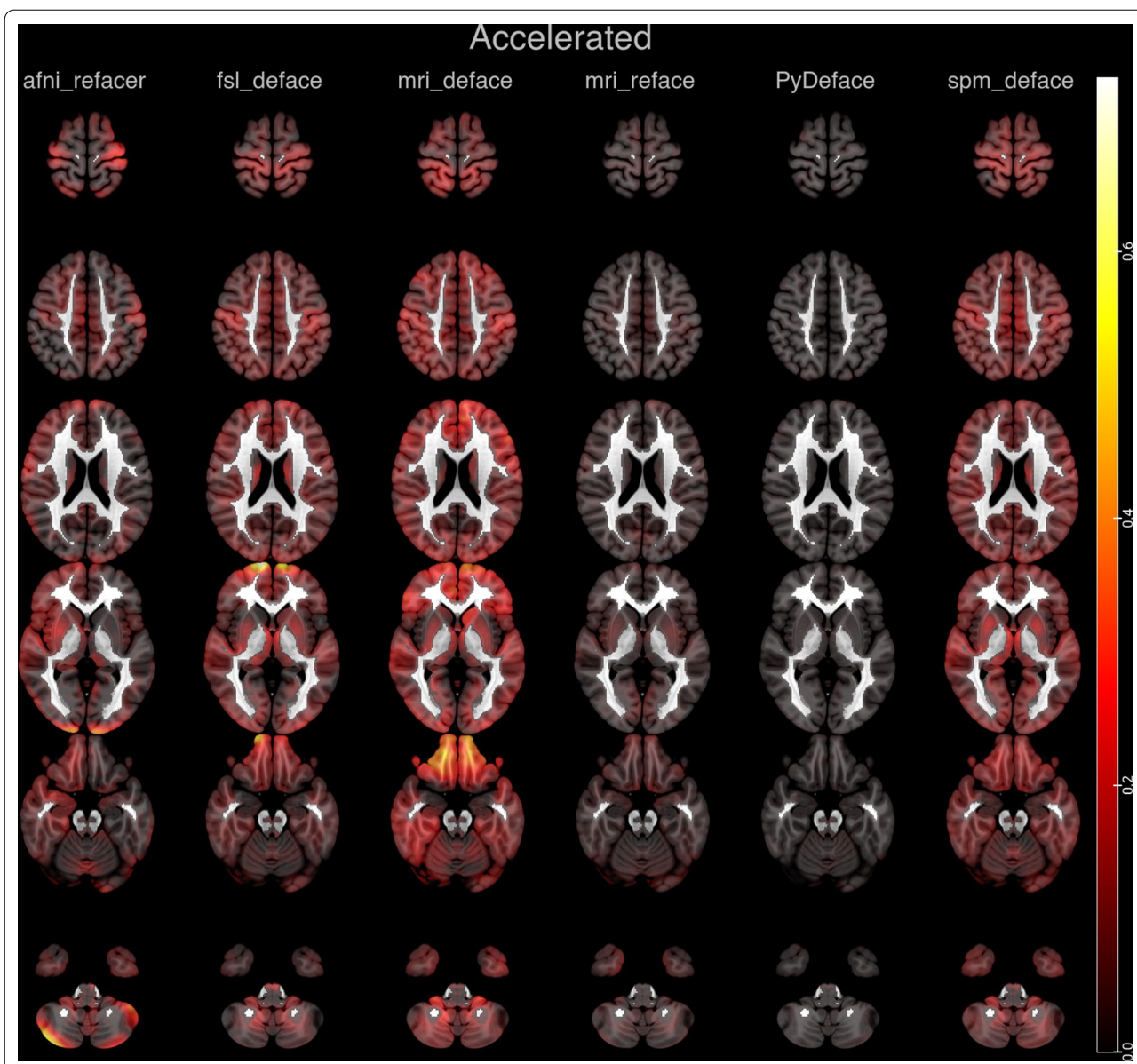


**Fig. 3** Absolute mean differences of the z-scores plotted as a heat map onto representative axial slices of the SPM152 standard template after applying the different defacing approaches on unaccelerated imaging

studies have shown modest to no effects of defacing on common neuroscientific analysis pipelines [3, 15].

Currently, there are two main approaches for anonymizing brain imaging data, “skull stripping,” i.e., removing everything from the image volume except for the brain, and removing just the facial features from the imaging volume, while retaining all other information. Automated skull stripping does suffer from limitations [37], and (accidental) removal of some voxels of the brain might severely limit any downstream analysis. Therefore, defacing is often the preferred method,

which is supposed to retain all information of the brain. The approaches tested in the current study generally register a brain scan to a common template and then apply a mask-based detection of the facial features to either remove these areas or replace the respective areas with a separately derived population average. In theory, defacing should therefore not interfere with, e.g., automated atrophy detection. However, we have noted deviations in a number of cases with sometimes large RMSEs, with no obvious changes to the brain in the defaced image volumes. Some neuroscientific



**Fig. 4** Absolute mean differences of the z-scores plotted as a heat map onto representative axial slices of the SPM152 standard template after applying the different defacing approaches on accelerated imaging

analysis pipelines, such as CAT12/SPM12, may use information from the background of the image volume in the detection of tissue classes (gray and white matter, cerebrospinal fluid), which might explain the observed deviations after defacing.

veganbagel denotes z-scores between  $-2.5$  to  $2.5$  as volume changes within expected limits [32]. With respect to these we found an extremely small mean RMSE in the benchmark assessments of full face unaccelerated 3D T1 imaging and the respective unaccelerated repeats. veganbagel may therefore be considered to

yield robust results in a heterogeneous dataset acquired on multiple different scanners.

We found that atrophy estimation results were susceptible to changes when defacing is conducted using afni\_refacer, fsl\_deface (pronounced in accelerated 3D T1 imaging), mri\_deface or spm\_deface (only in accelerated 3D T1 imaging), but not PyDeface. PyDeface resulted in a very small mean RMSE, no outliers noted with respect to the benchmark results and no clearly discernible pattern of deviations in the absolute mean differences maps. There were some outliers found by Grubbs's test, but



these must be placed in the context of a very small range of RMSE values, smaller, in fact, than the range of the benchmark RMSE values. In essence, no relevant changes in atrophy estimation are expected when processing brain scans defaced using PyDeface with *veganbagel*. Visualizing the regions of the brain most affected after defacing, *fsl\_deface* and *mri\_deface* demonstrated deviations mostly close to the face, but in accelerated imaging, more global deviations were observed for *fsl\_deface*, *mri\_deface* and *spm\_deface*, including the cerebellum and basal ganglia. Last, but not least, it has to be noted that in two cases both *afni\_refacer* and *spm\_deface* failed to yield image volumes usable for further analysis, while *mri\_deface* was not able to process a substantial number of the brain scans with the default settings. *mri\_reface* performed slightly better in this regard and showed very good overall results in our current study with regard to the RMSE, but two outliers, one of which with a relatively high RMSE, were noted after defacing accelerated 3D T1 imaging acquisitions.

Further considering privacy issues, *mri\_deface* and PyDeface sometimes retain parts of the orbits after defacing [3], and even imperfect, partial data may be of use for face recognition. [38] Schwarz et al. have shown that machine learning-based face recognition is still successful in 3% (*fsl\_deface*) to 10% (*mri\_deface* and PyDeface) of the cases after defacing by removing identifying parts of the face. However, the success rate of identifying subjects after defacing may be pushed to 28–38% when replacing missing parts of the face in defaced image volumes with a population average (*fsl\_deface* = 28%, *mri\_reface* = 30%, *mri\_deface* = 33% and PyDeface = 38%). [3]

Furthermore, we have noted differences after defacing unaccelerated and accelerated imaging. Accelerated imaging is often used to reduce scan time [39]. One of the benefits of accelerated imaging is increased patient comfort and compliance, which leads to less motion artifacts and may decrease the risk of study exclusion due to unusable imaging [30]. On the other hand, accelerated (parallel) imaging generally has a lower signal-to-noise ratio and may suffer from residual aliasing artifacts, altering tissue contrasts, as well as noise enhancements [39]. Depending on the underlying approach for defacing as well as brain atrophy estimation, these differences between unaccelerated and accelerated imaging may be further pronounced by local and distant effects of linear or nonlinear registration of an image volume to a template or by using information from the image background in order to classify tissue (e.g., in a Bayesian approach). Consequently, studies with different approaches to atrophy estimation have shown varying degrees of differences in atrophy assessments when using accelerated vs. unaccelerated imaging [28–31]. Likely, for the same reasons,

defacing of unaccelerated imaging generally was more robust, with the exception of PyDeface and to some degree *mri\_reface*, which performed well on both accelerated and unaccelerated imaging.

In synopsis, choosing the most suitable approach for defacing is a multifactorial decision. For example, *fsl\_deface* may currently provide a better defense than PyDeface against identification of a subject [3], but results of brain atrophy estimation may be biased in the frontal and frontobasal brain—especially in accelerated imaging. Further research is needed for different analysis pipelines and defacing approaches to carefully consider the trade-offs in result accuracy and privacy to choose the most suitable approach for the task at hand.

Our study is limited by restricting the analysis to only include AD patients from the ADNI database, and therefore, our results may not be transferable to other cohorts. However, ADNI was deliberately chosen, since it offers raw, unprocessed, non-defaced DICOM data of a wide variation of examinations from a large number of scanners from different vendors, which is expected to make our results more generalizable. Lastly, it has to be noted that different software versions of the defacing approaches or CAT12/SPM12 might lead to different results, and the respective release notes need to be monitored for major changes.

## Conclusion

Given the recent successes of applying face recognition algorithms to T1 imaging of the brain, some form of de-identification of MRI scans depicting facial features or the ears must be strongly considered when making data publicly available and possibly even when sending data to, e.g., cloud-based processing or analysis services. Especially PyDeface showed very good results with negligible impact on atrophy estimation. *mri\_reface* was found to be very promising and future versions should be re-evaluated. Furthermore, *veganbagel* demonstrated robust atrophy estimation results when comparing initial and repeat full face, unmodified imaging.

## Abbreviations

AD: Alzheimer's disease; ADNI: Alzheimer's Disease Neuroimaging Initiative; CAT: Computational anatomy toolbox; IQR: Interquartile range; MRI: Magnetic resonance imaging; QC: Quality control; RMSE: Root-mean-square error; SD: Standard deviation; SPM: Statistical parametric mapping; *Veganbagel*: Volumetric estimation of gross atrophy and brain age longitudinally.

## Acknowledgements

Computational infrastructure and support was provided by the Center for Information and Media Technology (ZIM) at the Heinrich Heine University of Duesseldorf (Germany). Collection and sharing of the Alzheimer's Disease Neuroimaging Initiative (ADNI) data used for evaluation in this study was funded by the (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical

Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (<https://fnih.org>). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

#### Authors' contributions

CR, DMH and JC contributed to the conception of the current study. CR and JC designed the experiment and analyzed the data. CR, DMH, RG and JC contributed to the interpretation of the data. CR conducted the experiments and created new software for the current study. CR drafted the manuscript and all authors made substantial contributions to the revisions of the manuscript. All authors read and approved the final manuscript

#### Funding

Open Access funding enabled and organized by Projekt DEAL. Robert Dahnke was funded by the DFG project DA 2167/1-1.

#### Availability of data and materials

The datasets analyzed during the current study are available in the Alzheimer's Disease Neuroimaging Initiative (ADNI) repository, <http://adni.loni.usc.edu>. `veganbagel` is available from <https://github.com/BrainImAccs/veganbagel>. `afni_refacer` is available from <https://afni.nimh.nih.gov>. `fsl_deface` is available from <https://fsl.fmrib.ox.ac.uk>. `mri_deface` is available from [https://surfer.nmr.mgh.harvard.edu/fswiki/mri\\_deface](https://surfer.nmr.mgh.harvard.edu/fswiki/mri_deface). `mri_reface` is available from [https://www.nitrc.org/projects/mri\\_reface](https://www.nitrc.org/projects/mri_reface). `PyDeface` is available from <https://github.com/poldracklab/pydeface>. `spm_deface` is available from <https://www.fil.ion.ucl.ac.uk/spm/>.

#### Declarations

##### Ethics approval and consent to participate

The study was approved by the ethics committee at the Medical Faculty of the Heinrich Heine University Düsseldorf (study number 2021-1424). Only publicly available data was used. The requirement for a written informed consent was therefore waived.

##### Consent for publication

Not applicable.

##### Competing interests

The authors declare that they have no competing interests.

##### Author details

<sup>1</sup>University Dusseldorf, Medical Faculty, Department of Diagnostic and Interventional Radiology, D-40225 Dusseldorf, Germany. <sup>2</sup>Department of Diagnostic and Interventional Neuroradiology, School of Medicine, Technical University of Munich, 81675 Munich, Germany. <sup>3</sup>Departments of Neurology and Psychiatry, Jena University Hospital, 07745 Jena, Germany. <sup>4</sup>Institut of Psychology, Friedrich Schiller University Jena, 07743 Jena, Germany. <sup>5</sup>Center of Functionally Integrative Neuroscience, Aarhus University, 8000 Aarhus, Denmark.

Received: 20 September 2021 Accepted: 19 February 2022

Published online: 26 March 2022

#### References

- Prior FW, Brunsden B, Hildebolt C et al (2009) Facial recognition from volume-rendered magnetic resonance imaging data. *IEEE Trans Inf Technol Biomed* 13:5–9. <https://doi.org/10.1109/titb.2008.2003335>
- Schwarz CG, Kremers WK, Therneau TM et al (2019) Identification of anonymous MRI research participants with face-recognition software. *N Engl J Med* 381:1684–1686. <https://doi.org/10.1056/nejmc1908881>
- Schwarz CG, Kremers WK, Wiste HJ et al (2021) Changing the face of neuroimaging research: comparing a new MRI de-facing technique with popular alternatives. *Neuroimage* 231:117845. <https://doi.org/10.1016/j.neuroimage.2021.117845>
- Emeršič Ž, Štruc V, Peer P (2017) Ear recognition: more than a survey. *Neurocomputing* 255:26–39. <https://doi.org/10.1016/j.neucom.2016.08.139>
- Essen DCV, Smith SM, Barch DM et al (2013) The WU-minn human connectome project: an overview. *Neuroimage* 80:62–79. <https://doi.org/10.1016/j.neuroimage.2013.05.041>
- Nooner KB, Colcombe SJ, Tobe RH et al (2012) The NKI-rockland sample: a model for accelerating the pace of discovery science in psychiatry. *Front Neurosci* 6:152. <https://doi.org/10.3389/fnins.2012.00152>
- Cox RW (1996) AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. *Comput Biomed Res Int J* 29:162–173. <https://doi.org/10.1006/cbmr.1996.0014>
- Milchenko M, Marcus D (2013) Obscuring surface anatomy in volumetric imaging data. *Neuroinformatics* 11:65–75. <https://doi.org/10.1007/s12021-012-9160-3>
- Bischoff-Grethe A, Ozyurt IB, Busa E et al (2007) A technique for the deidentification of structural brain MR images. *Hum Brain Mapp* 28:892–903. <https://doi.org/10.1002/hbm.20312>
- Alfaro-Almagro F, Jenkinson M, Bangerter NK et al (2018) Image processing and quality control for the first 10,000 brain imaging datasets from UK Biobank. *Neuroimage* 166:400–424. <https://doi.org/10.1016/j.neuroimage.2017.10.034>
- Gulban OF, Nielson D, Poldrack R et al (2019) poldracklab/pydeface: v2.0.0. Zenodo. <https://doi.org/10.5281/zenodo.3524401>
- Penny W, Friston K, Ashburner J et al (2006) *Statistical parametric mapping: the analysis of functional brain images*. Elsevier, Boston
- Bhalerao GV, Parekh P, Saini J et al (2021) Systematic evaluation of the impact of defacing on quality and volumetric assessments on T1-weighted MR-images. *J Neuroradiol*. <https://doi.org/10.1016/j.neurad.2021.03.001>
- de Sitter A, Visser M, Brouwer I et al (2020) Facing privacy in neuroimaging: removing facial features degrades performance of image analysis methods. *Eur Radiol* 30:1062–1074. <https://doi.org/10.1007/s00330-019-06459-3>
- Theyers AE, Zamyadi M, O'Reilly M et al (2021) Multisite comparison of MRI defacing software across multiple cohorts. *Front Psych* 12:617997. <https://doi.org/10.3389/fpsyg.2021.617997>
- Hedderich DM, Dieckmeyer M, Andrisan T et al (2020) Normative brain volume reports may improve differential diagnosis of dementing neurodegenerative diseases in clinical practice. *Eur Radiol* 30:2821–2829. <https://doi.org/10.1007/s00330-019-06602-0>
- Fumagalli GG, Basilico P, Arighi A et al (2018) Distinct patterns of brain atrophy in Genetic Frontotemporal Dementia Initiative (GENFI) cohort revealed by visual rating scales. *Alzheimers Res Ther* 10:46. <https://doi.org/10.1186/s13195-018-0376-9>
- Johnson EB, Gregory S (2019) Huntington's disease: brain imaging in Huntington's disease. *Prog Mol Biol Transl Sci* 165:321–369. <https://doi.org/10.1016/bs.pmbts.2019.04.004>
- Reetz K, Gaser C, Klein C et al (2009) Structural findings in the basal ganglia in genetically determined and idiopathic Parkinson's disease. *Mov Disord* 24:99–103. <https://doi.org/10.1002/mds.22333>
- Boxer AL, Geschwind MD, Belfor N et al (2006) Patterns of brain atrophy that differentiate corticobasal degeneration syndrome from progressive supranuclear palsy. *Arch Neurol* 63:81–86. <https://doi.org/10.1001/archneur.63.1.81>
- Sastre-Garriga J, Pareto D, Battaglini M et al (2020) MAGNIMS consensus recommendations on the use of brain and spinal cord atrophy measures in clinical practice. *Nat Rev Neurol* 16:171–182. <https://doi.org/10.1038/s41582-020-0314-x>

22. Scheltens P, Pasquier F, Weerts JG et al (1997) Qualitative assessment of cerebral atrophy on MRI: inter- and intra-observer reproducibility in dementia and normal aging. *Eur Neurol* 37:95–99. <https://doi.org/10.1159/000117417>
23. Kloppel S, Yang S, Kellner E et al (2018) Voxel-wise deviations from healthy aging for the detection of region-specific atrophy. *NeuroImage Clin* 20:851–860. <https://doi.org/10.1016/j.nicl.2018.09.013>
24. Mueller SG, Weiner MW, Thal LJ et al (2005) The Alzheimer's disease neuroimaging initiative. *Neuroimaging Clin N Am* 15:869-77-xi-xii. <https://doi.org/10.1016/j.nic.2005.09.008>
25. R Core Team (2020) R: a language and environment for statistical computing. R Foundation for Statistical Computing
26. Griswold MA, Jakob PM, Heidemann RM et al (2002) Generalized auto-calibrating partially parallel acquisitions (GRAPPA). *Magn Reson Med* 47:1202–1210. <https://doi.org/10.1002/mrm.10171>
27. Pruessmann KP, Weiger M, Scheidegger MB, Boesiger P (1999) SENSE: sensitivity encoding for fast MRI. *Magn Reson Med* 42:952–962. [https://doi.org/10.1002/\(sici\)1522-2594\(199911\)42:5%3c952::aid-mrm16%3e3.0.co;2-s](https://doi.org/10.1002/(sici)1522-2594(199911)42:5%3c952::aid-mrm16%3e3.0.co;2-s)
28. Vemuri P, Senjem ML, Gunter JL et al (2015) Accelerated vs. unaccelerated serial MRI based TBM-SyN measurements for clinical trials in Alzheimer's disease. *Neuroimage* 113:61–69. <https://doi.org/10.1016/j.neuroimage.2015.03.026>
29. Takao H, Amemiya S, Abe O, Initiative ADN (2021) Reproducibility of brain volume changes in longitudinal voxel-based morphometry between non-accelerated and accelerated magnetic resonance imaging. *J Alzheimer's Dis*. <https://doi.org/10.3233/jad-210596>
30. Manning EN, Leung KK, Nicholas JM et al (2017) A comparison of accelerated and non-accelerated MRI scans for brain volume and boundary shift integral measures of volume change: evidence from the ADNI dataset. *Neuroinformatics* 15:215–226. <https://doi.org/10.1007/s12021-017-9326-0>
31. Leung KK, Malone IM, Ourselin S et al (2015) Effects of changing from non-accelerated to accelerated MRI for follow-up in brain atrophy measurement. *Neuroimage* 107:46–53. <https://doi.org/10.1016/j.neuroimage.2014.11.049>
32. Caspers J, Heeger A, Turowski B, Rubbert C (2021) Automated age- and sex-specific volumetric estimation of regional brain atrophy: workflow and feasibility. *Eur Radiol* 31:1043–1048. <https://doi.org/10.1007/s00330-020-07196-8>
33. Abramian D, Eklund A (2019) Refacing: reconstructing anonymized facial features using GANS. In: 2019 IEEE 16th international symposium on biomedical imaging (ISBI 2019), pp 1104–1108. <https://doi.org/10.1109/ISBI.2019.8759515>
34. Grubbs FE (1950) Sample criteria for testing outlying observations. *Ann Math Stat* 21:27–58. <https://doi.org/10.1214/aoms/1177729885>
35. Xu L, Zhang P, Xu J et al (2010) High performance computing and applications. *Lect Notes Comput Sci*. [https://doi.org/10.1007/978-3-642-11842-5\\_66](https://doi.org/10.1007/978-3-642-11842-5_66)
36. Rorden C, Brett M (2000) Stereotaxic display of brain lesions. *Behav Neurol* 12:191–200. <https://doi.org/10.1155/2000/421719>
37. Kalavathi P, Prasath VBS (2016) Methods on skull stripping of MRI head scan images—a review. *J Digit Imaging* 29:365–379. <https://doi.org/10.1007/s10278-015-9847-8>
38. Elmahmudi A, Ugail H (2019) Deep face recognition using imperfect facial data. *Futur Gener Comput Syst* 99:213–225. <https://doi.org/10.1016/j.future.2019.04.025>
39. Deshmane A, Gulani V, Griswold MA, Seiberlich N (2012) Parallel MR imaging. *J Magn Reson Imaging* 36:55–72. <https://doi.org/10.1002/jmri.23639>

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Submit your manuscript to a SpringerOpen<sup>®</sup> journal and benefit from:

- Convenient online submission
- Rigorous peer review
- Open access: articles freely available online
- High visibility within the field
- Retaining the copyright to your article

---

Submit your next manuscript at ► [springeropen.com](https://www.springeropen.com)

---