



Facing privacy in neuroimaging: removing facial features degrades performance of image analysis methods

A. de Sitter¹ · M. Visser¹ · I. Brouwer¹ · K. S. Cover¹ · R. A. van Schijndel¹ · R. S. Eijelaar² · D. M. J. Müller³ · S. Ropele⁴ · L. Kappos⁵ · Á. Rovira⁶ · M. Filippi⁷ · C. Enzinger⁸ · J. Frederiksen⁹ · O. Ciccarelli¹⁰ · C. R. G. Guttman¹¹ · M. P. Wattjes^{1,12} · M. G. Witte² · P. C. de Witt Hamer³ · F. Barkhof^{1,13} · H. Vrenken¹ · on behalf of the MAGNIMS Study Group and Alzheimer's Disease Neuroimaging Initiative

Received: 18 April 2019 / Revised: 16 August 2019 / Accepted: 13 September 2019 / Published online: 5 November 2019
© The Author(s) 2019

Abstract

Background Recent studies have created awareness that facial features can be reconstructed from high-resolution MRI. Therefore, data sharing in neuroimaging requires special attention to protect participants' privacy. Facial features removal (FFR) could alleviate these concerns. We assessed the impact of three FFR methods on subsequent automated image analysis to obtain clinically relevant outcome measurements in three clinical groups.

Methods FFR was performed using QuickShear, FaceMasking, and Defacing. In 110 subjects of Alzheimer's Disease Neuroimaging Initiative, normalized brain volumes (NBV) were measured by SIENAX. In 70 multiple sclerosis patients of the MAGNIMS Study Group, lesion volumes (WMLV) were measured by lesion prediction algorithm in lesion segmentation toolbox. In 84 glioblastoma patients of the PICTURE Study Group, tumor volumes (GBV) were measured by BraTumIA. Failed analyses on FFR-processed images were recorded. Only cases in which all image analyses completed successfully were analyzed. Differences between outcomes obtained from FFR-processed and full images were assessed, by quantifying the intra-class correlation coefficient (ICC) for absolute agreement and by testing for systematic differences using paired *t* tests.

Results Automated analysis methods failed in 0–19% of cases in FFR-processed images versus 0–2% of cases in full images. ICC for absolute agreement ranged from 0.312 (GBV after FaceMasking) to 0.998 (WMLV after Defacing). FaceMasking

Electronic supplementary material The online version of this article (<https://doi.org/10.1007/s00330-019-06459-3>) contains supplementary material, which is available to authorized users.

✉ A. de Sitter
A.deSitter@vumc.nl

¹ Department of Radiology and Nuclear Medicine, Amsterdam Neuroscience Amsterdam UMC, location VUmc, Amsterdam, the Netherlands

² Department of Radiotherapy, The Netherlands Cancer Institute, Amsterdam, the Netherlands

³ Department of Neurosurgery, Amsterdam UMC, location VUmc, Amsterdam, the Netherlands

⁴ Department of Neurology, Medical University of Graz, Graz, Austria

⁵ Department of Neurology, University Hospital, Kantonsspital, Basel, Switzerland

⁶ Unitat de Resonància Magnètica (Servei de Radiologia), Hospital universitari Vall d'Hebron, Universitat Autònoma de Barcelona, Barcelona, Spain

⁷ Neuroimaging Research Unit, Institute of Experimental Neurology, Division of Neuroscience, San Raffaele Scientific Institute, UniSR, Milan, Italy

⁸ Division of Neuroradiology, Vascular and Interventional Radiology, Department of Radiology, Medical University of Graz, Graz, Austria

⁹ Department of Neurology, Glostrup University Hospital, Copenhagen, Denmark

¹⁰ UK/NIHR UCL-UCLH Biomedical Research Centre, Institute of Neurology, UCL, London, UK

¹¹ Center for Neurological Imaging, Department of Radiology, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA

¹² Department of Diagnostic and Interventional Neuroradiology, Hannover Medical School, Hannover, Germany

¹³ Institutes of Neurology & Healthcare Engineering, UCL, London, UK

yielded higher NBV ($p = 0.003$) and WMLV ($p \leq 0.001$). GBV was lower after QuickShear and Defacing (both $p < 0.001$).

Conclusions All three outcome measures were affected differently by FFR, including failure of analysis methods and both “random” variation and systematic differences. Further study is warranted to ensure high-quality neuroimaging research while protecting participants’ privacy.

Key Points

- *Protecting participants’ privacy when sharing MRI data is important.*
- *Impact of three facial features removal methods on subsequent analysis was assessed in three clinical groups.*
- *Removing facial features degrades performance of image analysis methods.*

Keywords Magnetic resonance imaging · Ethics · Database · Neuroimaging · Privacy

Abbreviations

AD	Alzheimer’s disease
BV	Unnormalized brain volume
EDSS	Expanded Disability Status Scale
FFR	Facial features removal
FLAIR	Fluid-attenuated inversion recovery
FN	False negative
FP	False positive
GBV	Tumor volume of glioblastoma patients
ICC	Intra-class correlation coefficient
LST-LPA	Lesion prediction algorithm in lesion segmentation toolbox
MCI	Mild cognitive impairment
MPRAGE	Magnetization-prepared rapid acquisition gradient echo
MR	Magnetic resonance
MRI	Magnetic resonance imaging
MS	Multiple sclerosis
NBV	Normalized brain volume
SI	Dice’s similarity index
T	Tesla
WML	White matter lesion
WMLV	White matter lesion volume

Introduction

Sharing participant image data can offer many benefits to neuroradiological research: a better understanding of diseases can be achieved by access to larger participant populations in combined multicenter datasets; researchers without access to their own data on a specific disease can still contribute to its understanding by using shared datasets; and methodological improvements can be stimulated by publicly shared benchmark datasets.

However, for shared data, it is crucial to protect participants’ privacy. Image files should not contain identifying information such as name, date of birth, or any national or hospital-based registration numbers. Such data are often saved in metadata or even filenames of magnetic resonance (MR) images and should be removed before sharing. Unfortunately, this is not

enough to alleviate privacy concerns, since typical structural magnetic resonance imaging (MRI) provides good enough skin to air contrast and spatial resolution to perform facial recognition from a 3D-rendered version of the image, whether by the human eye or using facial recognition software [1–5]. Therefore, in addition to identifying metadata, it has been suggested that facial features should also be removed, and this has been widely embraced [6–9]. However, it is not yet clear whether the removal of the facial features affects subsequent measurement of quantitative indices of brain pathology.

Therefore, the current study assessed the impact of facial features removal (FFR) on clinically relevant outcome measurements. We selected three FFR methods that are publicly available, well documented, and have been used in data sharing initiatives [10, 11]: QuickShear [12], FaceMasking [13], and Defacing [14]. We assessed their effects on clinically relevant outcome measures in three different diseases: normalized brain volumes (NBV) in Alzheimer’s disease (AD), white matter lesion volumes (WMLV) in multiple sclerosis (MS), and tumor volumes (GBV) in glioblastoma patients.

Materials and methods

Subject

Subjects in this study were obtained from three different dataset: for AD, a dataset from the ADNI study (<http://adni.loni.usc.edu/>) [15]; for MS, a multicenter dataset from the MAGNIMS Study Group (<https://www.magnims.eu/>) [16]; and for treatment-naïve glioblastoma patients, a clinical dataset from the PICTURE project collected in the Amsterdam UMC, location VUmc, in Amsterdam, the Netherlands. Primary studies were approved by the respective local ethics committee for all three datasets. A summary of the demographics is given in Supplementary Table 1.

Alzheimer’s disease

Data used in the preparation of this article were obtained from the ADNI database. The ADNI was launched in 2003 as a

public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial MRI, positron emission tomography, other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD).

From the ADNI1 dataset, we selected the subset of subjects who had a 3-Tesla (T) magnetization-prepared rapid acquisition gradient echo (MPRAGE) baseline MRI, which is a subset of the 562 subjects that are in the ADNI1 dataset [15, 17]. This subset included in total 110 (23% female) subjects with an average age of 75 (range 60–87) years. This dataset included 39 healthy elderly controls, 52 patients with mild cognitive impairment, and 19 patients with AD.

Multiple sclerosis

For MS, a multicenter dataset of the MAGNIMS Study Group was previously used to study iron accumulation in deep gray matter [18] and lesion segmentation software performance [16]. The dataset consisted of 70 patients (67% female), scanned in six different MAGNIMS centers. On average, the age was 34.9 (range 17–52) years. The mean disease duration from onset was 7.6 (range 1–28) years and the disease severity was measured using the Expanded Disability Status Scale (EDSS) on the day of scanning; patients had a median EDSS score of 2 (range 0.0–6.5) [19].

Glioblastoma

For glioblastoma, a total of 84 (38% female) patients were selected from a cohort treated at the Neurosurgical Center of the Amsterdam UMC, location VUmc, Amsterdam, the Netherlands, in 2012 and 2013. On average, the age was 61.4 (range 21–84) years. All patients had histopathologically confirmed WHO grade IV glioblastoma. The preoperative MRI was made on average within 1 week before resection.

MRI procedure

In the MS and AD datasets, all imaging was performed on 3-T whole-body MR systems, and for imaging of the glioblastoma dataset on 1.5- and 3-T MR systems. The protocol for the AD dataset included a 3D T1-weighted sequence, while the protocol for MS included a 3D T1-weighted sequence, as well as a 2D fluid-attenuated inversion recovery (FLAIR) sequence. The protocol for glioblastoma contained a 3D T1-weighted post contrast-enhanced scan, 3D FLAIR, and 2D T2-weighted and non-enhanced 2D T1-weighted sequence. In Table 1, more details are listed on data acquisition of the datasets.

Facial features removal methods

Three publicly available methods were selected: QuickShear [12], FaceMasking [13], and Defacing [14] (Fig. 1). For all three methods, default settings were used in this study. FaceMasking was applied on all MR modalities separately. QuickShear and Defacing can only remove facial features from 3D T1 images. To remove the facial features from the other images, the full 3D T1 image of each subject was registered to the other full images of the same subject, using FSL-FLIRT (<https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/FLIRT> [20]), with 12 degrees of freedom before applying the FFR methods. Using the resulting transformation matrices, the 3D T1 image without facial features was transformed to each of the other image spaces, and subsequently binarized and used as a mask to remove the face from the other images.

QuickShear

Starting from a user-supplied brain mask, QuickShear [12] uses two algorithms [21, 22] to create a plane that divides the MRI into two parts. One part contains the facial features, and the other part contains the remainder of the head, including the brain. After finding this plane, the intensity of all voxels on the “facial features” side of the plane is set to zero.

In this study, the brain mask was made with FSL-BET (<https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/BET>) using previously determined optimal settings [23, 24].

FaceMasking

FaceMasking [13] deforms the surface of the face with a filter. In this study, the normalized filtering method was used, which is the recommended filter. The method first selects the boundary of the skull and the face, and registers the image volume to an atlas with annotated face coordinates; then, the identified face region of interest layer is normalized and filtered and, finally, transformed back to the original image.

Defacing

Defacing [14] uses an algorithm that calculates the probability of voxels being brain tissue or part of the face, based on 10 annotated atlases of healthy controls. Voxels that are labeled as part of the face and have zero probability of being brain tissue are considered to contain facial features, and their signal intensities are set to zero in order to remove the facial features.

Clinical research outcome measurements

For all three datasets, commonly used, previously validated automated methods were used to obtain clinically relevant outcome measures on the full images (i.e., images without

Table 1 Details on the data acquisition of the AD, MS, and glioblastoma datasets

Dataset	Scanner brands	Scanner types	Field strength (Tesla)	Sequence	Sequence parameters			
					TR (ms)	TE (ms)	TI (ms)	Slice thickness (mm)
AD	Siemens GE Medical Systems Philips	Not known	3	3D T1	2300–3000	2.98	853–900	1.2
MS	Siemens	Trio	3	2D FLAIR	8000–11,000	69–136	2400–2800	3.0
	Philips	Achieva		3D T1	6.9–2300	2.8–298	815–900	1.0
Glioblastoma	Siemens	Sonata or Avanto	1.5 and 3	2D FLAIR	6500	355	2200	1.3
	GE Medical Systems	Signa HDxt or		3D T1*	2300–2700	4.5–5.0	950	1.0–1.5
	Toshiba	DISCOVERY		2D T1	520–600	8.0–12.0		5
	Philips	MR750		2D T2	5190–8670	93–101		5
		Titan3T Panorama HFO or Achieva						

*Post contrast (0.2 mmol/kg)

AD, Alzheimer's disease; MS, multiple sclerosis; FLAIR, fluid-attenuated inversion recovery; TR, repetition time; TE, echo time; TI, inversion time

FFR processing) as well as on all images after FFR. In the AD dataset, NBV and unnormalized brain volume (BV) were measured with SIENAX [23]. In the MS dataset, WMLV was measured by segmenting the lesions on the FLAIR images with the lesion prediction algorithm in the lesion segmentation toolbox (LST-LPA) software [25]. In the glioblastoma dataset, the GBV was measured by taking the union of the segmentation of the glioblastoma necrotic core and enhancing tumor generated by BraTumIA [26]. A short description of the methods is provided in the supplementary data.

To provide context to any observed differences between results from full images and images after FFR, reproducibility of SIENAX, LST-LPA, and BraTumIA was assessed. This was done by repeating the analysis on 10 native images per dataset, selected based on the results from the analyses of images after FFR to include in each case 5 images with large effects of FFR and 5 images with small effects of FFR.

Statistical analyses

First, we investigated whether the FFR methods would successfully process the data, and if the automated methods could

successfully process the data after FFR. A method was considered to have failed on a particular input image if the method gave an error or no output. Images were not excluded if the output quality was considered bad by human observers. The percentages of images for which the FFR methods produced output and the percentages of images for which FFR-processed images could be analyzed by the automated methods were calculated.

Next, the impact of the FFR on the outcome measures was evaluated. In order to allow a direct and fair comparison of metrics between FFR methods, only the subjects for whom all three FFR methods produced output and for which both the full images and all FFR-processed images could be analyzed by the subsequent image analysis method were included.

Volumetric analyses

The effect of FFR on volumes was evaluated by assessing changes in NBV and BV (AD dataset), WMLV (MS dataset), and GBV (glioblastoma dataset) in three different ways: in data distribution, variability, and systematic differences. First, to assess data distribution, histogram characteristics

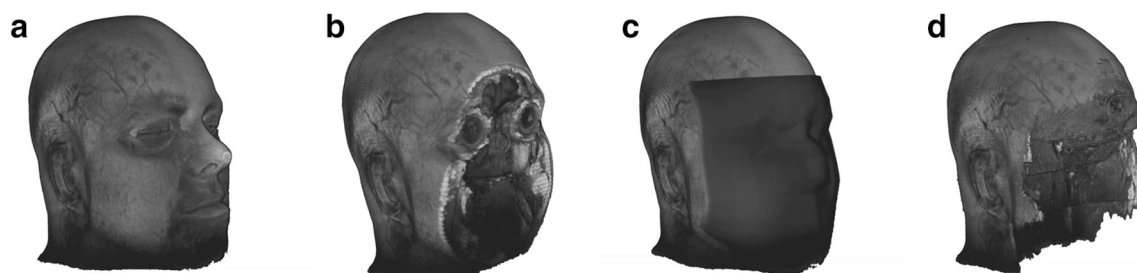


Fig. 1 Example 3D-rendered MRI: full (a) and after removal of facial features with QuickShear (b), FaceMasking (c), and Defacing (d). The subject gave written informed consent for using data and for displaying rendering in this figure

(median, first and third quartiles, means, and standard deviation) were calculated for four images (1 full; 3 FFR-processed) and difference characteristics (mL and percentage difference) were calculated for 3 FFR-processed images compared with the full image, and scatter plots and Bland-Altman plots were made. Second, to assess variability in the data, whether random we analyzed intra-class correlation coefficient (ICC) for absolute agreement between volumes obtained from full and FFR-processed images [27, 28] with the lower and upper bounds of 95% confidence interval [CI] in parentheses. Third, to assess systematic differences, two-tailed paired *t* tests were performed between volumes measured in full images and those obtained from each of the FFR-processed images, using a Bonferroni-corrected $p = 0.05$ as threshold for statistical significance.

Overlap analysis (MS and glioblastoma datasets)

In MS and glioblastoma datasets, we also compared voxelwise differences between the segmentations obtained with and without FFR, because the image analysis methods used in these datasets produce location-sensitive segmentations of the structures of interest. The full dataset is used as “gold standard” and is compared with each of the three FFR-processed datasets separately, quantifying spatial agreement using Dice’s similarity index (SI) [29]:

$$SI = (2 \times TP) / (2 \times TP + FP + FN)$$

where TP, FP, and FN are, respectively, true positive, false positive, and false negative volumes. SI can range from 0 to 1 and SI = 0 means no overlap and SI = 1, a perfect overlap. We calculated the median and first and third quartiles of SI, FP, and FN.

Results

Failure of pipelines

A simplified flowchart summarizing the study steps is shown in Fig. 2. An overview of the percentages of images for which the FFR methods and automated methods did not fail, i.e., executed without error and with output, is shown for each dataset in Table 2. FFR failed only in the glioblastoma dataset, specifically in 2% of cases for QuickShear and 1% of cases for FaceMasking. Automated method failures varied: while SIENAX completed successfully on all FFR-processed images (AD dataset), LST-LPA produced errors in 4% of cases for QuickShear and 19% of cases for Defacing (MS dataset); and BraTumIA in 17% of cases with QuickShear, 2% with FaceMasking, and 1% with Defacing (glioblastoma dataset). We excluded a subject from further analyses if at least one

FFR method failed on this subject. This resulted in 110, 55, and 66 subjects in the AD, MS, and GB datasets, respectively.

Volumetric analysis

Full image results

In all datasets, outcome measures obtained from the full images were in the expected range and showed expected distributions. The methods showed good reproducibility, as shown in Table 3, on 10 subjects per dataset, ICC (lower-upper band of 95% CI) of 0.988 (0.973–0.992), 0.998 (0.992–1.000), 0.996 (0.971–0.999), and 0.998 (0.998–1.000) for, respectively, the AD (NBV and BV), MS, and GB datasets. The top rows of Tables 4 and 5 provide the measured values of NBV, BV, WMLV, and GBV for the full images.

AD dataset

Both NBV and BV were affected by FFR processing, in terms of both variability and systematic differences. In Fig. 3, an example of effected SIENAX by FFR processing is given. Figure 4 a and b show scatter plots of NBV and BV for FFR-processed images versus full images in the AD dataset; corresponding Bland-Altman plots are provided in the supplementary section. These results suggest that FFR affected NBV variability more than BV variability, which is confirmed by the ICCs (Table 4): absolute agreement of NBV between FFR-processed images and full images ranged from 0.715 (Defacing) to 0.896 (FaceMasking), while for BV, absolute agreement ranged from 0.933 (Defacing) to 0.982 (QuickShear). Pairwise comparisons showed that NBV was typically overestimated after processing data with QuickShear (median [1st and 3rd quartiles] 1.26 [–4.40; 8.62] mL) and FaceMasking (5.91 [–1.57; 16.38] mL), and underestimated after Defacing (–2.17 [–10.80; 7.81] mL). BV was typically underestimated after FFR, with median [1st and 3rd quartiles] volume differences of –2.44 [–5.11; –1.22] mL (QuickShear), –0.52 [–2.47; 1.10] mL (FaceMasking), and –7.60 [–12.13; –4.71] mL (Defacing). The supplementary section provides means and standard deviations.

MS dataset

For WMLV, absolute agreement between FFR-processed images and full images was high, but there were small but significant systematic differences. In Fig. 5, an example of affected lesion segmentation by FFR processing is given. Figure 4c shows the WMLV scatter plot; corresponding Bland-Altman plots are provided in the supplementary section. The corresponding ICCs in Table 5 are all ≥ 0.988 . The median [1st and 3rd quartiles] WMLV values after FFR were 2.94 [1.35; 8.12] mL (QuickShear), 3.15 [1.60; 8.15] mL (FaceMasking), and

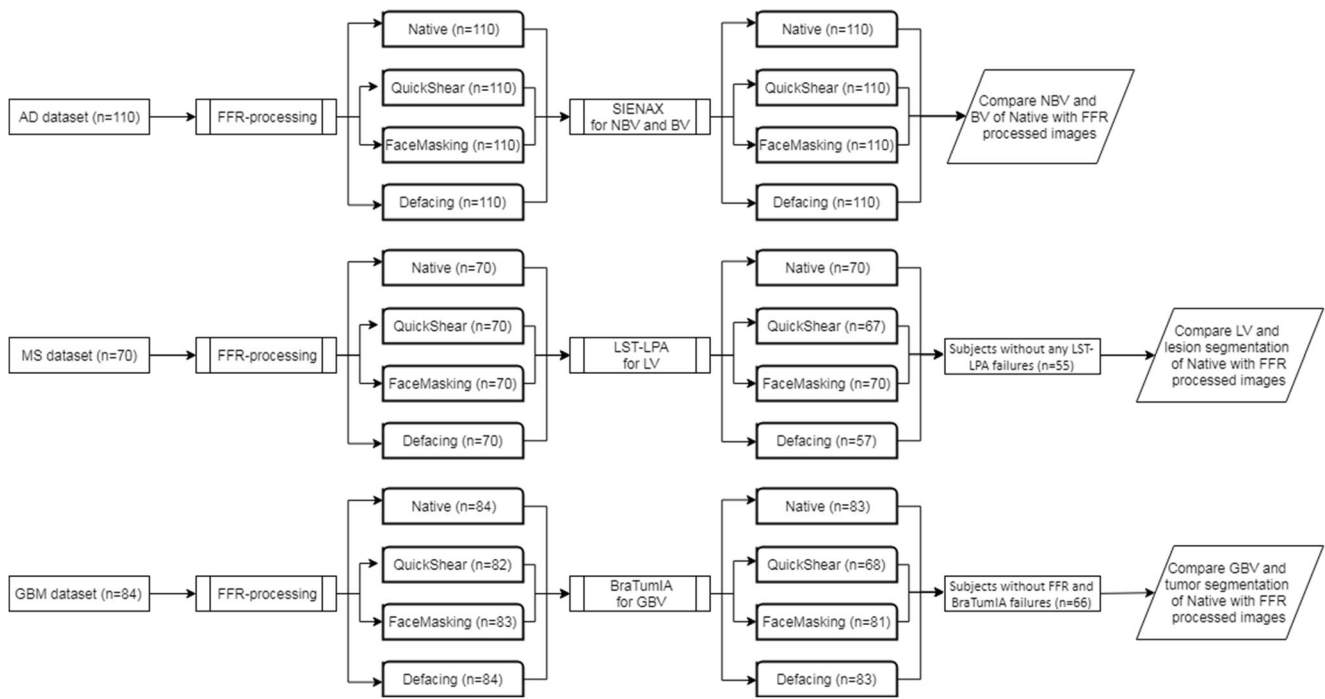


Fig. 2 A flowchart summarizing the study steps. Starting with 3 datasets which are FFR-processed, followed with automated (segmentation) methods, selection subjects, and comparing outcome measurements of the FFR-processed images with native images

2.77 [1.21; 7.77] mL (Defacing). For the full image, the WMLV value was 2.71 [1.32; 7.76] mL. In the case of FaceMasking, the WMLV is significantly higher than the WMLV of full images, Bonferroni-corrected $p < 0.001$. The supplementary section provides means and standard deviations.

Glioblastoma dataset

GBV appeared to be the outcome measure that is most strongly affected by FFR, with low to poor agreement and systematic volume underestimation after FFR by FaceMasking. In

Fig. 6, an example of affected glioblastoma segmentation by FFR processing is given. The scatter plot in Fig. 4d, the corresponding Bland-Altman plots in the supplementary section, and the ICCs in Table 5 show that FFR was affected, irrespective of tumor size. ICCs of 0.843, 0.312, and 0.810 between the full and FFR-processed images for QuickShear, FaceMasking, and Defacing indicate substantial effects of FFR. GBV were lower after FFR: differences with full image values were $-2.46 [-7.08; -0.54]$ mL (QuickShear), $-1.31 [-7.74; 0.57]$ mL (FaceMasking), and Defacing $-3.28 [-8.16; -0.72]$ mL (Defacing). The supplementary section provides means and standard deviations.

Table 2 Amount and percentage of images for which FFR methods completed successfully (left half of table) and for which the automated methods SIENAX, LST-LPA, and BraTumIA completed successfully (right half of table)

	Facial features removal			Measurement		
	AD (n = 110)	MS (n = 70)	Glioblastoma (n = 84)	AD (n = 110) SIENAX	MS (n = 70) LST-LPA	Glioblastoma (n = 84) BraTumIA
Full				110/110 100%	70/70 100%	83/84 99%
QuickShear	110/110 100%	70/70 100%	82/84 98%	110/110 100%	67/70 96%	68/82 83%
FaceMasking	110/110 100%	70/70 100%	83/84 99%	110/110 100%	70/70 100%	81/83 98%
Defacing	110/100 100%	70/70 100%	84/84 100%	110/110 100%	57/70 81%	83/84 99%

AD, Alzheimer’s disease; MS, multiple sclerosis

Table 3 Reproducibility of automated methods on 10 subjects per dataset. From left to right, the table lists median [1st and 3rd quartiles] for volumes; mean \pm std for volumes; *p* values for the pairwise comparison of volumes from first and second time processed full

images; ICC (absolute agreement (lower-upper band of 95% CI)) between volumes from full and FFR-processed images; and Dice's similarity index between segmentation from first and second time processed full images

<i>N</i> = 10		Volume		<i>p</i> value	ICC	SI
AD NBV (L)	First	1.34 [1.29; 1.38]	1.32 \pm 0.10	0.30	0.988 (0.973–0.992)	
	Second	1.34 [1.30; 1.40]	1.34 \pm 0.07			
AD BV (L)	First	1.09 [1.00; 1.13]	1.06 \pm 0.13	0.35	0.998 (0.992–1.000)	
	Second	1.08 [1.30; 1.40]	1.07 \pm 0.14			
MS LV (mL)	First	3.00 [1.64; 8.08]	6.39 \pm 6.87	0.16	0.996 (0.971–0.999)	0.95 [0.84–0.98]
	Second	3.82 [2.11; 8.35]	6.68 \pm 6.74			
GB GBV (mL)	First	39.89 [25.11; 85.92]	68.50 \pm 69.66	0.06	0.998 (0.998–1.000)	0.94 [0.89–0.96]
	Second	40.65 [25.28; 86.61]	69.38 \pm 70.16			

ICC, intra-class correlation coefficient; *N* = amount of subjects; *std*, standard deviation; *AD*, Alzheimer's disease; *MS*, multiple sclerosis; *SI*, Dice's similarity index; *FN*, false negative; *FP*, false positive

Overlap analysis

MS dataset

Table 5 also presents the spatial agreement of the WM lesion (WML) segmentations. The median [1st and 3rd quartiles] SI values of the WML segmentations from FFR-processed images with the corresponding segmentations from full images were 0.93 [0.86; 0.95] (QuickShear), 0.90 [0.72; 0.94] (FaceMasking), and 0.86 [0.39; 0.94] (Defacing). The FN volumes of FFR-processed images were 0.27 [0.11; 0.76] mL (QuickShear), 0.24 [0.09; 0.66] mL (FaceMasking), and 0.59 [0.17; 1.35] mL (Defacing). The FP volumes were 0.24 [0.11; 0.74] mL (QuickShear), 0.58 [0.12; 2.01] mL (FaceMasking), and 0.48 [0.20; 1.10] mL (Defacing). The supplementary section provides means and standard deviations.

Glioblastoma dataset

The SI, FN, and FP of the glioblastoma segmentation are shown in Table 5. The median [1st and 3rd quartiles] SI values of the FFR-processed glioblastoma segmentations with full image segmentations were 0.87 [0.75; 0.92] (QuickShear), 0.86 [0.74; 0.92] (FaceMasking), and 0.86 [0.74; 0.92] (Defacing). The FN volumes were 5.33 [2.70; 9.74] mL (QuickShear), 4.81 [2.00; 11.75] mL (FaceMasking), and 5.94 [2.79; 9.75] mL (Defacing). The FP volumes were 1.89 [0.79; 3.80] mL (QuickShear), 2.91 [1.79; 4.77] mL (FaceMasking), and 1.85 [0.79; 3.21] mL (Defacing). The supplementary section provides means and standard deviations.

Discussion

When sharing MRI data between research institutions, it is crucial to protect the privacy of participants. In addition to removing identifying metadata from MRI, facial features should also be removed. The current study evaluated how three publicly available FFR methods affect clinically relevant imaging outcome measures in AD, MS, and glioblastoma as derived using commonly used automated methods. Our results showed that the commonly used FFR methods can lead to subsequent failures of automated volumetric pipelines. Moreover, FFR can lead to substantial changes—both random (low ICC) and systematic (significant differences)—in volumes obtained by automated methods. The observed differences in outcome measures between full images and images after FFR cannot be attributed to random variation of SIENAX, LST-LPA, or BraTumIA, because the reproducibility of those methods was high.

The automated methods LST-LPA for WMLV and BraTumIA for GBV failed to successfully execute on multiple FFR-processed images. It should be mentioned that we applied the automated methods with their default settings and did not attempt to remedy the errors. We did, however, assess the failures and we suspect that the failures were related to image registration steps, because registration methods can be susceptible to (disease related) artifacts as recently highlighted by Dadar et al [30]. This recent study showed that registration used in the automated methods could have problems with higher levels of noise and non-uniformity in images and that head size could have an effect on registration methods. It is conceivable that if the face is removed or deformed, the level of noise and non-uniformity could change and lead to failures.

The possible importance of image registration in causing changes after FFR is further suggested by the higher variability of NBV compared with BV after FFR. To compute the

Table 4 Normalized brain volume and brain volume calculated with SIENAX in the AD dataset. From left to right, the table lists median [1st and 3rd quartiles] for volumes; Bonferroni-corrected *p* values for the pairwise comparison of volumes from FFR-processed and full images; volume differences between the full and FFR-processed images as mL and % differences (median [1st and 3rd quartiles]); and ICC (absolute agreement (lower-upper band of 95% CI)) between volumes from full and FFR-processed images

	Normalized brain volume			Brain volume						
	Volume (L)	<i>p</i> value	Difference	ICC (lower-upper)	Volume (L)	<i>p</i> value	Difference	ICC (lower-upper)		
			mL					Ml		
Full	1.39 [1.34; 1.44]				1.09 [1.01; 1.16]					
QuickShear	1.39 [1.35; 1.44]	0.266	1.26 [-4.40; 8.62]	0.835 (0.767–0.884)	1.09 [1.01; 1.16]	0.001	-2.44 [-5.11; -1.22]	0.982 (0.982–0.993)	-0.05 [-0.48; -0.11]	
FaceMasking	1.40 [1.35; 1.46]	0.003	5.91 [-1.57; 16.38]	0.896 (0.842–0.931)	1.09 [1.01; 1.16]	0.392	-0.52 [-2.47; 1.10]	0.973 (0.960–0.981)	-0.05 [-0.2; 0.10]	
Defacing	1.39 [1.35; 1.44]	1.000	-2.17 [-10.80; 7.81]	0.715 (0.610–0.795)	1.07 [0.99; 1.14]	<0.001	-7.60 [-12.13; -4.71]	0.933 (0.874–0.961)	-0.69 [-1.14; 0.43]	

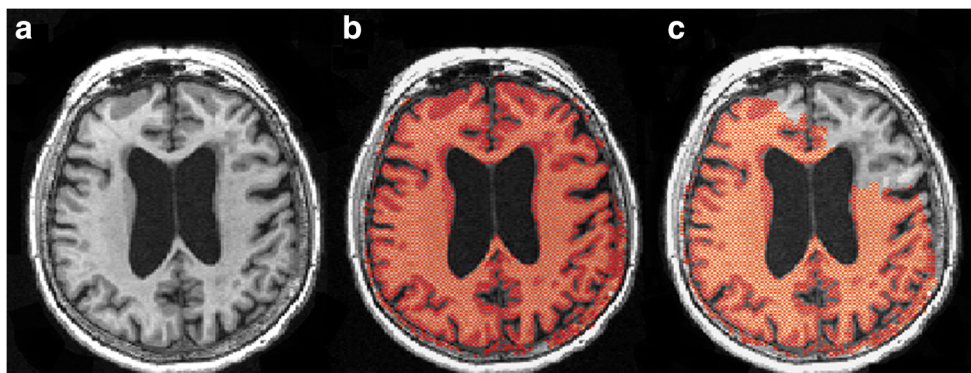
Only the images for which FFR was successful and for which the segmentation was successful are included
 ICC, intra-class correlation coefficient; *n*, number of subjects; AD, Alzheimer’s disease

Table 5 Lesion volume in the MS dataset and tumor volume in the glioblastoma dataset. From left to right, the table lists median [1st and 3rd quartiles] for volumes; Bonferroni-corrected *p* values for the pairwise comparison of volumes from FFR-processed and full images; volume differences between the full and FFR-processed images as mL and % differences (median [1st and 3rd quartiles]); [CC (absolute agreement (lower-upper band of 95% CI)) between volumes from full and FFR-processed images; Dice’s similarity index between segmentation from full and FFR-processed images; false negative between segmentations from full and FFR-processed images; and false positive between segmentations from full FFR-processed images

	MS; lesion (<i>n</i> = 55)			Glioblastoma; tumor (<i>n</i> = 66)			
	Volume (mL)	<i>p</i> value	Difference	ICC (lower-upper)	SI	FN (mL)	FP (mL)
			mL				
Full	2.71 [1.32; 7.76]						
QuickShear	2.94 [1.35; 8.12]	1.000	0.00 [-0.08; 0.04]	0.01 [-2.13; 1.08]	0.93 [0.86; 0.95]	0.27 [0.11; 0.76]	0.24 [0.11; 0.74]
FaceMasking	3.15 [1.60; 8.15]	<0.001	0.23 [-0.01; 1.26]	4.21 [-0.64; 25.58]	0.90 [0.72; 0.94]	0.24 [0.09; 0.66]	0.58 [0.12; 2.01]
Defacing	2.77 [1.21; 7.77]	0.197	0.00 [-0.13; 0.02]	-0.12 [-3.79; 0.83]	0.86 [0.39; 0.94]	0.59 [0.17; 1.35]	0.48 [0.20; 1.10]
Glioblastoma; tumor (<i>n</i> = 66)							
Volume (mL)				ICC	SI	FN (mL)	FP (mL)
Full	34.77 [19.54; 53.77]						
QuickShear	29.22 [16.74; 50.06]	<0.001	-2.46 [-7.08; -0.54]	0.843 (0.704–0.912)	0.87 [0.75; 0.92]	5.33 [2.70; 9.74]	1.89 [0.79; 3.80]
FaceMasking	31.65 [14.80; 51.13]	1.000	-1.31 [-7.74; 0.57]	0.312 (0.074–0.515)	0.86 [0.74; 0.92]	4.81 [2.00; 11.75]	2.19 [1.79; 4.77]
Defacing	28.47 [13.41; 49.21]	<0.001	-3.28 [-8.16; -0.72]	0.810 (0.560–0.901)	0.86 [0.74; 0.92]	5.94 [2.79; 9.75]	1.85 [0.79; 3.21]

Only the images for which FFR was successful and for which the segmentation was successful are included
 ICC, intra-class correlation; *n*, amount of subjects; MS, multiple sclerosis; SI, Dice’s similarity index; FN, false negative; FP, false positive

Fig. 3 An example of SIENAX affecting the BV by FFR processing showing 3D T1 images (a), 3D T1 images with the brain tissue segmentation shown in red (1.04 L) on the full image (b), and 3D T1 images with the brain tissue segmentation shown in red (0.87 L) on the FFR-processed image with Defacing (c)



NBV, SIENAX multiplies the BV (calculated in native subject space) by a volumetric scaling factor obtained from a linear registration of the brain image to a standard brain image, additionally using a derived skull image. FFR could affect the removal of non-brain tissue and identification of the skull, and

thereby cause a different registration result, culminating in altered NBV values. Differences in shapes of the head and face between people (related to, e.g., sex or ethnicity) may affect performance of standard FFR algorithms which may have ramifications especially for subsequent analysis methods

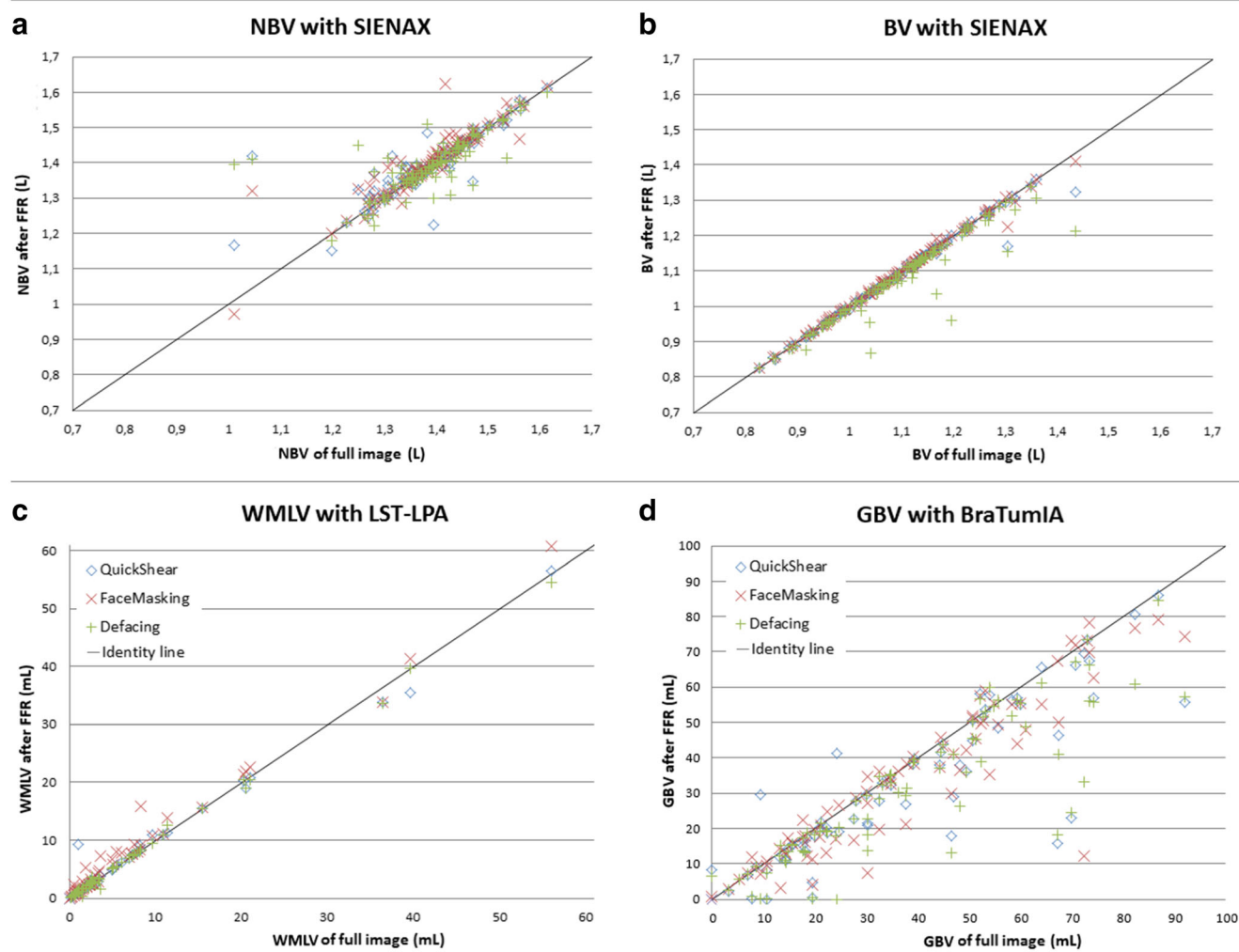
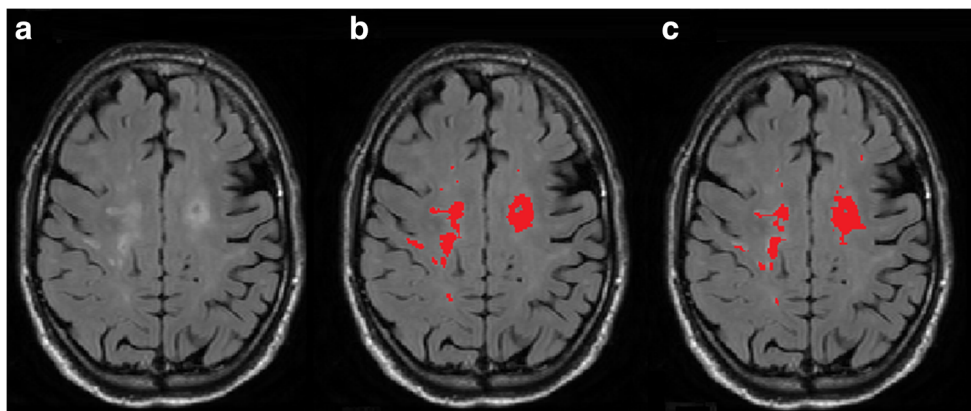


Fig. 4 Scatter plots of the normalized brain volume (a), brain volume (b), white matter lesion volume (c), and glioblastoma volume (d). The facial removal datasets are plotted against the full scan; QuickShear, blue diamond; FaceMasking, red cross; and Defacing, green plus sign. All

scatter plots have an identity line indicating perfect agreement. NBV, normalized brain volume; BV, brain volume; FFR, facial features removal; WMLV, white matter lesion volume; GBV, tumor volume; mL, milliliter; L, liter

Fig. 5 An example of lesion segmentation affected by FFR processing showing 2D FLAIR image (a), 2D FLAIR image with the lesion segmentation shown in red (8.30 mL) on full image (b), and 2D FLAIR image with the lesion segmentation shown in red (15.91 mL) on FFR-processed image with Defacing (c). Dice's similarity index between the complete 3D segmentations obtained from the full image and FFR-processed image was 0.48



that use the skull such as SIENAX. However, there are also cases in which the NBV was not affected by FFR, so maybe there is a cutoff on how much of the head can be removed without affecting the NBV measurement. In a further study, an optimum could be determined between the amount of facial features that should be removed for de-identification and the amount that should remain for correct analyses with automated methods.

Therefore, it would be interesting to study in more detail the effect of FFR methods on registration and other processing steps in a systematic way. Milchenko and Marcus [13] and Bischoff-Grethe et al [14] both addressed the effects of their FFR method on skull stripping; however, it would be interesting to analyze the effects of these methods on other processing steps as well as multiple skull stripping methods, all in the same dataset for an objective comparison. Next, to remedy those errors, analysis methods and processing steps should be made robust against the absence or distortion of facial features. As an example, facial features could be removed from fixed reference images in registration steps or in reference templates (e.g., of tissue probabilities) in image

processing pipelines. Moreover, it would also be helpful to study if changing the default settings of the automated methods improves the segmentation.

The measured volumes of the automated methods are affected—both random (low ICC) and systematic (significant differences). The random effects are mostly visible in the volume change of the NBV and GBV after FFR processing and the significant differences are visible in the volume change of BV and GBV. The Bland-Altman plots show that the volume changes are not dependent on the measured volumes. FFR affects not only the measured volumes of the automated methods but also the extent and precise spatial location of the WML and the glioblastomas, as demonstrated by the overlap analyses. For the WML in MS, median FP and FN fractions ranged between about 10 and 25% of the median total WMLV. Similar effects were observed for glioblastoma, with median FN fractions between about 15 and 20%, and median FP fractions between about 6 and 9%. Both the volumetric and spatial results indicate that the differences between full image segmentations and FFR-processed segmentations are substantial. This is unexpected, especially for LV and GBV, as given

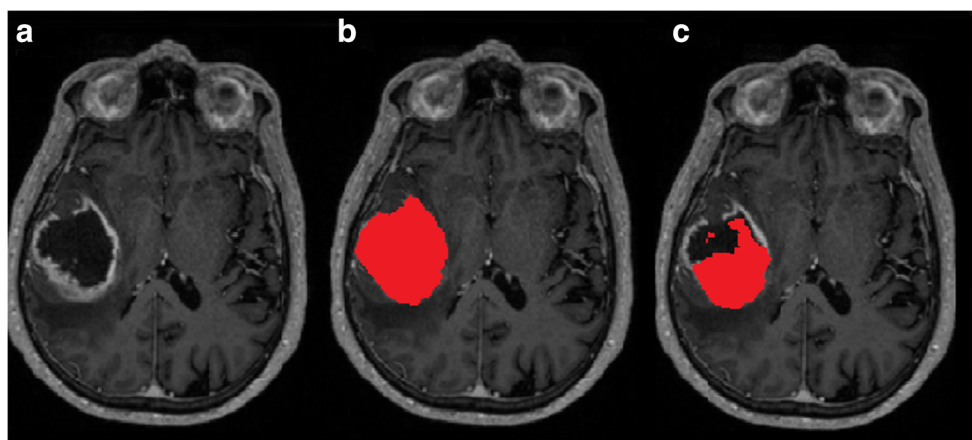


Fig. 6 An example of glioblastoma segmentation affected by FFR processing showing 3D T1 post contrast images (a), 3D T1 post contrast images with the glioblastoma segmentation shown in red (69.99 mL) on full image (b), and 3D T1 post contrast images with the

glioblastoma segmentation shown in red (22.87 mL) on FFR-processed image with QuickShear (c). Dice's similarity index between the complete 3D segmentations obtained from the full image and FFR-processed image was 0.48

that both the MS lesions and the glioblastoma are located within the region occupied by brain tissue that should not be, and judging from our visual inspections indeed was not, affected by the FFR methods. Both in MS and glioblastoma, the exact location and extent of pathological changes are of importance; therefore, these artifactual post FFR segmentation changes should be investigated in more detail and methods should be devised and tested to mitigate these effects.

Our results showed that the effects of FFR on current methods are a common problem across domains: brain volumes, MS lesion volumes, and glioblastoma volumes were all to some degree affected by FFR. The next step would be to investigate how to overcome such issues for SIENAX, LST-LPA, and BraTumIA, or in a broader sense, to study and mitigate sources of error after FFR for multiple methods aimed at brain volume, MS lesion, or glioblastoma segmentation. Another option would be to consider removing the facial features from fixed reference images in registration steps or in reference templates (e.g., of tissue probabilities) in image processing pipelines.

It should be noted that in this study, we did not test if the FFR methods indeed made the participant unrecognizable. However, we observed that the FFR methods in some cases seemed to leave parts of the face intact. We did not assess whether this made the person recognizable, because this would require a more rigorous setup outside the scope of this study. However, it would be important to establish guidelines on how to make participants unrecognizable, specifically which parts of the face should be removed or otherwise processed to ensure participants' privacy. Moreover, for protecting participants' privacy, it may be important to take into account that reconstruction of removed or deformed facial features may be possible [31].

In conclusion, this study highlighted a new challenge to the neuroimaging research community, which is to ensure high-quality neuroimaging research while protecting participants' privacy. Our results demonstrate that facial features removal of brain MRI can lead both to failure of automated analysis methods (mostly by LST-LPA and BraTumIA) and to changes in volumes obtained by the analysis methods, including both "random" variation (mostly by NBV and GBV) and systematic differences (mostly by BV and GBV). Therefore, volumetric image analysis methods need to be carefully assessed and optimized with regard to FFR methods, in order to ensure the reliability of clinical research outcomes while protecting participants' privacy in multicentric, collaborative studies. This could be done by improving image registration accuracy after FFR, addressing in more detail the effect of FFR methods on other processing steps or by developing methods that are tailored to images from which facial features have been removed.

Acknowledgments Part of the 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

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (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 (www.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.

MAGNIMS Study Group; MAGNIMS (Magnetic Resonance Imaging in MS) Study Group is a European study group of academics that share the interest in MS and imaging.

We want to thank Bob van Hoek that we could use the MR image of his head for the 3D render figure (Fig. 1).

Funding information The PICTURE project is supported by the program Innovative Medical Devices Initiative (IMDI) with project number 10-10400-96-14003, which is financed by the Netherlands Organization for Scientific Research (NWO). This research is also supported by a research grant from the Dutch Cancer Society (VU2014-7113).

The MS Center Amsterdam is supported by the Dutch MS Research Foundation through program grants (09-358d and 14-358e).

Frederik Barkhof and Olga Ciccarelli were supported by the NIHR Biomedical Research Centre at UCLH.

Compliance with ethical standards

Guarantor The scientific guarantor of this publication is Hugo Vrenken, h.vrenken@vumc.nl

Conflict of interest The authors of this manuscript declare relationships with the following companies:

A. de Sitter is employed on a project sponsored by a research grant from Teva Pharmaceuticals (grant to H. Vrenken and F. Barkhof).

M. Visser has nothing to disclose.

I. Brouwer is partly employed on a project sponsored by a research grant from Teva Pharmaceuticals (grant to H. Vrenken and F. Barkhof) and is partly employed on a project sponsored by a research grant from Novartis (grant to H. Vrenken and F. Barkhof).

K.S. Cover has nothing to disclose.

R.A. van Schijndel has nothing to disclose.

R.S. Eijgelaar has nothing to disclose.

D.M.J. Müller has nothing to disclose.

S. Ropele has nothing to disclose.

L. Kappos has served in the last 24 months as international or local principal investigator for the following drug studies: BOLD EXT., EXPAND (Siponimod, Novartis), DECIDE, DECIDE EXT. (Daclizumab HYP, Biogen), ENDORSE (DMF, Biogen), FINGORETT, FTY-UMBRELLA, INFORMS, INFORMS EXT LONGTERMS. (Fingolimod, Novartis), MOMENTUM (Amiselimod, Mitsubishi) OCRELIZUMAB PHASE II EXT., OPERA, ORATORIO and extensions (Ocrelizumab, Roche), REFLEXION (IFN β -1a, Merck), STRATA EXT, TOP (Natalizumab, Biogen), TERIFLUNOMIDE EXT, TERRIKIDS (Teriflunomide, Sanofi-Aventis), and ASCLEPIOS I/II (Ofatumumab, Novartis). L. Kappos is a member in the Editorial Boards of the following journals: “*Journal of Neurology*”, “*Multiple Sclerosis Journal*”, “*Neurology and Clinical Neuroscience*”, “*Multiple Sclerosis and Related Disorders*”, “*Clinical and Translational Neuroscience*”.

Á. Rovira serves on scientific advisory boards for Novartis, Sanofi-Genzyme, Icometrix, SyntheticMR, Bayer, Biogen, and OLEA Medical, and has received speaker honoraria from Bayer, Sanofi-Genzyme, Bracco, Merck Serono, Teva Pharmaceutical Industries Ltd., Novartis, Roche, and Biogen.

M. Filippi received speaker honoraria from Biogen Idec, Novartis, Genzyme, Sanofi-Aventis, Teva, Merck Serono, and Roche and receives research support from the Italian Ministry of Health and Fondazione Italiana Sclerosi Multipla.

C. Enzinger declares no conflict of interest.

J. Frederiksen has received no funding to support the presented work. She has served on scientific advisory boards for and received funding for travel related to these activities as well as honoraria from Biogen Idec, Merck Serono, Sanofi-Aventis, Teva, Novartis, and Almirall. She has received speaker honoraria from Biogen Idec, Teva, and Novartis. She has served as advisor on preclinical development for Takeda.

O. Ciccarelli has received research grants from the MS Society of Great Britain & Northern Ireland, National Institute for Health Research (NIHR) University College London Hospitals Biomedical Research Centre, EUH2020, Spinal Cord Research Foundation, and Rosetrees Trust. Professor Ciccarelli serves as a consultant for Novartis, Teva, and Roche and has received an honorarium from the AAN as Associate Editor of *Neurology* and serves on the Editorial Board of *Multiple Sclerosis Journal*.

C.R.G. Guttmann has received research grants from Sanofi and the National Multiple Sclerosis Society.

M.P. Wattjes has received speaker and/or consultancy fees from Biogen, Biologix, Bayer Healthcare, Celgene, Genilac, IXICO, Novartis, Merck Serono, Sanofi-Genzyme, Springer Healthcare, and Roche.

M.G. Witte has nothing to disclose.

P.C. de Witt Hamer has nothing to disclose.

F. Barkhof has received compensation for consulting services and/or speaking activities from Bayer Schering Pharma, Biogen Idec, Merck Serono, Novartis, Genzyme, Synthon BV, Roche, Teva, Jansen research and IXICO and is supported by the NIHR Biomedical Research Centre at UCLH.

H. Vrenken has received research grants from Novartis, Teva, Merck Serono, and Pfizer; consulting fees from Merck Serono; and speaker honoraria from Novartis; all funds were paid directly to his institution.

Statistics and biometry No complex statistical methods were necessary for this paper.

Informed consent Written informed consent was obtained for primary studies from all subjects (patients) in this study.

Ethical approval Institutional Review Board approval was obtained for primary studies.

Study subjects or cohorts overlap Some study subjects or cohorts have been previously reported in Wyman et al (2013) and de Sitter et al (2017). All patients of the paper from de Sitter et al (2017) were used and a subset of the data published in Wyman et al (2013). The difference in this study is that we use the dataset to study the effect of facial features removal, what is not done with this data yet.

Methodology

- prospective
- observational
- multicenter study

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

References

1. Gkoulalas-Divanis A, Grigorios L (2015) Medical data privacy handbook. Springer
2. Budin F, Zeng D, Ghosh A, Bullitt E (2008) Preventing facial recognition when rendering MR images of the head in three dimensions. *Med Image Anal* 12:229–239
3. Prior FW, Brunsten B, Hildebolt C et al (2009) Facial recognition from volume-rendered magnetic resonance imaging data. *IEEE Trans Inf Technol Biomed* 13:5–9
4. Parks CL, Monson KL (2017) Automated facial recognition of computed tomography-derived facial images: patient privacy implications. *J Digit Imaging* 30:204–214
5. Song X, Wang J, Wang A et al (2015) DeID - a data sharing tool for neuroimaging studies. *Front Neurosci* 9:325
6. Langer SG, Shih G, Nagy P, Landman BA (2018) Collaborative and reproducible research: goals, challenges, and strategies. *J Digit Imaging* 31:275–282
7. Holmes AJ, Hollinshead MO, O’Keefe TM et al (2015) Brain genomics superstruct project initial data release with structural, functional, and behavioral measures. *Sci Data* 2:150031
8. Liew SL, Anglin JM, Banks NW et al (2018) A large, open source dataset of stroke anatomical brain images and manual lesion segmentations. *Sci Data* 5:180011
9. Van Essen DC, Smith SM, Barch DM et al (2013) The WU-Minn Human Connectome Project: an overview. *Neuroimage* 80:62–79
10. Kushida CA, Nichols DA, Jadmicek R, Miller R, Walsh JK, Griffin K (2012) Strategies for de-identification and anonymization of electronic health record data for use in multicenter research studies. *Med Care* 50(Suppl):S82–S101
11. Marcus DS, Wang TH, Parker J, Csernansky JG, Morris JC, Buckner RL (2007) Open Access Series of Imaging Studies (OASIS): cross-sectional MRI data in young, middle aged, nondemented, and demented older adults. *J Cogn Neurosci* 19: 1498–1507
12. Schimke N, Hale J (2011) Quickshear defacing for neuroimages. Proceedings of the 2nd USENIX conference on Health security and privacy USENIX Association
13. Milchenko M, Marcus D (2013) Obscuring surface anatomy in volumetric imaging data. *Neuroinformatics* 11:65–75
14. 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

15. Wyman BT, Harvey DJ, Crawford K et al (2013) Standardization of analysis sets for reporting results from ADNI MRI data. *Alzheimer's Dement* 9:332–337
16. de Sitter A, Steenwijk MD, Ruet A et al (2017) Performance of five research-domain automated WM lesion segmentation methods in a multi-center MS study. *Neuroimage* 163:106–114
17. Cover KS, van Schijndel RA, Versteeg A et al (2016) Reproducibility of hippocampal atrophy rates measured with manual, FreeSurfer, AdaBoost, FSL/FIRST and the MAPS-HBSI methods in Alzheimer's disease. *Psychiatry Res Neuroimaging* 252:26–35
18. Ropele S, Kilsdonk ID, Wattjes MP et al (2014) Determinants of iron accumulation in deep grey matter of multiple sclerosis patients. *Mult Scler* 20:1692–1698
19. Kurtzke JF (1983) Rating neurologic impairment in multiple sclerosis: an Expanded Disability Status Scale (EDSS). *Neurology* 33:1444–1452
20. Jenkinson M, Bannister P, Brady M, Smith S (2002) Improved optimization for the robust and accurate linear registration and motion correction of brain images. *Neuroimage* 17:825–841
21. Cormen TH, Leiserson CE, Rivest RL, Stein C (2007) Introduction to algorithms, second edition. The MIT Press, Cambridge, London
22. Andrew AM (1979) Another efficient algorithm for convex hulls in two dimensions. *Inf Process Lett* 9:216–219
23. Smith SM (2002) Fast robust automated brain extraction. *Hum Brain Mapp* 17:143–155
24. Popescu V, Battaglini M, Hoogstrate WS et al (2012) Optimizing parameter choice for FSL-Brain Extraction Tool (BET) on 3D T1 images in multiple sclerosis. *Neuroimage* 61:1484–1494
25. Schmidt P (2017) Bayesian inference for structured additive regression models for large-scale problems with applications to medical imaging. PhD thesis, LudwigMaximilians-Universität München. Available via <http://nbn-resolvingde/urn:nbn:de:bvb:19-203731>
26. Bauer S, Fejes T, Slotboom J, Wiest R, Nolte LP, Reyes M (2012) Segmentation of brain tumor images based on integrated hierarchical classification and regularization. *Proceedings of MICCAI BraTS Work*, pp 10–13
27. Weir JP (2005) Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. *J Strength Cond Res* 19:231–240
28. Cicchetti DV (1994) Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychol Assess* 6:284–290
29. Dice LR (1954) Measures of the amount of ecologic association between species. *Ecology* 26:297–302
30. Dadar M, Fonov VS, Collins DL, Alzheimer's Disease Neuroimaging Initiative (2018) A comparison of publicly available linear MRI stereotaxic registration techniques. *Neuroimage* 174:191–200
31. Abramian D, Eklund A (2018) Refacing: reconstructing anonymized facial features using GANs. *arXiv preprint arXiv:1810.06455*

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