NeuroImage 231 (2021) 117845

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/neuroimage

# Changing the face of neuroimaging research: Comparing a new MRI de-facing technique with popular alternatives

Christopher G. Schwarz[a,*], Walter K. Kremers[b], Heather J. Wiste[b], Jeffrey L. Gunter[a,c], Prashanthi Vemuri[a], Anthony J. Spychalla[a], Kejal Kantarci[a], Aaron P. Schultz[d], Reisa A. Sperling[d], David S. Knopman[e], Ronald C. Petersen[e], Clifford R. Jack Jr.[a], the Alzheimer's Disease Neuroimaging Initiative[†]

[a] Department of Radiology, Mayo Clinic, Rochester, MN, United States
[b] Department of Health Sciences Research, Mayo Clinic, Rochester, MN, United States
[c] Department of Information Technology, Mayo Clinic, Rochester, MN, United States
[d] Department of Neurology, Massachusetts General Hospital, Harvard Medical School, Boston, MA, United States
[e] Department of Neurology, Mayo Clinic, Rochester, MN, United States

ARTICLE INFO

*Keywords:*
Face Recognition
De-Facing
De-Identification
Anonymization
Reliability

ABSTRACT

Recent advances in automated face recognition algorithms have increased the risk that de-identified research MRI scans may be re-identifiable by matching them to identified photographs using face recognition. A variety of software exist to de-face (remove faces from) MRI, but their ability to prevent face recognition has never been measured and their image modifications can alter automated brain measurements. In this study, we compared three popular de-facing techniques and introduce our *mri_reface* technique designed to minimize effects on brain measurements by replacing the face with a population average, rather than removing it. For each technique, we measured 1) how well it prevented automated face recognition (i.e. effects on exceptionally-motivated individuals) and 2) how it altered brain measurements from SPM12, FreeSurfer, and FSL (i.e. effects on the average user of de-identified data). Before de-facing, 97% of scans from a sample of 157 volunteers were correctly matched to photographs using automated face recognition. After de-facing with popular software, 28-38% of scans still retained enough data for successful automated face matching. Our proposed *mri_reface* had similar performance with the best existing method (*fsl_deface*) at preventing face recognition (28-30%) and it had the smallest effects on brain measurements in more pipelines than any other, but these differences were modest.

## 1. Introduction

It has long been hypothesized that de-identified brain images may potentially be re-identified by reconstructing the participant's face from the scan and applying face recognition. Consequently, many algorithmic techniques have been developed for removing or distorting face imagery ("de-facing") to prevent face recognition (Alfaro-Almagro et al., 2018; Bischoff-Grethe et al., 2007; Fonov and Collins, 2018; Gulban et al., n.d.; Hanke, 2015; Milchenko and Marcus, 2013; Schimke and Hale, 2011; Silva et al., 2018). A 2009 study tested whether human volunteers could successfully match participant photographs to the correct MRI-based face reconstruction, and only 40% of volunteers performed the matching with success rates greater than random chance (Prior et al., 2009). A 2012 study was the first to test automatic face recognition software (Google Picasa, launched in 2009), finding that 27.5% of CT-based reconstructions could be matched to the correct participant photographs (Mazura et al., 2012). However, adoption of de-facing software across research neuroimaging studies has been mixed, and many large studies and data-sets have widely distributed images without attempting to prevent face recognition. Studies may have chosen to avoid de-facing for multiple reasons: 1) desire to share minimally-altered data to maximize potential scientific utility; 2) concern that de-facing techniques may reduce the quality of measurements obtained from the images; 3) belief that researchers receiving the data would act in good faith and honor Data Use Agreements they signed in order to receive the data; or

* Correspondence to: Mayo Clinic, Diagnostic Radiology, 200 First Street SW, Rochester, Minnesota 55905, United States.
*E-mail address:* schwarz.christopher@mayo.edu (C.G. Schwarz).

† A portion of 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.

https://doi.org/10.1016/j.neuroimage.2021.117845
Received 11 August 2020; Received in revised form 1 December 2020; Accepted 5 February 2021
Available online 11 February 2021

4) belief that the threat of loss of privacy via face recognition was small enough to be an acceptable level of risk for their scientific potential.

Between 2013 and 2018, identification performance of face recognition algorithms (for their designed purpose of matching photographs to photographs) improved by approximately 20x industry-wide as older algorithms were replaced by convolutional neural networks and deep learning (Grother et al., 2018). Our recent study found that when applying more-recent face recognition algorithms, photos of 70/84 (83%) of participants were matched to the correct MRI (Schwarz et al., 2019), demonstrating that face recognition now poses a much greater risk to the privacy of research participants than previously estimated. These findings have led to increased concern that a motivated individual could violate research agreements by: 1) requesting access to de-identified MRI via data sharing, 2) reconstructing participants' faces, and 3) matching these to public photographs to re-identify participants. This re-identification would also re-identify all protected health information (PHI) released by the study about each participant, such as diagnoses, genetic information, neuropsychiatric measures, family/personal history, etc.

## 2. Study design

An ideal technique for de-facing images should: 1) minimize the risk of face recognition; and 2) *not* significantly alter automated brain measurements on the de-faced images. Other recent works have suggested that some popular de-facing techniques may provide inadequate protection (Abramian and Eklund, 2019), and may substantially alter or impede automated brain measurements (de Sitter et al., 2020). In this study, we compare several popular de-facing techniques and introduce our *mri_reface*, which we designed to minimize effects on brain measurements by replacing the face rather than removing it. We believe both of these criteria are equally important, but their consequences affect very discrete sets of people and therefore we first evaluate them separately and in very different contexts before we compare de-facing techniques according to the combined findings. We used 3D Fluid-Attenuated Inversion Recovery (FLAIR) scans for testing criterion 1 because these give the best facial reconstructions (see supplementary material), but in contrast we used T1-weighted scans for criteria 2 because these are the most commonly analyzed images from public data sets.

### 2.1. Validation criterion 1: Protection from face recognition

The hypothetical worst-case scenario caused by re-identification of research participants would be if a highly-skilled and highly-motivated individual were to re-identify and extort research participants using their sensitive protected health information. Regardless of whether they targeted a single high-profile participant or large numbers of average participants, public disclosure of such an event would cause a devastating loss of public trust and participation in medical research, as vulnerable individuals would not trust research scientists to keep their participation and health information confidential. Therefore, we designed Validation Criterion 1 to compare techniques according to how well they could potentially prevent re-identification by a motivated individual with extensive skills and knowledge of MR image processing. Our design tests the scenario where a motivated individual has reason to believe that someone is part of a study, and they attempt to identify that participant within the study's de-identified FLAIR image data.

### 2.2. Validation criterion 2: Minimizing effects on brain measurements

The overwhelming majority of people who gain access to de-identified research data will have no motivation to re-identify research participants; they will only use the data for its intended scientific purposes in accordance with applicable data use agreements. For most users, this means downloading de-identified T1-weighted images from a large public data-set (e.g. Alzheimer's Disease Neuroimaging Initiative (ADNI)), running standard popular software pipelines, and analyzing resulting numeric data. This large majority of users are not affected by whether the data can be re-identified but only by whether the de-identification process hinders their analyses. Therefore, we designed Validation Criteria 2 to compare and quantify how each de-identification technique affects these standard analyses by the average user.

## 3. Materials and methods

### 3.1. Standard de-facing techniques

We include three of what we believe to be the most popular de-facing software in our comparison: *mri_deface* (FreeSurfer) (Bischoff-Grethe et al., 2007), *pydeface* (Gulban et al., n.d.), and *fsl_deface* (Alfaro-Almagro et al., 2018). Although it is also relatively popular, we did not include *mask_face* (Milchenko and Marcus, 2013) because previous work has already demonstrated that it provides inadequate protection (Abramian and Eklund, 2019). We provide a comparative example of all tested methods in Fig. 1.

**mri_deface:** *mri_deface* is a program included with FreeSurfer that locates face voxels using linear registration and a pre-defined mask created from manual labeling of scans from 10 subjects. It was designed for use with T1-w images. Face voxels are removed (set to zero intensity) only if they are not within 7mm of estimated brain tissue (Bischoff-Grethe et al., 2007). We used *mri_deface* version 1.22 (the current release), with default settings.

**pydeface:** *pydeface* (Gulban et al., n.d.) was initially released in 2017 and has become popular among the python neuroimaging community. Like *mri_deface*, it includes its own pre-defined mask of face voxels which are located on the input image using linear registration (FSL's linear registration tool (FLIRT) (Jenkinson et al., 2002)) and removed (set to zero). We used the un-named version automatically installed through the python package manager (pip) on December 6, 2019, with default settings (only image input/output name were specified). Its documentation does not specify what types of MRI are supported.

**fsl_deface:** The UK Biobank study uses a customized image processing pipeline based on FSL (Alfaro-Almagro et al., 2018), which includes a de-facing approach also based on FSL tools. It was designed for use with T1-w images. This de-facing approach was later extracted from the larger processing pipeline and released as part of the main FSL package as *fsl_deface*. Like *mri_deface* and *pydeface*, this method uses linear registration (also FLIRT) to locate its own pre-defined mask of face voxels on the target image, then sets voxels in the mask to zero. Unlike *mri_deface* and *pydeface,* this method also removes the ears. We used *fsl_deface* as included in FSL version 6.0.3 (the current version), with default settings (only image input/output name were specified).

### 3.2. Proposed face-replacement technique: mri_reface

**Method overview:** We also propose and compare our in-house de-facing technique, *mri_reface*: rather than removing or blurring the face voxels, we replace them (i.e. perform a digital face "transplant") with voxels from a population-average face. The goal of this approach was to provide increased protection from re-identification while minimizing effects on brain measurements by generating de-faced images that better resemble the natural images that each measurement pipeline was designed for. Another notable detail vs the other compared methods is that our implementation is based on nonlinear registration rather than only linear, allowing for more precision in localizing face regions to ensure their accurate removal without altering brain regions. We also replace the ears, to prevent ear recognition (Emeršič et al., 2017), and regions of air that may contain image artifacts with identifiable features. Our current implementation supports T1-, T2-, and FLAIR-weighted images. We detail our approach below.

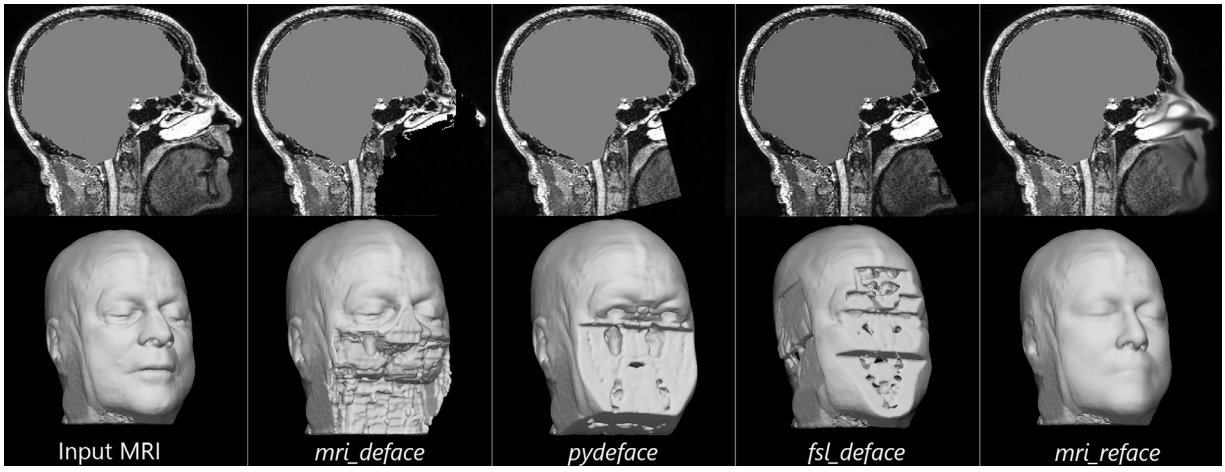


**Fig. 1.** Comparison of tested de-facing techniques on an input 3D FLAIR scan. Top: sagittal MRI slice (brain is omitted for participant privacy) Bottom: corresponding face reconstruction. Note that *mri_deface* retained the eyes and part of the nose. Our reconstruction process removes floating disconnected voxels, so the remaining nose is not visible on the corresponding render. *Pydeface* retained the top of the eyes. Among the three standard methods, only *fsl_deface* removed the ears, and entirely removed the eyes. In our proposed *mri_reface*, all face regions and ear regions were replaced with an average face and ears. This volunteer consented to allow publication of their photographs and corresponding MRI-based reconstructions for illustration purposes.

**Fig. 2.** Population-average MRI and de-identification mask. Left: Our population-average T1-weighted MRI template Center: Reconstruction of the population-average face from the template Right: De-identification mask (overlaid) of face, ear, and behind-head (red, yellow, and orange respectively) voxels to be replaced.

**Creation of population-average face templates:** First, we created a population-average template for each MRI contrast, using scans of 177 Mayo Clinic participants age 30-89 (stratified by age-decade and sex; 120 cognitively unimpaired and 57 with clinically-diagnosed Alzheimer's disease), imaged using 3D T1-weighted MPRAGE sequences on Siemens Prisma scanners. We co-registered the scans using an unbiased, group-wise approach (Avants et al., 2010) with high-dimensional symmetric normalization (Avants et al., 2008), and we normalized their intensities prior to voxel-wise averaging. In brief, a voxel-wise mean was computed from all 177 MRI, then each MRI was non-linearly deformed (its geometry was locally compressed and expanded) to match the mean image as closely as possible. The process was then repeated for several iterations until convergence (i.e. the average is no longer blurry). Finally, we used ANTs to transform the new template to the space of our Mayo Clinic Adult Lifespan Template (MCALT) (Schwarz et al., 2017a), to match our in-house atlases. This process provides an unbiased average MRI (Fig. 2, Left), which includes a geometrically average face (Fig. 2, Center), from the sampled population. On this average MRI, we then manually traced our de-identification mask: a mask of image voxels potentially containing face and ear structural information (Fig. 2, Right). We also created average T2-weighted and FLAIR templates from the same imaging sessions by linearly registering each of these images to their corresponding T1-weighted MRI and transforming them through the linear and nonlinear parameters computed above from the T1-weighted scans.

**Transforming between input image and template:** (Fig. 3 A-C) Given a target MRI to de-identify, our method performs ANTs symmetric non-linear registration (Avants et al., 2008) between the input image and the population-average template image, in order to transform the template to match the input image and identify all face/ear voxels. The input image modality is specified by the user, and the matching template is used from the included library. The registration uses a mask (*–warp-mask* option) in the template space that includes only voxels that are part of the head with those in the face and ear masks removed. Performing the registration using this mask calculates parameters that use only linear registration in the air, face, and ears, but use nonlinear registration in the rest of the head and brain. We then apply these parameters to transform and resample the template to the space of the original input image, which leaves the template face un-warped (aligned; modified only linearly) while warping the rest of the template image. At this point, the edges of the transformed face/ear regions align with those of the target image, but within those regions, the contour is that of the original average template, i.e., a population-average face rather than the face in the target image.

**Intensity normalization:** (Fig. 3 D) Before copying face voxels from the template image to the target image, the transformed template image must be intensity-normalized to match the target image. First, global intensity normalization is performed by a linear transform with the difference between air voxels for the intercept and the ratio between tissue voxels (white matter for T1-weighted images and gray matter for T2-weighted or FLAIR images) for the slope. Because this global intensity normalization alone would not account for local image intensity variations (field inhomogeneity), we apply differential bias correction (DBC) for smooth local intensity normalization between the images (Vemuri et al., 2015). To perform DBC, both images are smoothed with a Gaussian 12mm FWHM isotropic Gaussian kernel and sampled on an isotropic grid of approximately 10mm, omitting points where the difference between the images is large enough (differ by > 50%) to suggest that they likely do not contain analogous tissue. We compute the ratio of the images at each remaining voxel and use the *scatteredInterpolant* function in Matlab to interpolate between sampled points. Finally, we smooth the interpolated field with an 8mm FWHM Gaussian kernel and multiply the transformed template by the result to produce an image that is intensity-normalized to match the input image both globally and locally.

**Face and ear replacement:** We then de-identify the target image by replacing all voxels in the face and ear regions (Fig. 3 E) with those from

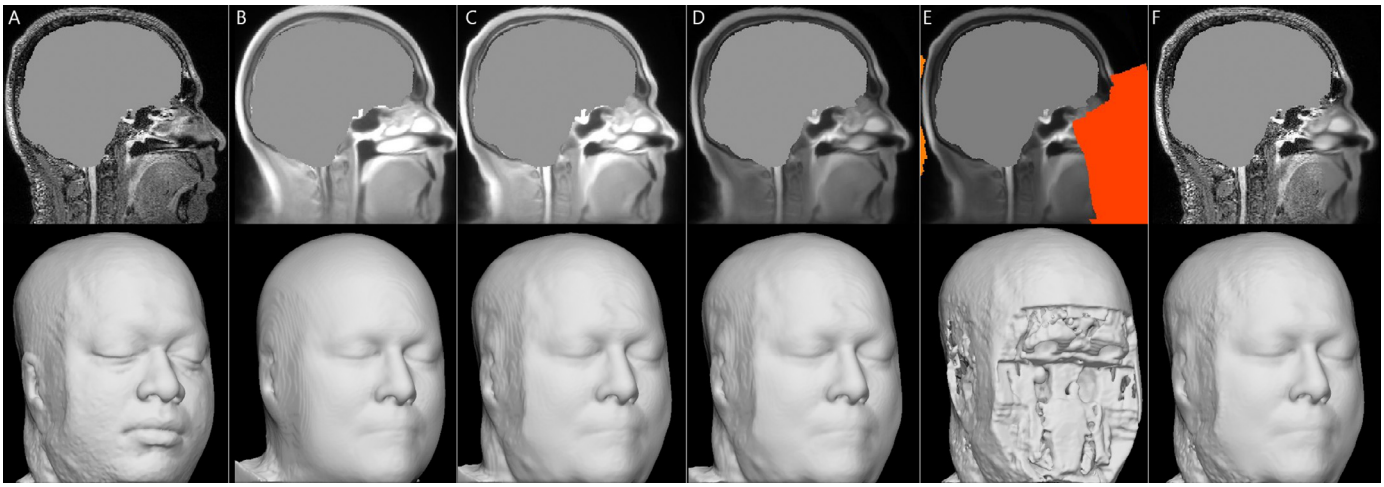


**Fig. 3.** Steps in our proposed face replacement (*mri_reface*) approach: Top: MRI voxel slice Bottom: Image of face reconstructed from above image. A) Input image (brain is omitted for participant privacy) B) Template co-registered (affine) to input image C) Template warped (nonlinear) to input image (only affine transformation in face/ear regions) D) Image C after DBC intensity normalization E) Mask of regions to be replaced F) Output image, a blend of images A and D as defined by E. This volunteer consented to allow publication of their photographs and corresponding MRI-based reconstructions for illustration purposes.

the transformed template, thereby "transplanting" the template face and ears onto the target. The binary mask of face and ear regions to replace is smoothed (8mm FWHM Gaussian kernel), and the spatial transition between original and replaced parts of the image is interpolated according to this smoothed mask (Fig. 3 F). All voxels within the mask of TIV (as defined in the template space) are never altered, regardless of their proximity to the face or the smoothed/interpolated transitional area.

**Removal of identifiable artifacts in air:** MRI voxels in front of the face and behind the head may contain identifiable information. For example, participant movement of the eyes/eyelids (more common) or nose/mouth (less common) during the scan can cause faint aliases of their contours in front of the face. To address this, we first compute a robust mean of air voxels in the input image, as identified by a mask in the template space, and replace air voxels >10x the mean (i.e. brighter than the surrounding air and may contain identifiable artifacts) with the corresponding voxels from the template. Relatedly, if the image field of view is too large to contain the nose/mouth/chin, these features are sometimes aliased into the air behind the head. Such locations are not typically modified by other de-facing methods, but they could be automatically un-wrapped and re-attached when generating MRI-based face reconstructions (as we do for face reconstructions in this work). To prevent this, we calculate the median of non-zero voxels in a mask-defined region of air voxels behind the head, then replace all voxels with intensity >2x the median with intensities from a random normal distribution.

### 3.3. Validation criterion 1: Protection from face recognition

We assessed the performance of each de-facing method according to two distinct criteria. The first was the relative ability of each de-facing method to prevent re-identification of research participants if attempted by an exceptionally skilled and motivated individual. We believe that such an individual would have the skills needed to attempt reconstructing faces from images with partially-removed faces, and thus we compared de-facing methods by their ability to prevent re-identification in this scenario.

#### 3.3.1. Validation data-set

This data-set included 84 individuals from our previous face recognition study (Schwarz et al., 2019), as well as 73 additional volunteers from continuing recruitment. In total, we recruited 157 volunteers (ages 34-93; mean=63.0, SD=16.3) stratified by sex and age-decade who had an existing brain MRI (3D FLAIR sequence) within the previous six months as part of their existing enrollment in the Mayo Clinic Study of Aging (MCSA (Petersen et al., 2010; Roberts et al., 2008)) or Alzheimer's Disease Research Center studies. All participants provided informed consent for this specific study, which was approved by the Mayo Clinic Institutional Review Board.

We photographed each individual's faces under indoor lighting conditions using standard iPads (Apple Inc., Cupertino, CA; models Air 2 and 6$^{th}$ generation). Participants were instructed to look directly at the camera, and then approximately 10 degrees up, down, left, and right, for a total of five photos, designed to provide somewhat-unique photos of each individual suitable for face recognition with minimal participant burden. Photos were manually cropped loosely around the head and converted to grayscale to better match MRI (which does not capture color). Our image cropping retained the head/hair/ears and removed only distant background and torso, in order to reduce unnecessary image size and speed up repeated image uploading during testing.

Sagittal 3D FLAIR head MRI were acquired using Siemens Prisma scanners using standard protocols matching those from ADNI3: resolution 1.0×1.0×1.2mm, repetition time=4800ms, echo time=441ms, and inversion time=1650ms. We used 3D FLAIR (rather than 3D T1-weighted) scans because these sequences provided more-recognizable face reconstructions (see supplementary material).

#### 3.3.2. Validation methods

**Standard Face Reconstructions**: We refer to generating a synthetic image of a face from an MRI scan as "face reconstruction". For our "standard" face reconstructions, we used the same process as previously described (Schwarz et al., 2019). In brief, a threshold was automatically chosen to binarize each image based on Otsu's method (Otsu, 1979), aliased nose parts behind the head were automatically detected and reattached to the face for applicable images, and any remaining, floating regions disconnected from the head (e.g. motion artifacts) were removed. Finally, each volume was converted to a surface using an automated threshold based on and the *nii_nii2gii* utility provided with *surf_ice* (Rorden, n.d.). For each reconstruction, 81 2D render images (analogous to a synthetic photograph) were automatically generated using *surf_ice* with the "Phong_Matte" shader under a variety of simulated lighting and viewing angles and saved as .png files (Schwarz et al., 2019). A random subset of 10 of these render images (chosen consistently across methods) was selected for each scan and used to train the face recognition software to recognize each MRI-based face reconstruction.

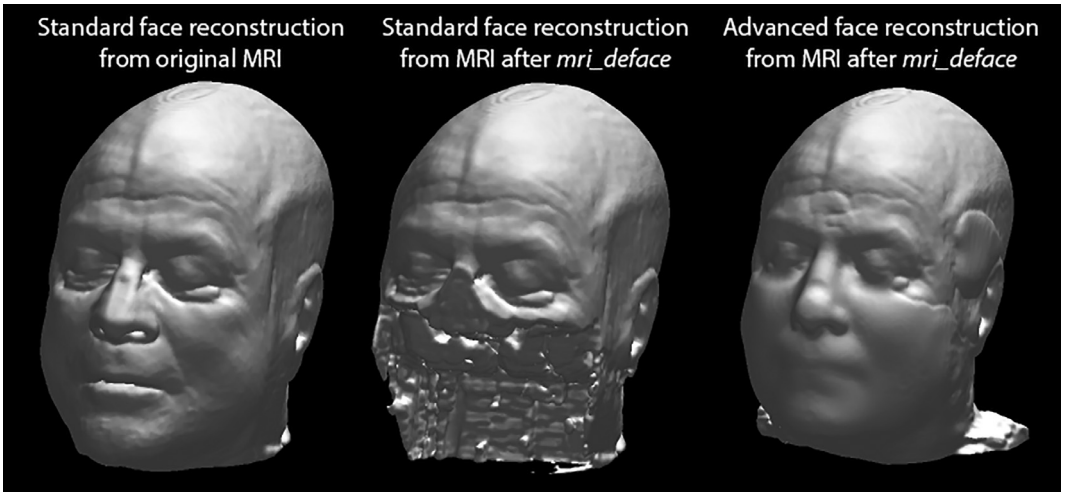


**Fig. 4.** Left: For images where faces have not been removed, we used our "standard" face reconstruction with minimal preprocessing. Center: *mri_deface and py_deface* frequently (but not always) retain the eyes, but the removed nose/mouth can prevent testing automated face recognition because no face is detected. Right: To test whether a highly skilled and motivated individual could perform automated face recognition using only the remaining facial features, we applied our "advanced" face reconstruction, where we filled-in missing regions with those from the average template.

**Advanced Face Reconstructions (for de-faced scans with partial faces):** For images where faces have been removed by de-face techniques, our "standard" face reconstructions (Fig. 4, left) produce images without faces or with only partial faces, which can prevent detection of a face by automatic algorithms (Fig. 4, center). However, we noticed that *mri_deface* and *pydeface* frequently (but not always) retain the eyes. Because some face recognition algorithms can attain good match performance with only partial faces (Elmahmudi and Ugail, 2019), we hypothesized that these de-faced images may still contain enough information for a skilled individual to perform partially-effective automatic face recognition. To test this hypothesis, we created our "advanced" face reconstructions as follows. We used ANTs (Avants et al., 2008) to warp a template MRI (the same as the average faces used in our own approach) to each de-faced scan, performed global intensity normalization, and replaced all missing voxels (those that were subthreshold in the input de-faced image but super-threshold in the template image) with those from the template. This replaced all removed voxels with those from the average face, allowing us to then use our standard method to create a face reconstruction containing a composite of: 1) the image regions that were not removed, and 2) the average face (in the areas that were removed) (Fig. 4, right). We used this process only when assessing Criterion 1 (preventing face recognition) with de-facing methods that remove face regions.

**Face Recognition Testing:** We used the same process for testing face recognition detailed in our previous publication (Schwarz et al., 2019), with the exception that we now use Microsoft Azure's recently updated *recognition_02* model with higher accuracy. In brief, we used the Microsoft Azure Cognitive Face API (Microsoft Corporation, 2019), which is designed to match an input face photo with one of a user-defined set of possible faces. This is analogous to a digital police "line-up", where the software is trained to recognize a "training set" of faces, and then a new "test image" is input with the question "which of the training faces does it best match?" The details of how the software works are unpublished proprietary technology using pre-trained models for face detection and encoding, and it operates as a cloud-based service that is available to the public. We trained an instance of Azure's "PersonGroup" classifier to recognize each participant based only on their MRI-based face reconstructions (i.e. the MRI faces comprised the "line-up" of potential faces to be recognized). Then, we input each of the five photographs for each participant and queried the software using the "FaceIdentify" function. For each photograph, Azure returned a ranked list of the 50 best matches (with match confidence scores for each) from among the MRI-based face reconstructions in the training set. We summed the match confidence scores for each potential match across each participant's five photographs and ranked each candidate MRI according to these summed scores. The resulting ranking reflected the combined set of five photos for each participant.

### 3.4. Validation criterion 2: Minimizing effects on brain measurements

Our second criterion for assessing de-facing methods was measuring how each affects the ability of an average data user to generate and compare brain measurements using SPM12, FreeSurfer, and FSL. We considered the results of each pipeline on the unmodified image to be the gold standard, and we compared the de-facing methods according to how much each pipeline's measurements on the de-faced images deviated from these original measurements. Aside from de-facing, no other pre-processing was performed on images before inputting them to each pipeline. This criterion was designed to reflect the experience of the typical downloader of shared de-identified images from public data sets (e.g. ADNI) who would run standard measurement pipelines on the de-identified (de-faced) data. We used ADNI data for this criterion because it is one of the largest public repositories of neuroimaging data and because it allowed us to construct a large data-set with a balanced set of scans from multiple MRI vendors.

#### 3.4.1. Validation data-set

We constructed a data-set of 300 3-Tesla T1-weighted accelerated MRI scans from the ADNI database (adni.loni.usc.edu). For up-to-date information, see the ADNI homepage ("ADNI Home," 2013). One hundred scans were included from each of the three MRI vendors (GE, Siemens, and Philips). Among each set of 100 within-vendor scans, 50 participants were cognitively unimpaired (CU) and 50 participants (individually sex- and age-matched within 3 years) had clinically-diagnosed Alzheimer's disease. Details of ADNI image acquisition parameters have been previously published (Jack et al., 2010). We used T1-weighted scans because these are what the most popular analysis pipelines are designed to analyze.

#### 3.4.2. Software measurement pipelines

We compare the effects of each de-facing method on three different software pipelines, each described below.

**SPM12:** We constructed this automated pipeline to represent a traditional GM volumes analysis with Statistical Parametric Mapping version

12 (SPM12) and default settings. For each input T1-weighted image, Unified Segmentation (Ashburner and Friston, 2005) was performed as implemented in SPM12 ("Segment" function) with the default priors and settings. The deformation parameters produced by Unified Segmentation were used to resample the atlas from MNI space into the space of the input image using nearest-neighbor interpolation. Per-region GM volumes were measured in subject space as the sum of the estimated GM probabilities in each voxel within the ROI, multiplied by the per-voxel volume. To produce per-region GM volume measurements, we used the Neuromorphometrics atlas included with SPM12 (omitting regions that are primarily WM or CSF). Total volume measurements were produced from segmentation outputs by using the included "Tissue Volumes" utility to sum total GM, WM, and CSF volumes. We summed GM+WM for total tissue volume (TTV) and GW+WM+CSF for total intracranial volume (TIV).

**FreeSurfer:** We used FreeSurfer (Fischl, 2012) version 6.0 in the cross-sectional stream. We ran *recon-all* using the flags *"-all -3T -notal-check -no-isrunning*" and analyzed regional GM volume, cortical thickness, TTV, and TIV measurements from the *aparc* and *aseg* outputs (omitting non-GM regions). We used *EstimatedTotalIntraCranialVol* for TIV and *BrainSegVolNotVent* for TTV (even though this excludes brainstem and cerebellum, it is used analogously and we are not comparing values across pipelines).

**FSL-UKBB:** To measure the effects of de-facing on FMRIB Software Library (FSL) tools, we included the UK Biobank Pipeline version 1 (Alfaro-Almagro et al., 2018), which uses FSL tools and was written with input from several primary authors of FSL. In their standard pipeline, segmentations are performed on images after applying the internal de-facing, but from *bb_struct_init* we separated the portions of code related to de-facing (described above) from the portions related to image segmentation (described here) so that we could test each independently and in combination with the other methods. We also removed the gradient distortion correction step, as our images have this already applied during preprocessing. This pipeline does not produce per-region GM volume measurements analogous to the other tested pipelines, so to create comparable measurements we transformed the HarvardOxford cortical and subcortical atlases (Desikan et al., 2006) included with FSL, using the same FNIRT (Andersson et al., 2008) nonlinear registration parameters the pipeline previously computed, and over each region we summed the per-voxel values previously computed by FAST (Zhang et al., 2001), then multiplied by the per-voxel volume. We used outputs from *SienaX* (Smith et al., 2002) for TTV ("BRAIN") and TIV ("VSCALING"). *SienaX* does not directly produce a TIV as a volume, but it provides the *VSCALING* ratio to be used for scaling tissue volume measurements; since that is arguably the primary use of TIV measurements, we analyzed this value in place of TIV (we performed all comparisons within each pipeline, so the difference in units or scale across pipelines is not an issue).

#### 3.4.3. Validation methods

The 300 images in the validation data-set were used as input to each of the three pipelines above, to produce per-region measurements of gray matter volumes and (where applicable) cortical thickness. We then input each image through each pipeline *after* using each de-facing method (without any other pre-processing). The per-region numeric values produced for the unmodified images were treated as the gold standard, and we measured systematic (bias) and non-systematic (noise) deviations from these values when measured from the de-faced images as described below.

**Quality Control (QC):** Trained medical image analysts visually examined the output segmentations produced by each of the three processing pipelines for each of the 300 unmodified scans. For any combination of scan+pipeline (900 total) where numeric results for unmodified images were not produced (missing, zero, or NaN) or where analysts judged the visual segmentation or atlas images to be gross failures, that scan was removed from analyses of the effects of de-facing on that particular pipeline.

**Statistical methods:** We measured the differences between per-region brain measurements produced by each pipeline for each image before vs. after each de-facing method. Analyses used *R* statistical software (R Development Core Team, 2008) version 3.6.2 with *tidyverse* packages (Wickham, 2017). For each combination of measurement pipeline, output regional brain measurement, and de-facing method, we measured a) non-systematic error (noise) with intra-class correlation coefficient (ICC) and b) systematic error (bias) across the 300 measurements from each scan before vs. after de-facing. For ICC, we used the *ICC* function from the R *pysch* package (Revelle, 2019) to calculate the fixed-raters *ICC3* variant that is not sensitive to differences in means between raters (i.e. is not sensitive to systematic error). We then separately measured the systematic error (bias) of the de-faced image measurements as the percent difference between the x=y line and a linear least-squares fit (*lm* function) of the original vs. de-faced measurements, taken at the "centercept" point (mean value across the x axis, i.e. all measurements from the unmodified image) (Wainer, 2000). We then summarized these ICC and bias values (across all regions, within each combination of pipeline and de-facing method) using median values and box plots. We measured *p* values for pair-wise differences between de-facing methods using paired Wilcoxon Signed Rank tests (*wilcox.test()* in R). Non-GM atlas regions were omitted from summary measures of GM volumes. The total numbers of summarized atlas GM regions for each pipeline were: 116 with SPM12, 78 with FreeSurfer (70 for cortical thickness), and 64 for FSL-UKBB.

**Simulated test-retest:** To provide a reference for these values, we also measured the differences in regional measurements produced by modifying the input image headers (Nifti image format *s-form* matrix and *q-form* values) to simulate the head moving 5mm downward and rotating (pitch) 2 degrees upward. The image voxels were not resampled or otherwise modified; only the image headers were altered. This technique simulates the effects of varying participant position in the scanner upon downstream measurements (Schwarz et al., 2017b), and we include it in the comparison of de-facing methods as "simulated test-retest". Because the image voxels are identical, this simulation measures variability in biomarker measurement software but excludes the variability in imaging hardware and patient motion that would both be present in "true" test-retest MRI scan-pairs. However, since correctly-working de-facing software would modify only non-brain regions and leave all brain voxels completely identical to the original image, we expected their effects on brain measurements to be smaller than or similar to those of this simulated test-retest experiment.

**ADNI test-retest data set:** We also compared the effects of de-facing to the effects of a more-traditional MRI test-retest (scan-rescan) experiment using back-to-back T1-weighted MRI scan-pairs of 117 ADNI participants. Identical 3T MPRAGE sequences were used for both scans, and patients were not repositioned (Jack et al., 2008). This measure of test-retest uses a discrete set of participants vs. our other experiments and represents a more traditional measure of test-retest precision that includes variance from scanner noise and patient movement, in addition to variability of the biomarker measurement software. Because the voxels in brain regions are re-imaged and thus not identical, we expected these test-retest effects on brain measurements to be larger than both the simulated test-retest experiments (above) and those of the de-facing software.

## 4. Results

### 4.1. Validation criterion 1: Protection from face recognition

We present the results of face recognition testing, both with and without each of the de-facing techniques, in Table 1.

**Standard face reconstruction:** Using unmodified (non-de-faced) MRI, 97% (153/157) of participants were automatically matched to their correct corresponding MRI. Our proposed *mri_reface* reduced the rate to 30%. When using the standard face reconstructions with

**Table 1.**
Rates of automatically matching 5 photos of each participant to their correct corresponding MRI-based face reconstruction, using Microsoft Azure, before and after each de-facing technique.

| | Standard Face Reconstruction(using the input MRI only)(person attempting the matching is less skilled in MR image processing) | | Advanced Face Reconstruction (missing face regions automatically replaced with an average template)(person attempting the matching is highly skilled in MR image processing) |
|---|---|---|---|
| | MRI-based face recons where any face was detected | Participants correctly matched photos→MRI | Participants correctly matched photos→MRI |
| Original Images | 157/157 (100%) | 153/157 (97%) | N/A |
| *mri_deface* | 18/157 (11%) | 16/157 (10%) | 52/157 (33%) |
| *Pydeface* | 20/157 (13%) | 16/157 (10%) | 59/157 (38%) |
| *fsl_deface* | 5/157 (3%) | 5/157 (3%) | 44/157 (28%) |
| *mri_reface* | 157/157 (100%) | 47/157 (30%) | N/A |

*mri_deface, pydeface*, and *fsl_deface*, the missing face parts prevented face detection in a large majority of images (Table 1, column 1). Consequently, with these methods only 3-13% of participants' MRI-based reconstructions could be included in the "training set" of potential matches (the software cannot "learn" to recognize a face where it believes none exists). The resulting match rates across all photos (including those participants that could not be included in the training set) were between 3% (*fsl_deface*) and 13% (*pydeface*) (Table 1, column 2). These results may represent success rates of re-identifying de-faced data when attempted by an individual who lacks the skill in image processing to overcome the face detection issue.

**Advanced face reconstruction:** When we used our advanced face reconstructions (Fig. 4), faces were almost-always detected, and this allowed for face recognition upon the partially de-faced images with success rates of 33% after *mri_deface*, 38% after *pydeface*, and 28% after *fsl_deface* (Table 1 column 3). We directly compare these rates with the 30% from *mri_reface* when using the standard reconstructions, because its outputs already contain (altered, but standard-reconstructable) faces. These results may represent success rates of re-identifying de-faced data when attempted by a highly-skilled individual who adapts their reconstruction approach to each de-facing method.

**Analyses of which participants were correctly matched:** All but four participants were correctly matched without de-facing, so this was not enough to establish trends. The common set of participants that were correctly recognized after all four de-facing methods methods (using advanced reconstructions for *fsl_deface, pydeface*, and *mri_deface*) was only 8/157 (5.1%). Across the 7 de-faced experiments in Table 1 (basic reconstructions only for mri_reface and basic + advanced for the 3 alternatives), we summed for each participant how many times (out of 7) they were correctly identified, and we tested this number for demographic correlations. Men were identified more often than women (t-test $p<.001$), and identified participants had larger head sizes (Spearman $p=0.02$). Trends were also observed toward increased correlations with older age (Spearman $p=0.15$), taller height (Spearman $p=0.07$), and greater weight (Spearman $p=0.07$). No correlation was observed with body mass index (BMI; Spearman $p=0.74$). We present plots of these data in supplementary material.

### *4.2. Validation criterion 2: Minimizing effects on brain measurements*

Effects of each de-facing method on GM volume and cortical thickness measures are plotted in Fig. 5. In supplementary material, we also provide the raw values from Fig. 5, the ICC and bias values for all regions individually, and a corresponding plot of un-signed (magnitude) bias. We show several examples of instances where de-facing procedures produced large changes in segmentation results in Fig. 6.

**QC Results:** Scans from 9 participants were removed from analyses with the SPM12 pipeline because its segmentations of the original, unmodified images were visually judged by trained image analysts as grossly invalid. By the same criteria, 10 scans were removed from analyses with the FSL-UKBB pipeline, and 7 were removed from analyses with FreeSurfer 6.0.

**Rates of failure to generate measurements:** In Table 2, we present the number of scans (out of 300) where each de-facing method caused a pipeline to produce no measurements or produce only zero or NaN measurements. Scans where this occurred for the unmodified image (see above) are counted only in the first row. Our proposed *mri_reface* caused the smallest number of failures (1 vs 2-3).

**Global measurements:** First we measured how each de-facing algorithm affected global measurements of each scan: total tissue volume (TTV; sum of gray matter and white matter, used as a global biomarker) and total intracranial volume (TIV; aka ICV, frequently used as a nuisance covariate in tissue volumes analyses). TIV and TTV measurements with all pipelines and all de-facing techniques had biases <1% (all but one combination were <0.5%) in magnitude. ICC's of TIV and TTV measurements were >0.99 with SPM12 and >0.96 with FreeSurfer. ICC's of TIV's from the FSL-UKBB pipeline were 0.89 when used with *fsl_deface*, but were >0.95 with all other combinations. Overall, these global measurements were only minimally affected by de-facing techniques.

**ICC (non-systematic error) of regional measurements:** There was no de-facing method that outperformed all others across all pipelines. SPM-based regional measurements with de-facing methods had ICC values >0.9, except for left/right frontal pole with *fsl_deface* (left/right ICC=0.84/0.86). FreeSurfer-based measurements all had ICC's>0.83, except for left/right frontal pole with *fsl_deface* (thickness left/right ICC=0.70/0.67; volume left/right ICC=0.63/0.65). The FSL-UKBB pipeline had ICC's >0.88, except for left pallidum GM volumes with *mri_reface, pydeface*, and *fsl_deface* (ICC=0.77, 0.77, 0.81 respectively). Our *mri_reface* had the largest median ICC for GM volumes from SPM12 (0.996, $p<.001$ vs. *mri_deface*) and from FreeSurfer (0.977, $p<.001$ vs. *fsl_deface*), but for both of these the difference in median ICC from the next-best method was <0.005. For FreeSurfer cortical thickness, the largest median ICC was also with *mri_reface* (0.964), but the difference between it and the next best method (*pydeface*) was < 0.001 and not statistically significant ($p=0.547$). The lowest-performing regional measurement with *mri_reface* had ICC=0.78 (left pallidum volume with the FSL-UKBB pipeline), and all other regional ICCs with *mri_reface* were >0.87.

ICCs with only perturbing image geometry headers (simulated test-retest) were greatly higher than all de-face methods with the SPM and FSL-based pipelines. With FreeSurfer, this modification actually added more noise to cortical thickness measurements than any tested de-face technique despite not altering the image voxels, and for GM volumes it added more noise than all de-facers except *mri_deface*. Conversely, this header modification had no effect on the FSL-UKBB pipeline. ICCs for the test-retest data set were substantially lower than all defacers, for all pipelines.

**Bias (systematic error) of regional measurements:** As with ICC, there was no de-facing method that outperformed all others across all

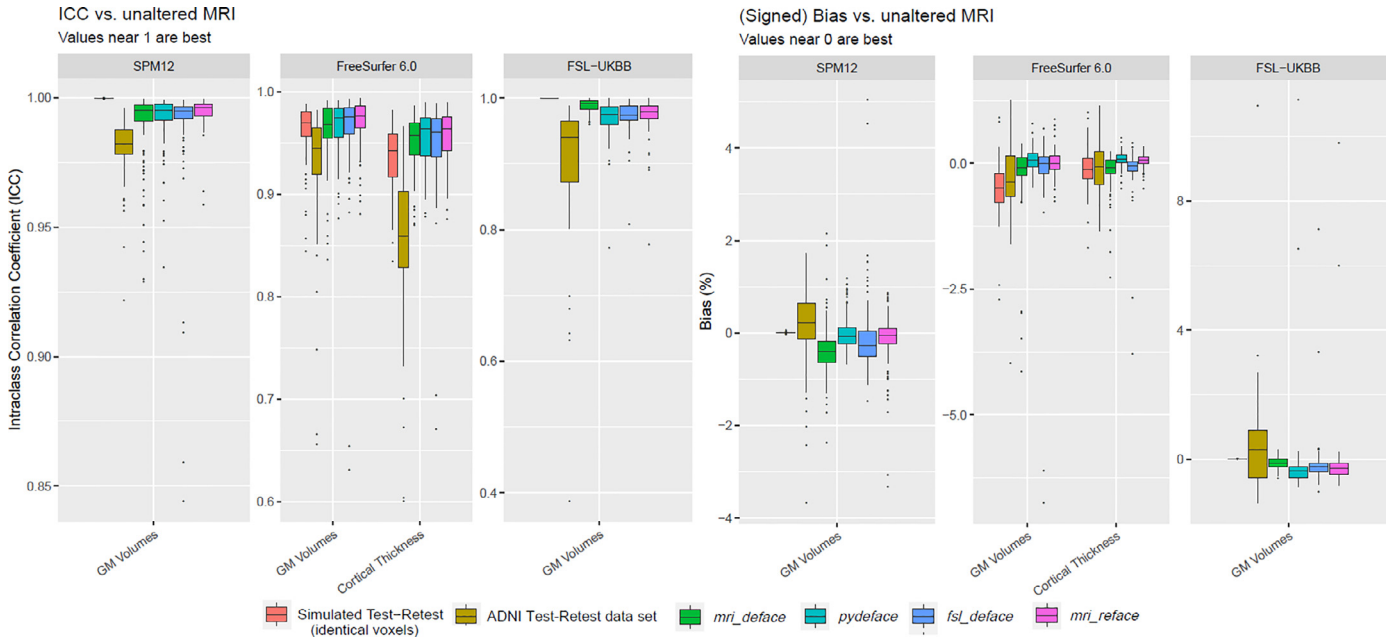


**Fig. 5.** Effects of de-facing methods on regional measurements of GM volume and Cortical Thickness from SPM, FreeSurfer, and FSL. Left: Intra-class correlation measurements (ICC) measure non-systematic error (noise) of measurements from unmodified vs de-faced scans with each de-facing method, and two different measures of test-retest error. Each plot shows the summary of ICC across atlas regions, where ICC between original vs. de-faced measurements were independently calculated for each region across scans of 300 participants. Higher ICC values indicate less noise. Right: Bias measures the systematic error as the percentage offset between the 1=1 line and a fit linear line, evaluated at the centercept (center of the x axis) for each region. Values near 0 are best. Each plot shows the summary of bias across atlas regions. We also provide the raw values for these plots, and a corresponding plot of un-signed (magnitude) bias, in supplementary material.

**Table 2.**
Numbers of scans (out of 300) where each pipeline + method combination produced no measurements or produced only zero or NaN measurements. Scans that failed using the original (unmodified) image are counted only in the first row.

| Method / Pipeline | SPM12 | FreeSurfer 6.0 | FSL-UKBB | Sum across pipelines |
|---|---|---|---|---|
| Original (unmodified) | 9 | 7 | 10 | 26 |
| *mri_deface* | 2 | 0 | 1 | 3 |
| *Pydeface* | 2 | 0 | 0 | 2 |
| *fsl_deface* | 0 | 1 | 1 | 2 |
| *mri_reface* | 0 | 1 | 0 | 1 |

pipelines. Across all pipelines, the greatest measurement biases occurred in GM volume measurements of the pallidum with FSL-UKBB, where volumes measured with de-faced data were an average of 6-11% larger than with unaltered images. The next-largest measurement biases occurred in measurements of the frontal pole; FreeSurfer measured these regions (both volume and thickness) as an average of 3-6.75% smaller after the *fsl_deface* and *mri_deface* methods, and SPM12 measured these as an average of 4.5-5.0% larger after *fsl_deface*. After these exceptions, all measurements had average biases <3.3%. Our *mri_reface* had the smallest median biases (closest to 0) with SPM12 GM volumes ($p$=.084 vs. *pydeface* (not significant)). With FreeSurfer GM volumes, *pydeface* had the smallest median biases ($p$=.034 vs. *mri_reface*), and with FreeSurfer cortical thickness, *fsl_deface* had the smallest median biases ($p$<.001 vs. *mri_reface*). For FSL-UKBB GM volumes, the best method was *mri_deface* ($p$<.001 vs. *fsl_deface* at rank 2; $p$<.001 vs. *mri_reface* at rank 3). Modifying the image header geometry had minimal systematic bias effects on the SPM- or FSL-based pipelines, but FreeSurfer volume and thickness measurements were biased to larger magnitudes (mostly toward smaller measurements) with modifying the image header than with any de-face method. In the test-retest data set, GM volumes in the repeat scan vs. the first scan were substantially larger for the SPM12 and FSL-UKBB pipelines, and substantially smaller for FreeSurfer. Despite their differing directions, the magnitudes of the biases for all three GM volume pipelines were substantially greater than the effects of any de-facing pipeline. However, test-retest biases with FreeSurfer cortical thickness were relatively small and were comparable with the de-face methods.

**Hippocampal and Entorhinal Measurements:** These regions were included in the above analyses, but we also examine them separately because they are of specific interest to studies of aging and Alzheimer's disease (in supplementary material, we provide the quantitative results for all regions). Across all de-facing methods, ICC's of hippocampal volume measurements were >0.99 with SPM12, >0.96 with FreeSurfer, and >0.96 with FSL-UKBB. Entorhinal cortical thickness values had ICC's >0.94 with FreeSurfer. Across all de-facing methods, median biases were <1% for all hippocampal volume and entorhinal thickness measures. Although the effects of de-facing on these regions were relatively small, it is notable that they were affected at all given their distance from altered face or ear regions. For each pipeline's hippocampal volume measurements, and FreeSurfer's entorhinal thickness measurements, we also measured the effects of these de-facing techniques on: a) Area Under the Receiving Operating Characteristic Curve (AUROC) separation of cognitively unimpaired vs. Alzheimer's disease participants; b) Spearman correlation (rho) with each participant's Mini-Mental State Exam (MMSE) score (Folstein et al., 1975); or c) Spearman correlation with each participant's Clinical Dementia Rating-Sum of Boxes (Lynch et al., 2006). We assessed AUROC using the *roc* function from the *pROC* package (Robin et al., 2011), and Spearman correlation (with confidence intervals) using the *SpearmanRho* function from the *desctools* package

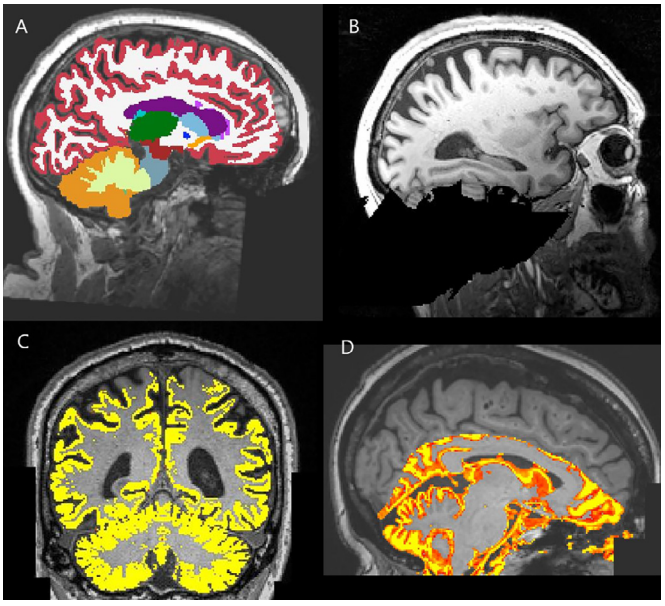


**Fig. 6.** Examples of quantification errors due to de-facing. In all instances, outputs from the unmodified images did not have these errors. A) Pydeface + FreeSurfer: voxels in frontal lobe were segmented as non-brain tissue (not colored), despite not being adjacent to the tissue removed by de-facing. B) mri_deface: voxels in the cerebellum were removed by defacing while the face itself was left intact (affects all pipelines). C) fsl_deface + SPM12: gray matter in precentral gyrus and other superior regions was misclassified (not marked yellow) as a result of de-facing. De-facing did not remove any tissue nearby these regions. D) fsl_deface + FSL-UKBB: gray matter in most of the cortex was misclassified (not marked red/yellow) as a result of de-facing.

(Signorell, 2019). Compared within each pipeline across de-facing methods, 95% confidence intervals for each measure greatly overlapped and no meaningful differences or consistent trends were observed (these analyses are shown only in supplementary material).

**Run times:** We recorded the run times for each de-facing program when we ran each of the 300 ADNI T1-w images in our validation dataset using our local compute cluster of 28 "white box" servers with Intel Xeon processors (manufactured between 2010 and 2020) with 24-40 cores (assigned one core/job) and 128GB+ RAM. Mean (sd) runtimes in minutes for each method were: mri_deface: 5.06 (1.59); pydeface: 1.80 (0.52); fsl_deface: 2.32 (1.05); mri_reface: 51.95 (17.65).

## 5. Discussion

### 5.1. Face recognition rates

**Recognition rates without de-facing**: Since our previous study (Schwarz et al., 2019), our measured face recognition rates on unaltered MRI have improved from 83% of 84 participants to 97% of 157 participants. The improved detection rate with the recently-updated Microsoft algorithm (see further discussion in supplementary material) is prima facie evidence of the inevitable improvement of face recognition technology with time. The threat of individual re-identification will continue to increase with advancing face recognition technology, suggesting that in the future some form of de-facing should be considered a necessary part of image de-identification prior to external data sharing.

**Recognition rates across de-facing methods:** We found that a highly-skilled individual could re-identify up to 38% of scans, even *after* popular de-facing programs. Even with our standard face reconstructions (i.e. assuming re-identification attempted by a less-skilled individual), recognition rates after these popular programs were as high as 10%. These results suggest that for the goal of preventing face recognition, effective de-facing is a much more difficult problem than previously appreciated. We also found that after de-facing, men were still recognized significantly more often than women. Unsurprisingly (because these variables are themselves correlated with sex), we also found correlations of varying significance with larger head size (TIV), weight, and height. However, these correlations disappeared when using BMI, suggesting that this is an effect of participants' sex or gender rather than their relative size or physical fitness. We expected that larger participants would be *less* recognizable due to larger heads increasingly making contact with (being deformed by) the MRI head coil, but we found the opposite effect. These results are difficult to interpret because we cannot separate (due to too-few participants that were mismatched using non-defaced scans) between potential demographic biases in the de-facing software vs. potential demographic biases in the face recognition system. We will explore this further in future work.

### 5.2. Effects of de-facing on brain measurements

Another surprising result was that effects on downstream regional measurements were measureable, even in regions distant from any modified parts of the image. Although frontal pole, orbitofrontal, and temporal pole regions (relatively near to the face) often had the largest differences, some of the largest effects also occurred in measurements of deep grey structures, far from any modified voxels. Original validations of many de-facing techniques have focused on ensuring that minimal brain voxels are modified (Alfaro-Almagro et al., 2018; Bischoff-Grethe et al., 2007; Schimke and Hale, 2011), but our findings agree with another recent study (de Sitter et al., 2020) in showing that measurable effects can also occur in measurements from unmodified, distant brain regions. These distant effects may arise from both local and non-local effects on linear or nonlinear registration between each scan and standard templates, or from generative/Bayesian segmentation frameworks where the estimated probability of a given tissue type at each location is defined relative to the appearance of all other tissue types in the image. Therefore, creating de-facing software that does not alter brain measurements is *also* much more difficult than previously appreciated.

**Comparisons with test-retest:** It is reasonable to expect that ideal de-facing software should produce effects on brain measurements that are smaller than those of re-scanning the same participant without significant time passage (test-retest). We tested this expectation against two different measurements of test-retest. The "ADNI test-retest data set" is a "true" test-retest experiment that includes scans of 117 (different) ADNI participants who were scanned back-to-back. These test-retest scans did not reposition participants between the scans, so its effects are likely underestimated vs. typical test-retest (scan-rescan) experiments. Even so, all of the de-facing pipelines achieved the goal of having smaller effects than (this underestimate of) test-retest: ICCs were substantially higher and biases were somewhat smaller (nearer to zero).

However, all the tested brain measurements involve (only) brain regions, which all the de-facing programs are designed to not modify. In theory, one would expect identical brain voxels to produce identical brain measurements, but this is not the case. Consequently, we argue that ideal de-facing software should have effects on brain measurements that are more comparable to test-retest *with identical brain voxels* (i.e. measuring variance of the measurement software without variance in the imaging/participant) than with standard test-retest experiments (which measure both sources of variance). To this end, our *simulated* test-retest data used the same 300 scans as all the other experiments; geometry in the image header was modified but no resampling was performed and all image voxels (not just the brain) were identical to the original. All the de-facing programs had effects on the SPM- and FSL-based pipelines that were substantially larger than this standard of (same-voxels) test-retest, but the opposite was true for FreeSurfer. In total, effects of de-facing programs on brain measurements were smaller than test-retest, but still measureable and (mostly) larger than our simulated identical-voxel test-retest. We expected them to lie in between

these two measures, but closer to the simulated than "true" test-retest variants.

**Comparisons with effect sizes of the biology of interest:** We found that the effects of de-facing were extremely negligible in cross-sectional comparisons of AD pathology in the hippocampus and entorhinal cortex. However, the statistical power of those particular comparisons (before de-facing) is very large, and these regions were also among the most robust in all the brain in our test-retest measurements, so these case-control comparisons represent a best-case scenario. We found much larger effects in other regions (mainly frontal and subcortical), where de-facing with some techniques biased *average* GM volumes by as much as 11%. By comparison, the rate of annual hippocampal atrophy in AD participants is only 2-4% (Josephs et al., 2017) and many brain pathologies have effects far smaller than AD. Thus, even though the effects of de-facing were smaller than test-retest, they were sometimes large and problematic in comparison to biological effect sizes.

### 5.3. Comparing de-facing methods

When face recognition was performed using the advanced face reconstructions as-needed, the de-facing methods that best prevented face recognition were *fsl_deface* (44/157=28%) and *mri_reface* (47/157=30%). We designed this scenario to reflect recognition performance when the person attempting re-identification has the skills to exploit more remaining information in the images. However, when we used standard face reconstructions with minimal image processing, *fsl_deface* had the lowest face recognition rate (3%). This low rate is largely due to the fact that only 3% of reconstructions were detected as faces i.e. recognition could not be performed on the rest. The *mri_reface* and *fsl_deface* methods reliably removed the eye regions, which is likely why they prevented face recognition substantially better than *mri_deface* or *pydeface*. The *mri_deface* and *pydeface* methods also do not attempt to remove the ears, although our study did not attempt ear recognition. Our *mri_reface* is also the only method that attempts to remove aliased face regions (artifacts) in front or behind the head, although our study did not attempt to exploit that information during face recognition.

No de-facing method consistently outperformed all others in minimizing effects on brain measurements. Median ICC's with *mri_reface* were the best for GM volumes with SPM12 and with FreeSurfer. Although both differences were statistically significant, their magnitudes vs. the next-best methods were <.005 ICC. For FreeSurfer cortical thickness, ICC values were also the smallest with *mri_reface*, but the difference between it and the next-best method (*pydeface*) was within 0.001 and not statistically significant. Its measurement biases were the nearest to zero for SPM12 GM volumes, and they were second-smallest for FreeSurfer GM volumes (the smallest was *pydeface*). It was also caused fewer instances (1 vs 2-3) where pipelined to fail to produce measurements for any scan that had worked prior to de-facing. We attribute these advantages to its generation of outputs that resemble natural images with realistic image statistics. One cost to these improvements is that *mri_reface* has a significantly longer run-time (≈50 minutes) than the tested alternatives (<5 minutes). These costs occur primarily in computing the nonlinear registration (vs. affine with other methods) needed for more-accurate localization of face regions.

Taken together, our proposed *mri_reface* was the second best method for preventing re-identification via face recognition, and it had the smallest effects on brain measurements in more pipelines than any other method. We hypothesize that its relatively better protection from face recognition is because it was designed to remove (replace) more/larger identifiable portions of input images (than *mri_deface* or *pydeface*), and its smaller effects on brain measurements are because it replaces rather than removes voxels to produce output images that better resemble those that SPM, FreeSurfer, and FSL were designed for. *fsl_deface* also performed well; it was the best for preventing face recognition (28% vs 30% with *mri_reface*), and it was also among the leading methods for minimizing measurement effects for each pipeline.

### 5.4. Strengths and limitations of current study

**Strengths:** We took a comprehensive approach to comparing MRI-defacing techniques by simultaneously considering 1) their efficacy at preventing re-identification via face recognition, and 2) their effects on brain measures from output scans; previous studies comparing de-facing techniques have considered only the latter (de Sitter et al., 2020). We also compared these measurement effects with two different variants of test-rest experiments.

**Face recognition scenarios:** Our face recognition testing methodology of matching participant photos to MRIs reflects a hypothetical situation where a highly motivated and skilled individual has reason to believe that a participant is part of a particular study and wishes to find them among shared de-identified study data. The opposite problem, where someone has a de-identified MRI and wishes to re-identify it using face recognition from among potentially all humans on earth, is of course much more difficult. However, the 97% recognition rate on the current scenario suggests that MRI scans of the brain can provide comparable face-recognizing information with that of standard photographs, and because existing technologies can successfully identify face photos from databases of > 12 million people with failure rates <1% (Grother et al., 2018), we hypothesize that identifying a fully-unknown brain MRI is not implausible with current or near-future technology. Moreover, the ability to re-identify an individual will only improve as technology inevitably advances.

**Alternate face recognition algorithms:** We did not compare multiple face recognition algorithms. We used Microsoft Azure because of an existing Mayo Clinic service agreement with Microsoft that allowed us to test the Azure face recognition services under a secure private platform without potentially exposing data to the public. Microsoft's face recognition algorithms have also ranked #1 in ongoing NIST face recognition software comparisons (Grother et al., 2018). Because most leading face recognition algorithms are proprietary cloud-based technologies, securing the data used for testing presents complicated legal and technical challenges. Thus, although we acknowledge that testing the relative performance of de-facing techniques against multiple face recognition systems would be of value, we must leave this exercise for future work. We also did not examine human-based face matching because comprehensive prior work (Prior et al., 2009) has shown that it has much lower matching rates (40% of participants could perform the matching with success rates exceeding statistical chance) than with automatic face recognition.

**De-facing approaches:** We included what we believed to be the most popular de-facing approaches currently in use, and we tested each using its default (or only available) settings. It is possible that alternate settings or pre-processing approaches that we were not aware of may have improved the other software's performances. Future work would ideally include a "grand challenge"-like competition, where software authors would each run their own approaches on a shared data-set and the results would be compared by third parties.

**Demographics:** We performed tests using images of older adults from studies of aging and Alzheimer's disease. It is possible that the performance of either criterion may not generalize to other populations. Our *mri_reface* and *fsl_deface* were both designed for older adults, so pediatric populations may be especially challenging. In future work, we plan to test images from other populations and adapt our method as needed (e.g. by generating population-specific templates).

## 6. Conclusion

Without de-facing, automated face recognition was able to match participant photographs to the correct MRI with 97% accuracy, suggesting that in the future some form of de-facing should be considered a necessary part of image de-identification prior to external data transmission or sharing. After de-facing with popular programs, recognition rates still ranged from 28%-38%. Effects on SPM12, FreeSurfer 6, and FSL's re-

gional brain measurements of de-faced scans were modest but measureable: smaller than standard test-retest but larger than would be expected given that by-design these software alter only non-brain voxels. Compared to tested popular de-facers, our proposed *mri_reface* performed second best at preventing face recognition (30% vs 28% with *fsl_deface*) and modestly reduced effects on brain volume/thickness measurements. Still, our proposed method's improvements were very modest, and further work is needed to greatly improve MRI de-facing techniques while minimizing and reducing their impacts on brain measurements.

## Acknowledgements

The authors give their thanks to all the volunteers, participants, and coordinators who contributed to this research, with special thanks to Steven M. Smith, Josie M. Williams, Paul D. Lewis, Steven J. Demuth, Soudabeh Kargar, and Zuzana Nedelska. Thank you to Stephen Weigand for statistical advice. We gratefully thank our funding sources: NIH grants R56 AG068206, U01 AG006786, P50 AG016574, R01 AG034676, R37 AG011378, R01 AG041851, R01 NS097495, R01 AG056366, U01 NS100620; The GHR Foundation; The Elsie and Marvin Dekelboum Family Foundation; The Alexander Family Alzheimer's Disease Research Professorship of the Mayo Clinic; The Liston Award; The Schuler Foundation; and The Mayo Foundation for Medical Education and Research. We also gratefully acknowledge the support of NVIDIA Corporation with the donation of the Quadro P6000 GPU used in generating 3D facial reconstructions for this research.

## Data and Code Availability Statement

ICC and bias calculations for every regional measurement from each pipeline are provided in supplementary material. ADNI images are available directly through ADNI (http://adni.loni.usc.edu/). The face recognition data-set contains participant photos, which are considered primary identifiers; to protect participant privacy and comply with the IRB and consent forms, these data cannot be shared. MRI and other data from the Mayo Clinic Study of Aging and the Alzheimer's Disease Research Center are available to qualified academic and industry researchers by request to the Mayo Clinic Study of Aging/Alzheimer's Disease Research Center Executive Committee. Our *mri_reface* software is still in development, but it will be provided to qualified researchers by request and will later be released as free for non-commercial research use. The *mri_deface, pydeface*, and *fsl_deface* de-face software are each available from their respective authors. Microsoft Azure face recognition is a proprietary cloud-based service by Microsoft.

Disclosure Statements

Dr. Schwarz receives funding from the National Institutes of Health, related and unrelated to this study, and has a related US patent pending.

Dr. Kremers received grant funding from NIH for this study, and from NIH, DOD, AstraZeneca, Biogen and Roche unrelated to this study.

Ms. Wiste reports no disclosures.

Dr. Gunter receives funding from the NIH, and has a related US patent pending.

Dr. Vemuri receives funding from the NIH.

Mr. Spychalla reports no disclosures.

Dr. Kantarci serves on the data safety monitoring board for Takeda Global Research and Development Center, Inc.; data monitoring boards of Pfizer and Janssen Alzheimer Immunotherapy; and receives research support from Avid Radiopharmaceuticals and Eli Lilly, the Alzheimer's Drug Discovery Foundation, and NIH.

Dr. Schultz reports no disclosures.

Dr. Sperling has received research funding from NIH, Alzheimer's Association, GHR Foundation, Eli Lilly, Janssen, and Eisai. She has served as a consultant for AC Immune, Acumen, Cytox, Janssen, Neurocentria, Prothena, Renew.

Dr. Knopman served on a Data Safety Monitoring Board for the DIAN study. He serves on a Data Safety monitoring Board for a tau therapeutic for Biogen, but receives no personal compensation. He is an investigator in clinical trials sponsored by Biogen, Lilly Pharmaceuticals and the University of Southern California. He serves as a consultant for Samus Therapeutics, Third Rock and Alzeca Biosciences but receives no personal compensation. He receives research support from the NIH.

Dr. Petersen serves on scientific advisory boards for Elan Pharmaceuticals, Wyeth Pharmaceuticals, and GE Healthcare; receives royalties from publishing Mild Cognitive Impairment (Oxford University Press, 2003); and receives research support from NIH.

Dr. Jack serves on an independent data monitoring board for Roche and has consulted for Eisai, but he receives no personal compensation from any commercial entity. He receives research support from NIH and the Alexander Family Alzheimer's Disease Research Professorship of the Mayo Clinic and has a related US patent pending.

## Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.neuroimage.2021.117845.

## References

Abramian, D., Eklund, A., 2019. Refacing: reconstructing anonymized facial features using GANs. In: Proc. Int. Symp. Biomed. Imaging.

ADNI Home [WWW Document], 2013. URL www.adni-info.org (accessed 1.1.15).

Alfaro-Almagro, F., Jenkinson, M., Bangerter, N.K., Andersson, J.L.R., Griffanti, L., Douaud, G., Sotiropoulos, S.N., Jbabdi, S., Hernandez-Fernandez, M., Vallee, E., Vidaurre, D., Webster, M., McCarthy, P., Rorden, C., Daducci, A., Alexander, D.C., Zhang, H., Dragonu, I., Matthews, P.M., Miller, K.L., Smith, S.M., 2018. Image processing and Quality Control for the first 10,000 brain imaging datasets from UK Biobank. Neuroimage 166, 400–424. doi:10.1016/j.neuroimage.2017.10.034.

Andersson, J., Smith, S., Jenkinson, M., 2008. FNIRT-FMRIB's non-linear image registration tool. Annual Meeting of the Organization for Human Brain Mapping (OHBM). Wiley.

Ashburner, J., Friston, K.J., 2005. Unified segmentation. Neuroimage 26, 839–851. doi:10.1016/j.neuroimage.2005.02.018.

Avants, B.B., Epstein, C.L., Grossman, M., Gee, J.C., 2008. Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain. Med. Image Anal. 12, 26–41. doi:10.1016/j.media.2007.06.004.

Avants, B.B., Yushkevich, P., Pluta, J., Minkoff, D., Korczykowski, M., Detre, J., Gee, J.C., 2010. The optimal template effect in hippocampus studies of diseased populations. Neuroimage 49, 2457–2466. doi:10.1016/j.neuroimage.2009.09.062.

Bischoff-Grethe, A., Ozyurt, I.B., Busa, E., Quinn, B.T., Fennema-Notestine, C., Clark, C.P., Morris, S., Bondi, M.W., Jernigan, T.L., Dale, A.M., Brown, G.G., Fischl, B., 2007. A technique for the deidentification of structural brain MR images. Hum. Brain Mapp. 28, 892–903. doi:10.1002/hbm.20312.

de Sitter, A., Visser, M., Brouwer, I., Cover, K.S., van Schijndel, R.A., Eijgelaar, R.S., Müller, D.M.J., Ropele, S., Kappos, L., Rovira, Filippi, M., Enzinger, C., Frederiksen, J., Ciccarelli, O., Guttmann, C.R.G., Wattjes, M.P., Witte, M.G., de Witt Hamer, P.C., Barkhof, F., Vrenken, 2020. Facing privacy in neuroimaging: removing facial features degrades performance of image analysis methods. Eur. Radiol. 30, 1062–1074. doi:10.1007/s00330-019-06459-3.

Desikan, R.S., Ségonne, F., Fischl, B., Quinn, B.T., Dickerson, B.C., Blacker, D., Buckner, R.L., Dale, A.M., Maguire, R.P., Hyman, B.T., Albert, M.S., Killiany, R.J., 2006. An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. Neuroimage 31, 968–980. doi:10.1016/j.neuroimage.2006.01.021.

Elmahmudi, A., Ugail, H., 2019. Deep face recognition using imperfect facial data. Futur. Gener. Comput. Syst. 99, 213–225. doi:10.1016/j.future.2019.04.025.

Emeršič, Ž., Štruc, V., Peer, P., 2017. Ear recognition: More than a survey. Neurocomputing 255, 26–39. doi:10.1016/j.neucom.2016.08.139.

Fischl, B., 2012. FreeSurfer. Neuroimage 62, 774–781. doi:10.1016/j.neuroimage.2012.01.021.

Folstein, M.F., Folstein, S.E., McHugh, P.R., 1975. Mini-mental state" A practical method for grading the cognitive state of patients for the clinician. J. Psychiatr. Res. 12, 189–198 https://doi.org/0022-3956(75)90026-6 [pii].

Fonov, V.S., Collins, L.D., 2018. *BIC Defacing Algorithm*. bioArxiv. https://doi.org/10.1101/275453.

Grother, P., Ngan, M., Hanaoka, K., 2018. *Ongoing face recognition vendor test (FRVT) part 2*: Gaithersburg, MD. https://doi.org/10.6028/NIST.IR.8238

Gulban, O.F., Nielson, D., Poldrack, R., Lee, J., Gorgolewski, C., Vanessasaurus, Ghosh, S., n.d. poldracklab/pydeface. https://doi.org/10.5281/zenodo.3524400

Hanke, M., 2015. mridefacer.

Jack, C.R.J., Bernstein, M.A., Borowski, B.J., Gunter, J.L., Fox, N.C., Thompson, P.M., Schuff, N., Krueger, G., Killiany, R.J., DeCarli, C.S., Dale, A.M., Weiner, M.W., Carmichael, O.W., Tosun, D., Weiner, M.W., 2010. Update on the magnetic resonance imaging core of the Alzheimer's disease neuroimaging initiative. Alzheimer's Dement. 6, 212–220. https://doi.org/10.1016/j.jalz.2010.03.004.Update

Jack, C.R.J., Bernstein, M.A., Fox, N.C., Thompson, P., Alexander, G., Harvey, D., Borowski, B., Britson, P.J., Whitwell, J.L., Ward, C., Dale, A.M., Felmlee, J.P., Gunter, J.L., Hill, D.L.G., Killiany, R., Schuff, N., Fox-Bosetti, S., Lin, C., Studholme, C., DeCarli, C.S., Krueger, G., Ward, H.A., Metzger, G.J., Scott, K.T., Mallozzi, R., Blezek, D., Levy, J., Debbins, J.P., Fleisher, A.S., Albert, M., Green, R., Bartzokis, G., Glover, G., Mugler, J., Weiner, M.W., 2008. The Alzheimer's Disease Neuroimaging Initiative (ADNI): MRI methods. J. Magn. Reson. Imaging JMRI 27, 685–691. doi:10.1002/jmri.21049.

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. doi:10.1006/nimg.2002.1132.

Josephs, K.A., Dickson, D.W., Tosakulwong, N., Weigand, S.D., Murray, M.E., Petrucelli, L., Liesinger, A.M., Senjem, M.L., Spychalla, A.J., Knopman, D.S., Parisi, J.E., Petersen, R.C., Jack, C.R., Whitwell, J.L., 2017. Rates of hippocampal atrophy and presence of post-mortem TDP-43 in patients with Alzheimer's disease: a longitudinal retrospective study. Lancet Neurol. 16, 917–924. doi:10.1016/S1474-4422(17)30284-3.

Lynch, C.A., Walsh, C., Blanco, A., Moran, M., Coen, R.F., Walsh, J.B., Lawlor, B.A., 2006. The Clinical Dementia Rating Sum of Box Score in Mild Dementia. Dement. Geriatr. Cogn. Disord. 21, 40–43. doi:10.1159/000089218.

Mazura, J.C., Juluru, K., Chen, J.J., Morgan, T.A., John, M., Siegel, E.L., 2012. Facial recognition software success rates for the identification of 3D surface reconstructed facial images: Implications for patient privacy and security. J. Digit. Imaging 25, 347–351. doi:10.1007/s10278-011-9429-3.

Microsoft Corporation, 2019. *Microsoft Azure Face API Documentation [WWW Document]*. URL https://docs.microsoft.com/en-us/azure/cognitive-services/face/(accessed 12.10.18).

Milchenko, M., Marcus, D., 2013. Obscuring surface anatomy in volumetric imaging data. Neuroinformatics 11, 65–75. doi:10.1007/s12021-012-9160-3.

Otsu, N., 1979. A Threshold Selection Method from Gray-Level Histograms. IEEE Trans. Syst. Man. Cybern. 9, 62–66. doi:10.1109/TSMC.1979.4310076.

Petersen, R.C., Roberts, R.O., Knopman, D.S., Geda, Y.E., Cha, R.H., Pankratz, V.S., Boeve, B.F., Tangalos, E.G., Ivnik, R.J., Rocca, W.A., 2010. Prevalence of mild cognitive impairment is higher in men. The Mayo Clinic Study of Aging. Neurology 75, 889–897. doi:10.1212/WNL.0b013e3181f11d85.

Prior, F.W., Brunsden, B., Hildebolt, C., Nolan, T.S., Pringle, M., Vaishnavi, S.N., Larson-Prior, L.J., 2009. Facial Recognition From Volume-Rendered Magnetic Resonance Imaging Data. IEEE Trans. Inf. Technol. Biomed. 13, 5–9. doi:10.1109/TITB.2008.2003335.

R Development Core Team, 2008. *R: A Language and Environment for Statistical Computing*[WWW Document]. URL http://www.r-project.org

Revelle, W., 2019. psych: *Procedures for Psychological, Psychometric, and Personality Research*.

Roberts, R.O., Geda, Y.E., Knopman, D.S., Cha, R.H., Pankratz, V.S., Boeve, B.F., Ivnik, R.J., Tangalos, E.G., Petersen, R.C., Rocca, W.A., 2008. The Mayo Clinic Study of Aging: design and sampling, participation, baseline measures and sample characteristics. Neuroepidemiology 30, 58–69. doi:10.1159/000115751.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., Müller, M., 2011. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Informatics 12, 77.

Rorden, C., n.d.*Surf Ice [WWW Document*]. URL https://www.nitrc.org/projects/surfice/ (accessed 12.11.2018).

Schimke, N., Hale, J., 2011. Quickshear Defacing for Neuroimages Neuroimages. 2nd USENIX Conference on Health Security and Privacy.

Schwarz, C.G., Gunter, J.L., Ward, C.P., Vemuri, P., Senjem, M.L., Wiste, H.J., Petersen, R.C., Knopman, D.S., Jack, C.R., 2017a. The Mayo Clinic Adult Lifespan Template: Better Quantification Across the Lifespan. Alzheimer's Dement. 13, P792. https://doi.org/10.1016/j.jalz.2017.06.1071

Schwarz, C.G., Jones, D.T., Gunter, J.L., Lowe, V.J., Vemuri, P., Senjem, M.L., Petersen, R.C., Knopman, D.S., Jack, C.R., 2017b. Contributions of imprecision in PET-MRI rigid registration to imprecision in amyloid PET SUVR measurements. Hum. Brain Mapp. 38, 3323–3336. doi:10.1002/hbm.23622.

Schwarz, C.G., Kremers, W.K., Therneau, T.M., Sharp, R.R., Gunter, J.L., Vemuri, P., Arani, A., Spychalla, A.J., Kantarci, K., Knopman, D.S., Jack, C.R.J., 2019. Identification of Anonymous MRI Research Participants with Face Recognition Software. N. Engl. J. Med..

Signorell, A., 2019. *DescTools: Tools for Descriptive Statistics*.

Silva, J.M., Guerra, A., Silva, J.F., Pinho, E., Costa, C., 2018. Face De-Identification Service for Neuroimaging Volumes. In: Proc. - IEEE Symp. Comput. Med. Syst. 2018-June, pp. 141–145. doi:10.1109/CBMS.2018.00032.

Smith, S.M., Zhang, Y., Jenkinson, M., Chen, J., Matthews, P.M., Federico, A., De Stefano, N., 2002. Accurate, Robust, and Automated Longitudinal and Cross-Sectional Brain Change Analysis. Neuroimage 17, 479–489. doi:10.1006/nimg.2002.1040.

Vemuri, P., Senjem, M.L., Gunter, J.L., Lundt, E.S., Tosakulwong, N., Weigand, S.D., Borowski, B.J., Bernstein, M.A., Zuk, S.M., Lowe, V.J., Knopman, D.S., Petersen, R.C., Fox, N.C., Thompson, P.M., Weiner, M.W., Jack, C.R., 2015. Accelerated vs. unaccelerated serial MRI based TBM-SyN measurements for clinical trials in Alzheimer's disease. Neuroimage 113, 61–69. doi:10.1016/j.neuroimage.2015.03.026.

Wainer, H., 2000. The Centercept: An Estimable and Meaningful Regression Parameter. Psychol. Sci. 11, 434–436. doi:10.1111/1467-9280.00284.

Wickham, H., 2017. *tidyverse: Easily Install and Load the "Tidyverse"* [WWW Document]. URL https://cran.r-project.org/package=tidyverse

Zhang, Y., Brady, M., Smith, S., 2001. Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. Med. Imaging, IEEE Trans. 20, 45–57. doi:10.1109/42.906424.