Conditional De-Identification of 3D Magnetic Resonance Images

Lennart Alexander Van der Goten[1,2]
lavdg@kth.se
Tobias Hepp[3,4]
tobias.hepp@tuebingen.mpg.de
Zeynep Akata[3,4]
zeynep.akata@uni-tuebingen.de
Kevin Smith[1,2]
ksmith@kth.se

[1] KTH Royal Institute of Technology
Stockholm, SWEDEN
[2] Science for Life Laboratory
Solna, SWEDEN
[3] Max Planck Institute for Intelligent Systems
Tübingen, GERMANY
[4] University of Tübingen
Tübingen, GERMANY

*for the Alzheimer's Disease Neuroimaging Initiative**

Abstract

Privacy protection of medical image data is challenging. Even if metadata is removed, brain scans are vulnerable to attacks that match renderings of the face to facial image databases. Solutions have been developed to de-identify diagnostic scans by obfuscating or removing parts of the face. However, these solutions either fail to reliably hide the patient's identity or are so aggressive that they impair further analyses. We propose a new class of de-identification techniques that, instead of removing facial features, remodels them. Our solution relies on a conditional multi-scale GAN architecture. It takes a patient's MRI scan as input and generates a 3D volume conditioned on the patient's brain, which is preserved exactly, but where the face has been de-identified through remodeling. We demonstrate that our approach preserves privacy far better than existing techniques, without compromising downstream medical analyses. Analyses were run on the OASIS-3 and ADNI corpora.

1 Introduction

The digitalization of heath records has increased the risk of –and impact of– large scale data leaks. Although data compliance standards have been enacted to protect health records (HIPAA and GDPR), privacy of medical data is a growing concern. Three-dimensional scans such as magnetic resonance images (MRI) and computed tomography (CT), for example, contain an intrinsic privacy risk [20]. Detailed renderings of the head can be crafted from

*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

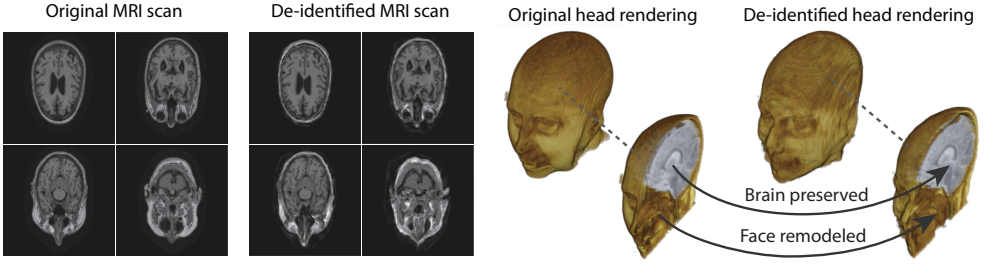

Figure 1: *De-identification through facial remodeling.* Face renderings from medical scans such as these MRIs represent a privacy vulnerability. To address this, we define a new class of de-identification techniques that aims to realistically *remodel* privacy-sensitive regions, such as the face, while *preserving* essential regions, such as the brain (illustrated OASIS-3).

3D scans using techniques such as volumetric raycasting, as in Figure 1. This vulnerability can expose the patient's identity if the renderings are matched to a face database [20, 21].

To prevent these types of attack, medical scans are currently de-identified using crude *removal-based* techniques [1, 23, 25] which seek to remove privacy-sensitive parts of the head (examples in Figure 3). However, as we demonstrate, these existing techniques fail to reliably hide the patient's identity – or they are so aggressive that they impair further medical analyses. A better solution is needed.

One might ask *why de-identify the face when one can just remove everything except the brain?* This approach, known as skull-stripping [27], does provide excellent privacy guarantees, but unfortunately **renders the scan useless for many types of clinical analysis**. Automated tools for analyzing MRI scans rely on landmarks within the head, and fail when they are removed [3]. Furthermore, skull-stripping corrupts measurements of important tissues and fluids, such as extra-cranial CSF [1]. For these reasons, *remodeling the head* rather than deleting privacy-sensitive regions is desirable, because it protects privacy and at the same time ensures robustness of downstream medical analyses.

Therefore, in this work, we define a new class of de-identification techniques that *remodels* the privacy-sensitive regions without altering the content of medically relevant data (see Figure 1). Under such a remodeling approach, the face, eyes, oral and nasal cavities, *etc.* should exhibit realistic appearance and structure of appropriate size, but should otherwise be independent of the original data. To solve this task, we propose a novel model called *Convex Privacy* GAN, or CP-GAN, that conditions on a convex hull of the skull extracted from the scan to be de-identified. The generator learns to synthesize volumes that *preserve* medically-sensitive regions such as the brain, while non-invertibly *remodeling* privacy-sensitive characteristics from the original scan.

The main contributions of this work are as follows: (1) We define a novel methodology to ensure privacy in medical imagery in which medically relevant regions are preserved and privacy-sensitive regions are de-identified. (2) We propose CP-GAN, a conditional multi-scale volumetric GAN that realizes a solution to the aforementioned methodology. (3) Through human- and model-based experiments, we show that CP-GAN preserves privacy in MRI scans more reliably than removal-based techniques without adversely affecting downstream analyses. In addition, we make technical contributions towards the generation of the convex hull and surface representations necessary for the privacy conditioning of the GAN. Source code as well as a video demonstration can be found in the *supplementary material*.

2 Related Work

A handful of de-identification techniques exist for MRI scans, which are conventionally used for sharing and distribution of MRI data. These existing methods rely on a *removal* approach to privacy. DEFACE [1] estimates the probabilities of voxels belonging to the face based on an atlas of healthy control subjects. The scan is de-identified by setting intensities of voxels whose probabilities are small enough to zero. QUICKSHEAR [25] is a fast but simple approach that computes a hyperplane to divide the MRI into two regions: one containing facial structures, and the other containing the brain of the scan. Voxels in the first part are set to zero. FACE MASK [23] uses a filtering method to blur the facial features. These existing de-identification approaches are based on traditional computer vision techniques; we believe that the proposed algorithm is the first to adopt a learning-based approach.

While not a de-identification method, *Shin et al*. [28] recently proposed a *pix2pix*-inspired model [11] to generate synthetic abnormal MRI images with brain tumors. In this work, the authors argue that, in principle, their approach can be used to generate a completely artificial corpus where none of the scans can be attributed to actual patients. However, as the brain data is hallucinated, this method is not useful for our task.

The literature covering removal of privacy-sensitive information from image data largely focuses on de-identification of photographs of faces [14, 24]. Among these, *Deep Privacy* [9] is the closest to our approach as it was the first to suggest GANs to de-identify faces. It conditions on an *a priori* binary segmentation, guiding the generator to inpaint privacy-sensitive regions while preserving insensitive regions. Similar to our approach, *Deep Privacy* seeks to anonymize faces – but in 2D images. To identify face regions for conditional inpainting, it relies on an SSD detector [19]. We develop an alternative approach because a 3D analogue does not exist, and this allows us to comprehensively remodel interior regions of the head such as the neck and oral/nasal cavities. In particular, we define a convex hull enclosing the head and mask of the brain for conditioning. Finally, whereas *Deep Privacy* de-identifies conventional images of size 128×128, our goal is to generate much higher dimensional 3D volumes at 128^3 voxels – the equivalent of a 1448×1448 image.

3 Conditional De-Identification of 3D Images

Given a set of 3D images $(X^{(i)})_{i=1,\dots N} \overset{\text{i.i.d.}}{\sim} \mathcal{P}_X$ with values in $\mathcal{I}^{S\times S\times S}$ over some intensity space $\mathcal{I}\subset\mathbb{R}$, we are interested in finding a function of the form $Y=f_\Phi(\gamma(X))\sim\mathcal{P}_Y$ that maps a 3D image X to its de-identified counterpart Y. The task of the function $\gamma(X)$ is to filter out any sensitive information in order to make it impossible to infer the subject's identity given only Y; *i.e*. to create a *privacy preserving representation*. In this work we consider MRI data*, and choose $\gamma(X)$ to be a function of the convex hull of the head $c(X)\in\{0,1\}^{S\times S\times S}$ and the brain mask $b(X)\in\{0,1\}^{S\times S\times S}$.

Within this *remodeling*-based privacy framework, we impose three requirements: (i) *Anonymity*, *i.e*. $\gamma(X)$ is non-invertible; (ii) *Distribution preservation*, *i.e*. $\mathcal{P}_X$ and $\mathcal{P}_Y$ are stochastically indistinguishable and finally (iii) *Brain preservation*, *i.e*. $\forall(i,j,k): b(X)_{i,j,k} = 1 \implies X_{i,j,k} = f_\Phi(\gamma(X))_{i,j,k}$. In other words, we are interested in deriving a function f_Φ that maps some original scan X to some de-identified scan Y, while retaining medically relevant

*MRI scans can be acquired under different conditions, data availability limits us to common T1-weighted MRI.

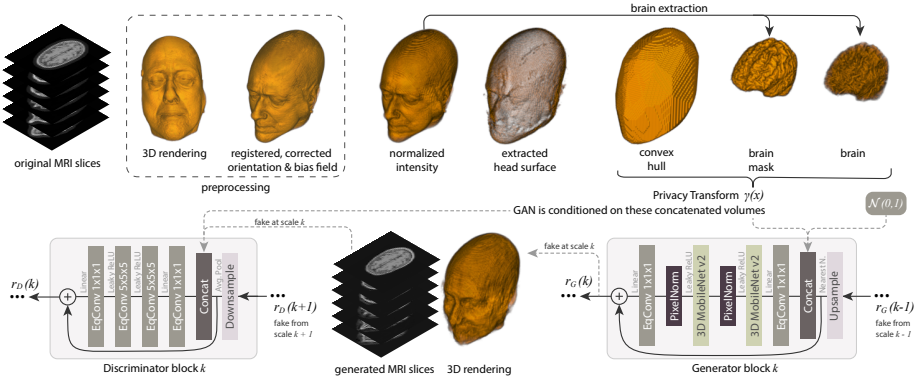

Figure 2: *Overview of our approach*: We apply standard preprocessing to register and correct the scans (illustrated OASIS-3). Then, using a novel technique, we construct a *head surface* representation from which we extract a convex hull $c(x)$, a brain mask $b(x)$, and the brain intensities $b(x) \circ x$. Combined, these form the privacy transform $\gamma(x)$ which serves as a conditioning variable of our model. CP-GAN learns to convert the distribution $\mathcal{P}_X$ of original MR scans to a de-identified counterpart $\mathcal{P}_Y$. Note that $\gamma(x)$ does not contain any privacy-sensitive information.

information (*e.g.* the brain) but preventing other information specific to X to leak into Y (*e.g.* the face). This makes it impossible to infer a person's identity from facial renderings. Figure 2 depicts the de-identification process, described below, including the privacy transform $\gamma(X)$ and the mapping function f_Φ implemented with a conditional multi-scale volumetric GAN.

3.1 The Privacy Transform $\gamma(X)$

The goal of the privacy transform is to non-invertibly change an individual MRI representation x into a form $\gamma(x)$ that removes detailed privacy-sensitive information and replaces it with a convex hull filled with 1's, smoothing away detailed face information (*e.g.* eyes, nose, and mouth). The transform guides the GAN, showing which regions should be hallucinated via a convex hull $c(x)$ and which regions should be retained through a brain mask $b(x)$. It also includes the brain data $b(x) \circ x$. The convex hull can afterwards be used to suppress noise patterns surrounding the head. Following the preprocessing (see *Appendix*), we define a function $c(x)$ that maps a scan $x \in \mathcal{I}^{S\times S\times S}$ to a binary convex hull volume of the same shape. As no efficient off-the-shelf algorithm exists, we developed a probabilistic solution that first constructs a surface representation from the MRI scan, and from this we compute the convex hull of the head. These steps are described below.

Surface Representation. To extract a surface representation Z from an MRI scan x, we compute maps where rays cast from each direction intersect the head at random rotations. We then rotate these measurements back to the reference coordinates and treat each as the probability of it belonging to the surface. The rotations are randomized to sample the subject from all sides uniformly. We begin by converting a given scan into a sequence of K *binarized* and *rotated* scans, *i.e.* $m^{(i)} = \mathrm{Rot}(\mathbb{1}[x \geq \delta]; R_i) \in \mathcal{I}^{S\times S\times S}$ for sampled rotations $R_1, \ldots, R_K \overset{\text{i.i.d}}{\sim} \mathcal{U}(\mathrm{SO}(3))$, where $\mathcal{U}(\mathrm{SO}(3))$ denotes the uniform distribution over all rotations in three-dimensional space and $\delta \in \mathcal{I}$ represents a suitably chosen binarization thresh-

old[†] for the binarization operator $\mathbb{1}[\cdot]$. Let us further introduce the concept of the $\zeta_{a,d}$-distance of some voxel at position (k_0,k_1,k_2) for some axis $a \in \{0,1,2\}$ and some direction $d \in \{-1,+1\}$:

$$\zeta_{a,d}(k_0,k_1,k_2) \hat{=} \begin{cases} (S-1)-k_a & \text{if } d=+1 \\ k_a & \text{otherwise.} \end{cases} \tag{1}$$

For fixed a and d, we can use this to create an *intersection map* $\Lambda^i_{a,d}$ for each binary image m^i_S:

$$\Lambda^{(i)}_{a,d}[k_0,k_1,k_2] = \mathbb{1}\left[\left(m^{(i)}_{k_0,k_1,k_2}=1\right) \wedge \left(\zeta_{a,d}(k_0,k_1,k_2) = \min_{\substack{s\in\{0,\dots,S-1\},\\ m^{(i)}_{k_{a-1}|s|k_{a+1}}=1}} \zeta_{a,d}(k_{a-1} \mid s \mid k_{a+1})\right)\right] \tag{2}$$

where $(k_{a-1} \mid s \mid k_{a+1})$ indicates that the a-th index is set to s and the two others to their associated value in k_0,k_1,k_2. We average the intersection map over all axis-direction combinations, *i.e.* $\Lambda^{(i)} = 1/6 \sum_a \sum_d \Lambda^i_{a,d}$. This process can be thought of as casting rays from each principle direction and recording the location of the intersection with the rotated, binarized head in $m^{(i)}$. Voxels on the surface of the head will exhibit high values of $\Lambda^{(i)}$. The final step is to back-rotate $\Lambda^{(1)},\dots,\Lambda^{(K)}$ to the reference coordinate system and average among the K randomly sampled rotations to create the *surface* representation:

$$Z = 1/K \sum_{i=1}^{K} Rot(\Lambda^{(i)}; R_i^{-1}) \in [0,1]^{S\times S\times S} \tag{3}$$

Note that Z is a random variable induced by the sampled rotations $R_1,\dots,R_K$. We interpret individual voxel values of Z as Bernoulli parameters characterizing the probability of some voxel belonging to the surface. This justifies binarizing Z by considering it as a three-dimensional Bernoulli tensor and sampling from it on a voxel-wise basis in the next step.

Convex Hull. From Z, we sample a set of non-zero indices and use Chan's Algorithm [2] to compute the triangles $\mathcal{T}$ making up the convex hull. We initialize a uniform volume filled with 1's, then randomly select a sufficient number of triangles (100 suffice) from $\mathcal{T}$. For each triangle, we find its corresponding hyperplane and the half-spaces within $c(x)$ defined by it. Voxels in the outward half-space of $c(x)$ are set to 0 while the rest are unchanged, yielding a binary convex hull volume.

Privacy Transform. The binary convex hull volume $c(x)$ instructs the GAN as to which regions should be hallucinated. A binary brain mask $b(x)$ obtained by applying [10] indicates which regions should be preserved. Together, these volumes along with the masked *continuous* values of the brain $b(x) \circ x$, are concatenated to make the privacy transform $\gamma(x)$. The GAN is conditioned on $\gamma(x)$ in the following subsection, as depicted in Figure 2.

3.2 Conditional De-identification GAN

The CP-GAN architecture depicted in Figure 2 is capable of generating volumes at multiple scales and passing gradients between each scale during training. We start from a 2D generation framework akin to MSG-GAN [15] and adapt it to our task by means of the following: *(1)* we incorporate *conditional* information via the privacy transform, *(2)* we make architectural improvements described below, *(3)* we use a new resampling strategy, *(4)* we adopt relativistic (non-averaging) R-LSGAN loss, and *(5)* we operate on 3D volumes. We use

[†]The threshold δ is chosen to be larger than the noise values surrounding the skull.

bottlenecks between scales as recently suggested by [17], in which the generator outputs single-channel maps instead of multi-channel maps. To reduce the memory footprint, we use modified MobileNetV2 convolutions as suggested in [8].

Both the generator $G_\Phi(\gamma(x))$ and the discriminator $D_\Theta(\gamma(x), v)$ are conditioned on $\gamma(x)$, where v either denotes a multi-resolutional original or fake sample. Regarding scales – suppose that S and s are powers of two that denote the maximum/minimum resolution synthesized by G_Φ. Then both G_Φ and D_Θ are defined to have $N_B = \log_2(S/s) + 1$ blocks (indexed by k) that either double (G_Φ) or halve (D_Θ) their input resolution. Here, we generate scales from $4\times4\times4$ to $128\times128\times128$.

Generator. The generator $G_\Phi = G_\Phi^{(N_B)} \circ \ldots \circ G_\Phi^{(1)}$ for $G_\Phi^{(k)} : \mathbb{R}^{r_G(k-1)} \times \mathbb{R}^{r_G(k)} \rightarrow \mathbb{R}^{r_G(k)}$ and $r_G(k) = 1 \times 2^{k-1}s \times 2^{k-1}s \times 2^{k-1}s$ synthesizes a sequence of fake images $g_1, \ldots, g_{N_B}$ of increasing resolutions as $g_k = G_\Phi^{(k)}(g_{k-1}, \gamma_k)$ for $k = 1, \ldots, N_B$ where $g_0 \sim \mathcal{N}(0,1)$ and $\gamma_k = \downarrow_{r_G(k)} \gamma(x)$ is $\gamma(x)$ downsampled to a resolution of $r_G(k)$.

Discriminator. The discriminator $D_\Theta = F \circ D_\Theta^{(N_B)} \circ \ldots \circ D_\Theta^{(1)}$ for $D_\Theta^{(1)} : \mathbb{R}^{r_D(1)} \times \mathbb{R}^{r_D(1)} \rightarrow \mathbb{R}^{r_D(1)}$ resp. $D_\Theta^{(k)} : \mathbb{R}^{r_D(k-1)} \times \mathbb{R}^{r_D(k)} \times \mathbb{R}^{r_D(k)} \rightarrow \mathbb{R}^{r_D(k)} (k > 1)$ and $r_D(k) = 1 \times S/2^{k-1} \times S/2^{k-1} \times S/2^{k-1}$ assigns a *scalar* to a sequence of images[‡] of decreasing resolutions $v_1, \ldots, v_{N_B}$ as $d_1 = D_\Theta(v_1, \gamma_1)$ resp. $d_k = D_\Theta(d_{k-1}, v_k, \gamma_k)$ for $k = 2, \ldots, N_B$ where $\gamma_k = \downarrow_{r_D(k)} \gamma(x)$ is $\gamma(x)$ downsampled to a resolution of $r_D(k)$ and F is a fully-connected layer that computes a scalar summary of the output of $D_\Theta^{(N_B)}$.

Resampling blocks. [16, 17] recently proposed to use bilinear interpolation for downsampling, but adapting this approach is problematic as it will create undesirable interpolation effects in the binary volumes. Therefore, we suggest a *probabilistic* interpretation of *average pooling* which guarantees that the proportion of non-zero voxels is preserved (in expectation) while maintaining voxel-wise correspondence to conventional average pooling performed on non-binary images. Specifics can be found in the *Appendix*.

Loss Function. We use the relativistic (*non-averaging*) *R-LSGAN* loss [12]: We opt for relativistic losses as they induce a lower memory footprint than, for instance, the widely-established WGAN-GP [6] requiring an additional forward/backward pass.

Brain Preservation. One of the requirements defined above in the Problem Definition is to perfectly preserve medically relevant information. Therefore, in a similar process to image inpainting in which original image content is masked and retained, we use the brain mask $b(x)$ to embed the original brain data into the volume synthesized by the generator, *i.e.* $f_\Phi(\gamma(x)) = b(x) \circ x + (1 - b(x)) \circ G_\Phi(\gamma(x))$ where $\circ$ denotes the *Hadamard* product.

4 Experiments

Above, we proposed a new and modern approach to de-identify medical image data. To judge its utility, we must address the following questions: *(1) Does remodeling preserve privacy better than existing removal-based de-identification methods? (2) Does our approach adversely affect the performance of common medical applications?* Below, we compare our approach to other de-identification methods to answer these questions experimentally.

[‡] $x_1, \ldots, x_{N_B}$ in case of an original image and $g_1, \ldots, g_{N_B}$ in case of a fake image

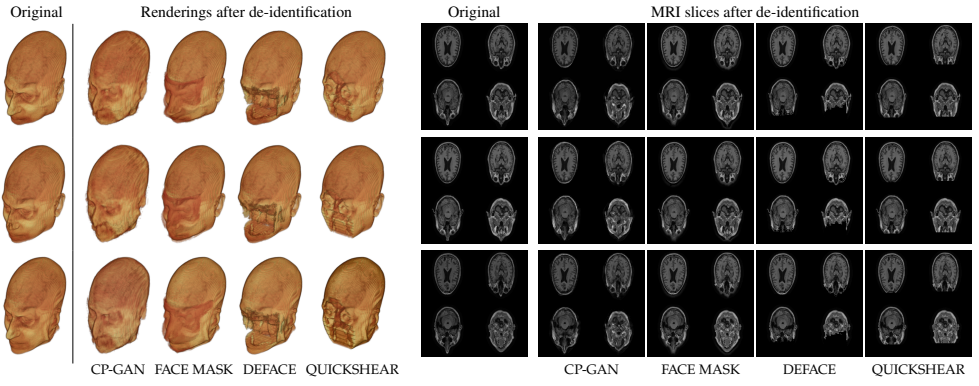

Figure 3: *De-identification examples*: De-identified face renderings *(left)* and the corresponding MRI scans *(right)* for several examples from OASIS-3. All methods perfectly preserve the brain, but only CP-GAN successfully de-identifies the patient while maintaining realistic appearance and structure.

4.1 Setup

Datasets. In this work, we use two standard, publicly available large-scale Alzheimer's disease imaging studies which feature T1-weighted volumetric MR scans of the head for each subject: A selection of 2,172 MRIs from ADNI [32, 33] and 2,168 MRIs from OASIS-3 [18]. Both datasets are split (80%-20% train-test) on a patient level to avoid data leakage by memorizing the patient. The ADNI data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative database (adni.loni.usc.edu). 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 magnetic resonance imaging (MRI), positron emission tomography (PET), 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). For up-to-date information, see www.adni-info.org. Scanner types and acquisition protocols differ between and within the datasets, details can be found in the *Appendix*.

Benchmark De-Identification Methods. We compare our result with three publicly available and widely-established methods for de-identification of MRI head scans, depicted in Figure 3. All methods have in common that they *(1)* are not deep-learning-driven, *(2)* require no additional training and *(3)*, are used on a day-to-day basis in neuroscience and clinical research. All procedures were applied with default settings on images of resolution $128\times128\times128$. The methods include QUICKSHEAR [25], FACE MASK [23], and DEFACE [1]. Descriptions of the methods are provided in the *Appendix*. We also include MRI WATERSHED [21], a skull-stripping method that removes everything except the brain.

Training. We use the AdamP [7] optimizer with a learning rate of $2\cdot10^{-3}$ and $\beta=(0,0.99)$ and a batch size of 2. See the *Appendix* for a complete list of hyperparameters.

4.2 Results

In this section, we present results on (1) studies comparing the identification rate of our model with existing de-identification methods, and (2) the effects of de-identification on common medical image analysis tasks. In the *Appendix*, we provide a comparison of execu-

	USER-BASED		MODEL-BASED	
	OASIS-3	ADNI	OASIS-3	ADNI
ORIGINAL	64.3 ±2.1	63.2 ±2.1	100.0 ±0.0	100.0 ±0.0
BLURRED	57.2 ±2.1	53.1 ±2.2	46.7 ±8.2	51.2 ±7.8
FACE MASK	48.7 ±2.3	47.5 ±2.2	68.5 ±7.6	63.3 ±8.2
DEFACE	54.0 ±2.3	46.7 ±2.3	70.7 ±6.4	73.1 ±7.0
QUICKSHEAR	45.6 ±2.3	39.9 ±2.2	44.0 ±7.4	47.0 ±6.7
CP-GAN	**28.3 ±2.0**	**22.4 ±1.8**	**29.5 ±5.9**	**27.4 ±5.7**
BLACK	18.9 ±1.6	20.5 ±1.8	19.7 ±4.7	19.8 ±4.4
MRI WATERSHED	20.7 ±2.0	18.7 ±1.6	21.0 ±4.8	21.6 ±4.4

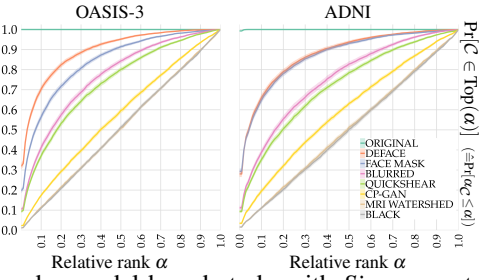

Figure 4: (*left*) A user-based study on AMT and a model-based study with *Siamese* networks to determine how well de-identified faces can be identified. Correct identification rates (±s.d.) are reported for two datasets (lower is better); (*right*) A Siamese network takes a face rendering as an input, and aims to identify the correct de-identified rendering in a retrieval scenario. We measure the probability of the correct face appearing in its top α-percentile ranked predictions. BLACK = *optimal*, ORIGINAL = *pessimal*.

tion times. Video results are provided in *supplementary material.*

De-identification quality user study. The privacy attack described in [21] relied on prospectively collected data, meaning the authors had access to CT scans as well as photographs of patient faces. Replicating that study for MRI scans is impossible, because photographs of ADNI and OASIS-3 patients do not exist. Therefore, we conduct a similarly-spirited study using *Amazon Mechanical Turk* in which workers are asked to defeat the various de-identification methods given renderings of MRI scans. Workers were presented with an unaltered rendering of a query patient along with five renderings de-identified using a single method[§] – one of which is a de-identified rendering of the query patient. The task was then to pick out the de-identified rendering which corresponds to the unaltered query rendering. We considered the following de-identification methods: QUICKSHEAR [25], FACE MASK [23], and DEFACE [1], and CP-GAN (ours). In addition to the four de-identification methods, we added four control tasks, ORIGINAL, which signifies the absence of any de-identification scheme, BLURRED, in which the 2D renderings are blurred to mildly obscure the patient identity, BLACK, which features the same all-black image for each option, and MRI WATERSHED [27] which completely removes all tissue except the brain. We asked 800 distinct questions per dataset. Each question was given to five workers, for a total of 4,000 assignments. The mean and the standard deviation are estimated by bootstrapping over 1,000 resamples.

In Figure 4 (*left*), we report the identification rate, or how often the workers were able to defeat each method, see *Appendix* for details. The upper performance bound from random guessing corresponds to 20%. The results substantiate the claim that CP-GAN performs extraordinarily well at de-identification. Our model outperforms the other de-identification methods by gaps of 17%–25% on both datasets. We note that for both datasets, CP-GAN performs close to the theoretical optimum of 20%.

De-identification quality model-based study. In a similar fashion to the last experiment, we assess the de-identification performance of the various models by attempting to defeat them. This time, however, we leverage a neural network to assess similarity.

To this end, we use a metric learning approach to train a Siamese network $\mathcal{S}(\cdot,\cdot)$ to quantify whether its two input renderings belong to the same patient or not.

Given two inputs x and y, the network $\mathcal{S}$ is constructed by applying a sub-network $\tilde{\mathcal{S}}$

[§] An exemplary question can be found in the *Appendix*

	Sørensen-Dice coefficient ↑								*Intersection-over-Union* (IoU) ↑							
	OASIS-3				ADNI				OASIS-3				ADNI			
	BRAIN	VCSF	WHITE	GREY	BRAIN	VCSF	WHITE	GREY	BRAIN	VCSF	WHITE	GREY	BRAIN	VCSF	WHITE	GREY
ORIGINAL	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
FACE MASK	0.991	0.984	0.989	0.996	0.986	0.977	0.976	0.987	0.982	0.968	0.978	0.992	0.973	0.955	0.953	0.975
DEFACE	0.993	0.986	0.986	0.995	0.982	0.966	0.965	0.981	0.985	0.972	0.973	0.990	0.965	0.934	0.932	0.963
QUICKSHEAR	0.994	0.989	0.990	0.997	0.986	0.975	0.972	0.985	0.987	0.978	0.980	0.994	0.972	0.952	0.946	0.971
CP-GAN	**0.995**	**0.991**	**0.992**	**0.998**	**0.989**	**0.979**	**0.978**	**0.989**	**0.989**	**0.981**	**0.983**	**0.996**	**0.977**	**0.960**	**0.957**	**0.978**
MRI WATERSHED	0.675	0.415	0.564	0.718	0.717	0.570	0.589	0.732	0.509	0.262	0.393	0.560	0.559	0.399	0.417	0.578

Table 1: *Brain segmentation:* We measure the Sørensen-Dice coefficient and the IoU computed between segmentations on the original scan and de-identified scans using standard software, SIENAX. We test on the whole brain, VCSF, white matter, and grey matter. Ideally, segmentation should not be affected by de-identification, indicated by a Dice score and IoU of 1. CP-GAN outperforms all other de-identification methods. Note that MRI WATERSHED, which removes everything but the brain, has a catastrophic effect.

(conv. block/flatten/fully-connected layer) on x and y independently, followed by summarizing both embeddings with the *Euclidean* distance, *i.e.* $\mathcal{S}(x,y) = \|\tilde{\mathcal{S}}(x) - \tilde{\mathcal{S}}(y)\|_2$. We use the *Triplet Margin* loss function as described in [31], choosing the margin to be equal to 5. We split the previously defined (hold-out) data set, and *randomly* select 80% of its patients as a training set $\mathcal{D}_i$ and the remaining 20% as its complement $\overline{\mathcal{D}_i}$, where $i = 1,\ldots,100$ denotes the i-th (resampled) fold. Specifics on the nature of the training can be found in the *Appendix*.

In Figure 4 (*left*) we report the ability of the network to defeat the de-identification methods in similar fashion to the *user-based* study. We first sample a patient p from $\overline{\mathcal{D}_i}$ from whom, in turn, we sample a scan $s \in S(p)$. Afterwards, we sample a method m and and consider the m-rendering of s to be the *correct* option $\mathcal{C}$. The remaining $5-1=4$ options are obtained by randomly selecting m-renderings from other patients' scans. Denoting the five options by $x_1,\ldots,x_5$, we obtain the predicted option by $k = \arg\min_{j=1,\ldots,5} \mathcal{S}(x_O, x_j)$ where x_O denotes the original rendering of s. As in the *user-based* study, CP-GAN outperforms the other de-identification methods. For FACE MASK and DEFACE, the network was able to de-identify between 16 to 20% more renderings than its human counterparts. The effect on QUICKSHEAR is more moderate.

In Figure 4 (*right*) we evaluate the Siamese network's ability to defeat de-identification methods in a retrieval-inspired setting. We observe that CP-GAN's de-identification capabilities are strikingly close to the *optimal* case while other methods perform substantially worse. For some given original rendering $x^{(\text{orig.})}$ and some method m, we analyze how often the correct choice $\mathcal{C}$ falls within the top α (relative) ranks, *e.g.* Top(α=0.1) is a subset of the 10% top-ranked scans. An *optimal* de-identification method induces a *uniformly random* rank $\alpha_{\mathcal{C}} \sim \mathcal{U}(0,1)$ of $\mathcal{C}$ (c.f. BLACK), whereas a *pessimal* method induces a *Dirac* placement $\alpha_{\mathcal{C}} = 0$ (c.f. ORIGINAL). Confidence bands (CI=0.95) are calculated over the 100 previously defined data splits $\overline{\mathcal{D}_i}$.

Overall, we conclude that CP-GAN outperforms established methods by a substantial double-digit margin, withstanding both human and model-based attacks. We note a gap between CP-GAN and optimal performance, which can be explained by a property of the model: it preserves head size. The attacker may exploit this to eliminate some candidates. We explore this concept further in the *Appendix*.

Effect of De-Identification on Medical Analyses. Beyond ensuring patient privacy, de-identification methods should not adversely affect software tools commonly used on medical scans. However, it has been shown that facial de-identification methods *do* adversely impact automated image analysis on MRI scans used in research and in the clinic [3]. In line with

this study, we conduct two experiments. In the first, we assess how the de facto standard brain tissue segmentation tool, SIENAX [29], performs on de-identified MRI scans in comparison to the originals. In Table 1, we report the (Sørensen-) Dice scores [4, 30] and IoU between the original and de-identified scans for various brain segmentation tasks. We observe that CP-GAN outperforms all of its contenders, proving that brain volume estimations are reliable after the subject is de-identified using CP-GAN. *Note that removing everything except the brain using* MRI WATERSHED *has a catastrophic effect*, replicating the effect observed in [6].

In the second experiment, we investigate whether de-identification adversely affects brain age estimation – an important task as the difference between predicted and chronological age has links to brain disease [13]. This is a challenging task for CP-GAN since age information captured in the MRI is filtered out in $\gamma(x)$ in contrast to the other methods that preserve head information in addition to the brain. Nonetheless, we find that our de-identification introduces less bias than DEFACE and QUICKSHEAR on both datasets, though FACE MASK slightly outperforms our model. Due to space limitations, these results appear in the *Appendix*.

5 Conclusion

In this work, we defined a new paradigm for de-identification of medical imagery and realized it for MRI scans. Our approach remodels privacy-relevant information while keeping medically-relevant information untouched. *It can be applied to other modalities, producing remodeled images that appear genuine and preserve relevant medical information, but without revealing privacy-sensitive information.* Our method protects privacy substantially better than existing methods, without compromising analyses typically found in research and clinical settings – a crucial deficiency of strong removal methods such as skull-stripping. A future research direction is to extend our approach to other MRI and CT modalities, adding new downstream tasks such as lesion and brain tumor segmentation [22, 26]. Incorporating other pulse sequences, such as *T2-weighting* or *FLAIR*, requires a re-training of the network as well as a sufficient number of training samples. Unfortunately, alternative pulse sequences are less readily available in comparison to the *T1-weighted* imagery used in this paper. Apart from this apparent scarcity, we however do not expect any change in complexity. We hope that the methods outlined here can help to better protect patient privacy.

Acknowledgements *This work was partially supported by the Swedish Research Council (VR) 2017-04609, the ERC (853489 - DEXIM) and by the DFG (2064/1 – Project number 390727645). 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.*

References

[1] Amanda Bischoff-Grethe, I. Burak Ozyurt, Evelina Busa, Brian T. Quinn, Christine Fennema-Notestine, Camellia P. Clark, Shaunna Morris, Mark W. Bondi, Terry L. Jernigan, Anders M. Dale, Gregory G. Brown, and Bruce Fischl. A technique for the deidentification of structural brain MR images. *Human Brain Mapping*, 28(9): 892–903, sep 2007. ISSN 10659471. doi: 10.1002/hbm.20312.

[2] T. M. Chan. Optimal output-sensitive convex hull algorithms in two and three dimensions. *Discrete and Computational Geometry*, 16(4):361–368, 1996. ISSN 01795376. doi: 10.1007/BF02712873.

[3] A. De Sitter, M. Visser, I. Brouwer, K. S. Cover, R. A. van Schijndel, R. S. Eijgelaar, D. M.J. Müller, S. Ropele, L. Kappos, Rovira, M. Filippi, C. Enzinger, J. Frederiksen, O. Ciccarelli, C. R.G. Guttmann, M. P. Wattjes, M. G. Witte, P. C. de Witt Hamer, F. Barkhof, and H. Vrenken. Facing privacy in neuroimaging: removing facial features degrades performance of image analysis methods. *European Radiology*, 30(2):1062–1074, feb 2020. ISSN 14321084. doi: 10.1007/s00330-019-06459-3.

[4] Lee R. Dice. Measures of the Amount of Ecologic Association Between Species. *Ecology*, 26(3):297–302, jul 1945. ISSN 00129658. doi: 10.2307/1932409. URL http://doi.wiley.com/10.2307/1932409.

[5] Christine Fennema-Notestine, I. Burak Ozyurt, Camellia P. Clark, Shaunna Morris, Amanda Bischoff-Grethe, Mark W. Bondi, Terry L. Jernigan, Bruce Fischl, Florent Segonne, David W. Shattuck, Richard M. Leahy, David E. Rex, Arthur W. Toga, Kelly H. Zou, and Gregory G. Brown. Quantitative evaluation of automated skull-stripping methods applied to contemporary and legacy images: Effects of diagnosis, bias correction, and slice location. *Human Brain Mapping*, 27(2):99–113, 2006. doi: https://doi.org/10.1002/hbm.20161. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/hbm.20161.

[6] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. In *Advances in neural information processing systems*, pages 5767–5777, 2017.

[7] Byeongho Heo, Sanghyuk Chun, Seong Joon Oh, Dongyoon Han, Sangdoo Yun, Youngjung Uh, and Jung-Woo Ha. Slowing down the weight norm increase in momentum-based optimizers. *arXiv preprint arXiv:2006.08217*, 2020.

[8] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *CoRR*, abs/1704.04861, 2017. URL http://arxiv.org/abs/1704.04861.

[9] Håkon Hukkelås, Rudolf Mester, and Frank Lindseth. DeepPrivacy: A Generative Adversarial Network for Face Anonymization. sep 2019. URL http://arxiv.org/abs/1909.04538.

[10] Juan Eugenio Iglesias, Cheng Yi Liu, Paul M. Thompson, and Zhuowen Tu. Robust brain extraction across datasets and comparison with publicly available methods. *IEEE*

Transactions on Medical Imaging, 30(9):1617–1634, sep 2011. ISSN 02780062. doi: 10.1109/TMI.2011.2138152.

[11] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-Image Translation with Conditional Adversarial Networks. *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017-January:5967–5976, nov 2016. URL http://arxiv.org/abs/1611.07004.

[12] Alexia Jolicoeur-Martineau. The relativistic discriminator: a key element missing from standard GAN. *7th International Conference on Learning Representations, ICLR 2019*, jul 2018. URL http://arxiv.org/abs/1807.00734.

[13] Benedikt Atli Jónsson, Gyda Bjornsdottir, TE Thorgeirsson, Lotta María Ellingsen, G Bragi Walters, DF Gudbjartsson, Hreinn Stefansson, Kari Stefansson, and MO Ulfarsson. Brain age prediction using deep learning uncovers associated sequence variants. *Nature communications*, 10(1):1–10, 2019.

[14] Amin Jourabloo, Xi Yin, and Xiaoming Liu. Attribute preserved face de-identification. In *2015 International conference on biometrics (ICB)*, pages 278–285. IEEE, 2015.

[15] Animesh Karnewar and Oliver Wang. MSG-GAN: Multi-Scale Gradient GAN for Stable Image Synthesis. mar 2019. URL http://arxiv.org/abs/1903.06048.

[16] Tero Karras, Samuli Laine, and Timo Aila. A Style-Based Generator Architecture for Generative Adversarial Networks. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2019-June:4396–4405, dec 2018. URL http://arxiv.org/abs/1812.04948.

[17] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and Improving the Image Quality of StyleGAN. dec 2019. URL http://arxiv.org/abs/1912.04958.

[18] Pamela J LaMontagne, Tammie L.S. Benzinger, John C. Morris, Sarah Keefe, Russ Hornbeck, Chengjie Xiong, Elizabeth Grant, Jason Hassenstab, Krista Moulder, Andrei Vlassenko, Marcus E. Raichle, Carlos Cruchaga, and Daniel Marcus. OASIS-3: Longitudinal Neuroimaging, Clinical, and Cognitive Dataset for Normal Aging and Alzheimer Disease. *medRxiv*, page 2019.12.13.19014902, dec 2019. doi: 10.1101/2019.12.13.19014902.

[19] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. SSD: Single Shot MultiBox Detector. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 9905 LNCS:21–37, dec 2015. doi: 10.1007/978-3-319-46448-0_2. URL http://arxiv.org/abs/1512.02325http://dx.doi.org/10.1007/978-3-319-46448-0{_}2.

[20] Eyal Lotan, Charlotte Tschider, Daniel K Sodickson, Arthur L Caplan, Mary Bruno, Ben Zhang, and Yvonne W Lui. Medical imaging and privacy in the era of artificial intelligence: Myth, fallacy, and the future. *Journal of the American College of Radiology*, 17(9):1159–1162, 2020.

[21] Jan C Mazura, Krishna Juluru, Joseph J Chen, Tara A Morgan, Majnu John, and Eliot L Siegel. Facial recognition software success rates for the identification of 3d surface reconstructed facial images: implications for patient privacy and security. *Journal of digital imaging*, 25(3):347–351, 2012.

[22] Raphael Meier, Urspeter Knecht, Tina Loosli, Stefan Bauer, Johannes Slotboom, Roland Wiest, and Mauricio Reyes. Clinical evaluation of a fully-automatic segmentation method for longitudinal brain tumor volumetry. *Scientific Reports*, 6:23376, 03 2016. doi: 10.1038/srep23376.

[23] Mikhail Milchenko and Daniel Marcus. Obscuring surface anatomy in volumetric imaging data. *Neuroinformatics*, 11(1):65–75, jan 2013. ISSN 15392791. doi: 10.1007/s12021-012-9160-3. URL http://www.ncbi.nlm.nih.gov/pubmed/22968671http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3538950.

[24] Elaine M Newton, Latanya Sweeney, and Bradley Malin. Preserving privacy by de-identifying face images. *IEEE transactions on Knowledge and Data Engineering*, 17 (2):232–243, 2005.

[25] Nakeisha Schimke, Mary Kuehler, and John Hale. Preserving privacy in structural neuroimages. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 6818 LNCS, pages 301–308. Springer, Berlin, Heidelberg, 2011. ISBN 9783642223471. doi: 10.1007/978-3-642-22348-8_26.

[26] Paul Schmidt, Christian Gaser, Milan Arsic, Dorothea Buck, Annette Förschler, Achim Berthele, Muna Hoshi, Rüdiger Ilg, Volker J. Schmid, Claus Zimmer, Bernhard Hemmer, and Mark Mühlau. An automated tool for detection of flair-hyperintense white-matter lesions in multiple sclerosis. *NeuroImage*, 59(4):3774 – 3783, 2012. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2011.11.032. URL http://www.sciencedirect.com/science/article/pii/S1053811911013139.

[27] F. Ségonne, A. M. Dale, E. Busa, M. Glessner, D. Salat, H. K. Hahn, and B. Fischl. A hybrid approach to the skull stripping problem in MRI. *NeuroImage*, 22(3):1060–1075, jul 2004. ISSN 10538119. doi: 10.1016/j.neuroimage.2004.03.032.

[28] Hoo-Chang Shin, Neil A Tenenholtz, Jameson K Rogers, Christopher G Schwarz, Matthew L Senjem, Jeffrey L Gunter, Katherine Andriole, and Mark Michalski. Medical Image Synthesis for Data Augmentation and Anonymization using Generative Adversarial Networks. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 11037 LNCS: 1–11, jul 2018. URL http://arxiv.org/abs/1807.10225.

[29] Stephen M. Smith, Mark Jenkinson, Mark W. Woolrich, Christian F. Beckmann, Timothy E.J. Behrens, Heidi Johansen-Berg, Peter R. Bannister, Marilena De Luca, Ivana Drobnjak, David E. Flitney, Rami K. Niazy, James Saunders, John Vickers, Yongyue Zhang, Nicola De Stefano, J. Michael Brady, and Paul M. Matthews. Advances in functional and structural MR image analysis and implementation as FSL. In *NeuroImage*, volume 23, pages S208–S219. Academic Press, jan 2004. doi: 10.1016/j.neuroimage.2004.07.051.

[30] T.J. Sørensen. *A Method of Establishing Groups of Equal Amplitude in Plant Sociology Based on Similarity of Species Content and Its Application to Analyses of the Vegetation on Danish Commons*. Biologiske skrifter. I kommission hos E. Munksgaard, 1948. URL https://books.google.de/books?id=rpS8GAAACAAJ.

[31] Daniel Ponsa Vassileios Balntas, Edgar Riba and Krystian Mikolajczyk. Learning local feature descriptors with triplets and shallow convolutional neural networks. In Edwin R. Hancock Richard C. Wilson and William A. P. Smith, editors, *Proceedings of the British Machine Vision Conference (BMVC)*, pages 119.1–119.11. BMVA Press, September 2016. ISBN 1-901725-59-6. doi: 10.5244/C.30.119. URL https://dx.doi.org/10.5244/C.30.119.

[32] Michael W. Weiner, Dallas P. Veitch, Paul S. Aisen, Laurel A. Beckett, Nigel J. Cairns, Robert C. Green, Danielle Harvey, Clifford R. Jack, William Jagust, John C. Morris, Ronald C. Petersen, Jennifer Salazar, Andrew J. Saykin, Leslie M. Shaw, Arthur W. Toga, and John Q. Trojanowski. The Alzheimer's Disease Neuroimaging Initiative 3: Continued innovation for clinical trial improvement, may 2017. ISSN 15525279. URL http://www.ncbi.nlm.nih.gov/pubmed/27931796http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5536850.

[33] Bradley T. Wyman, Danielle J. Harvey, Karen Crawford, Matt A. Bernstein, Owen Carmichael, Patricia E. Cole, Paul K. Crane, Charles Decarli, Nick C. Fox, Jeffrey L. Gunter, Derek Hill, Ronald J. Killiany, Chahin Pachai, Adam J. Schwarz, Norbert Schuff, Matthew L. Senjem, Joyce Suhy, Paul M. Thompson, Michael Weiner, and Clifford R. Jack. Standardization of analysis sets for reporting results from ADNI MRI data, may 2013. ISSN 15525279.