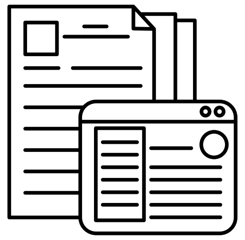
Beyond the Mask: The illusion of privacy in defaced brain MRI

Imen Bajar, Mohamed Maouche, Carole Frindel

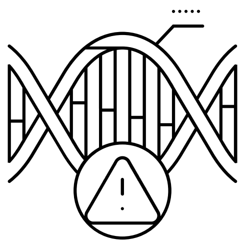
ISBI 2026 IEEE International Symposium on Biomedical Imaging

ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
Université du Québec
ÉTS Engineering for Industry
Inria

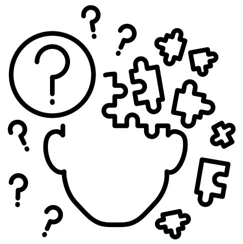
Why Medical Data Is Sensitive ?



- Personal identifiers (age, sex, ethnicity)

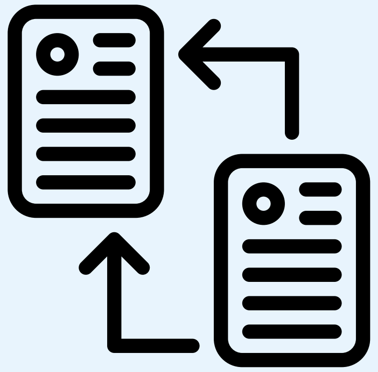


- Genetic disorders & predispositions



- Private conditions & vulnerabilities

Context



More Data Shared

Medical data
digitized & shared

More Privacy Risks

- Data breaches
- Cyber attacks

More sharing → More exposure → More re-identification risk

Context

- Hospitals face cybersecurity attacks leading to data breaches

French hospital suspends operations after cyber attacks

FRANCE

A hospital in Versailles, near Paris had to cancel operations and transfer some patients after being hit by a cyberattack over the weekend, France's health ministry said Sunday.

https://www.france24.com/en/france/20221205-french-hospital-suspends-operations-after-cyber-attacks

Institution	TRICARE	CHS Community Health Systems, Inc.	UCLA	…	l'Assurance Maladie Agir ensemble, protéger chacun
Date	09/2011	06/2014	07/2015		03/2022
N° Patients impacted	5 millions	4.5 millions	4.5 millions		510,000

https://www.upguard.com/blog/biggest-data-breaches-in-healthcare

Context

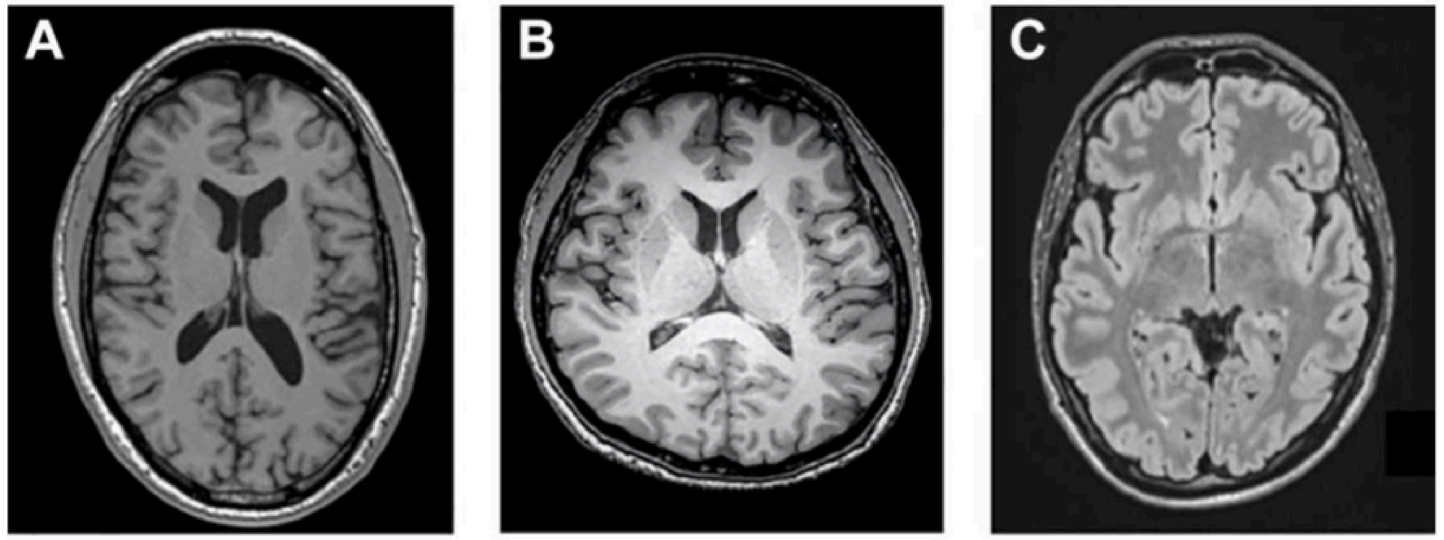
How vulnerable are medical images

and

can they pose a privacy threat?

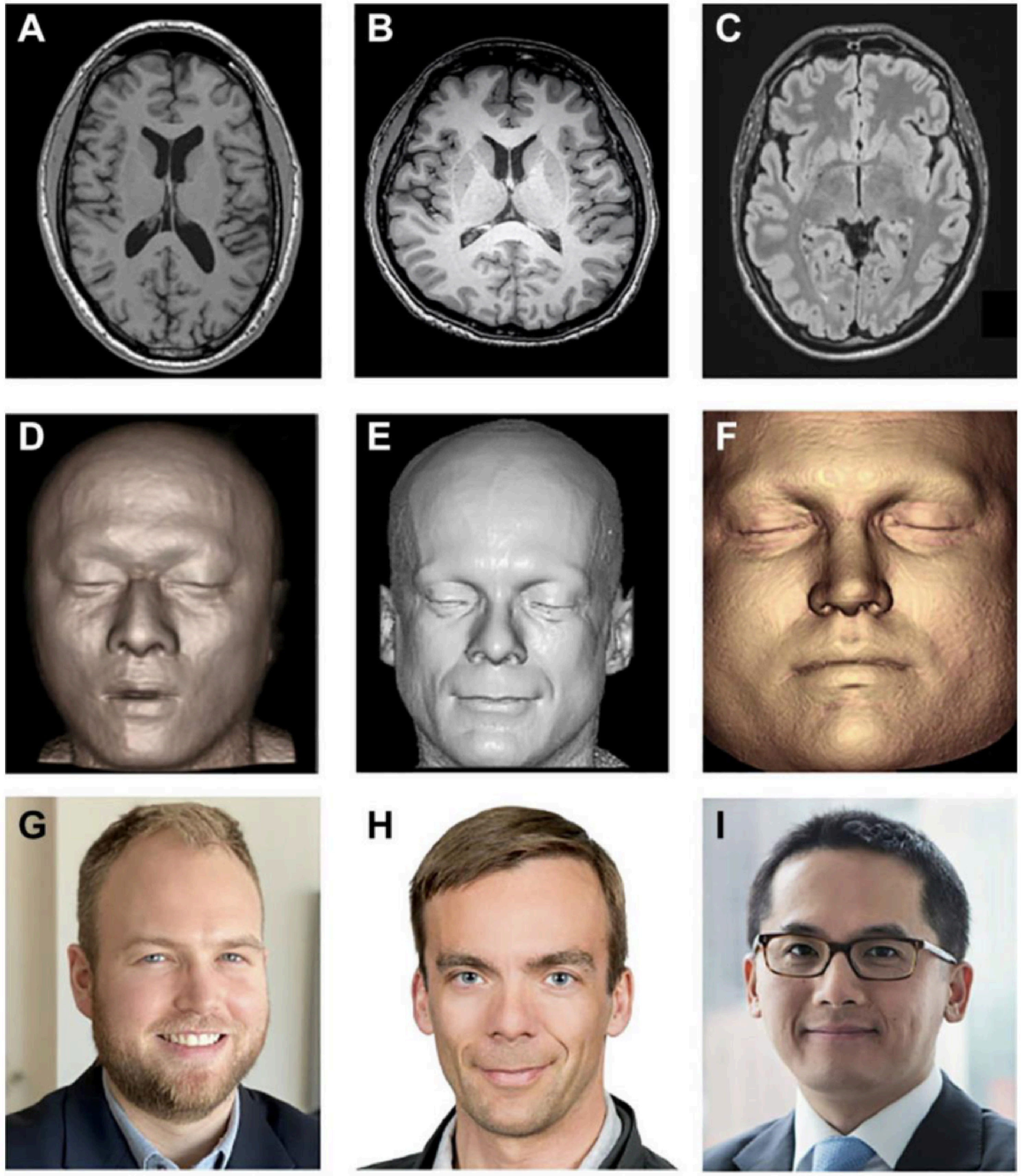
More sharing → More exposure → More re-identification risk

Can You Match These MRI Scans to These People?



Parker, William, et al. "Canadian association of radiologists white paper on de-identification of medical imaging: part 2, practical considerations." Canadian Association of Radiologists Journal 72.1 (2021): 25-34.

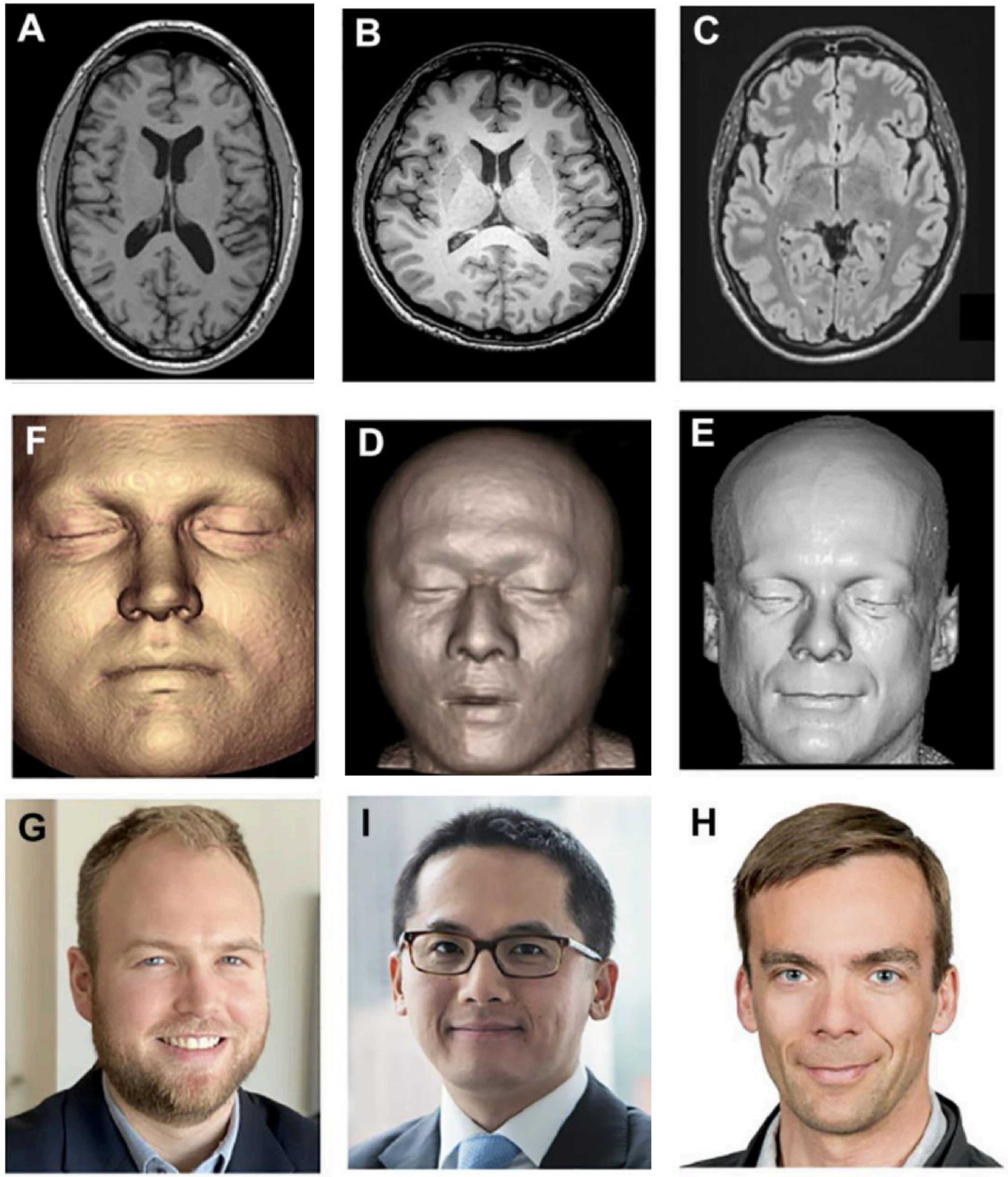
Adding the Missing Step: Reconstructed Faces



What about now?

Parker, William, et al. "Canadian association of radiologists white paper on de-identification of medical imaging: part 2, practical considerations." Canadian Association of Radiologists Journal 72.1 (2021): 25-34.

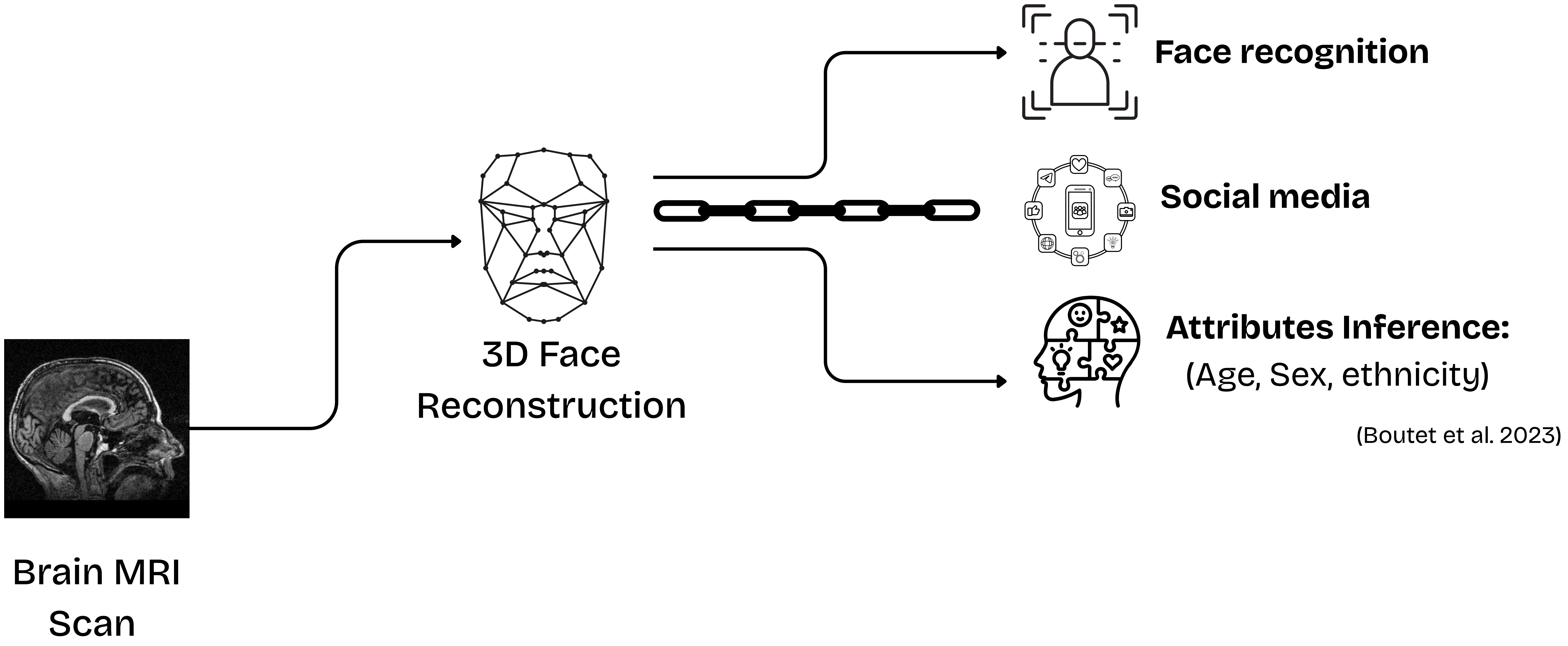
Face Reconstruction Changes the Problem



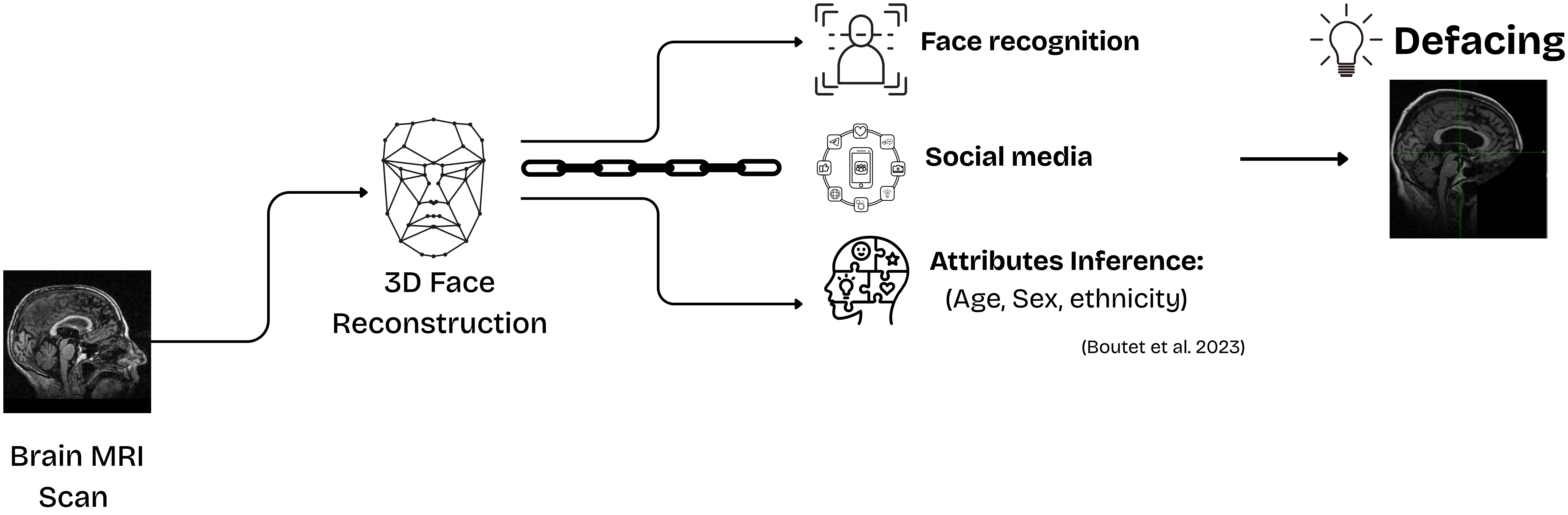
Key message:
A whole-head MRI may not look identifying, but it can be **transformed into a recognizable facial representation**

Parker, William, et al. "Canadian association of radiologists white paper on de-identification of medical imaging: part 2, practical considerations." Canadian Association of Radiologists Journal 72.1 (2021): 25-34.

Context

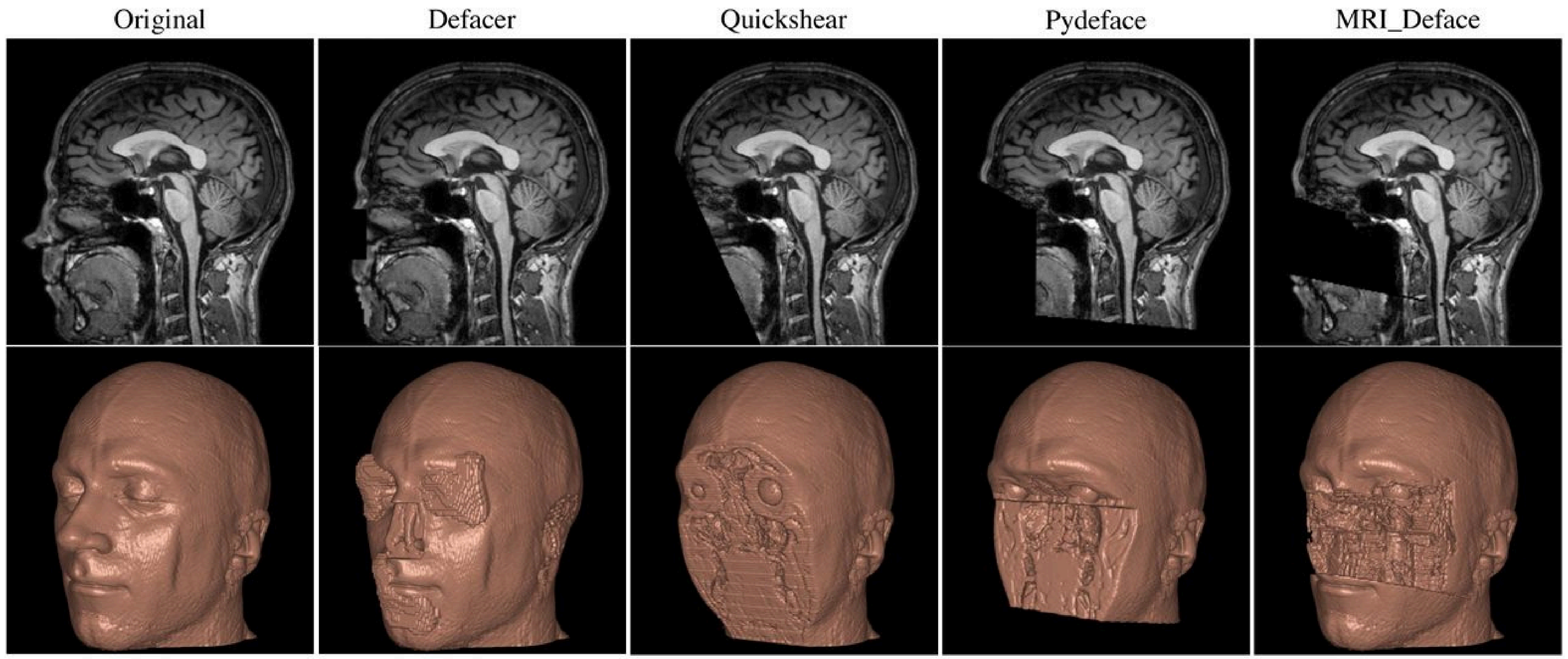

Face recognition
Social media
Attributes Inference:
(Age, Sex, ethnicity)
(Boutet et al. 2023)
3D Face
Reconstruction
Brain MRI
Scan

Context


Face recognition
Defacing
Social media
Attributes Inference:
(Age, Sex, ethnicity)
3D Face
Reconstruction
Brain MRI
Scan
(Boutet et al. 2023)

What is defacing ?

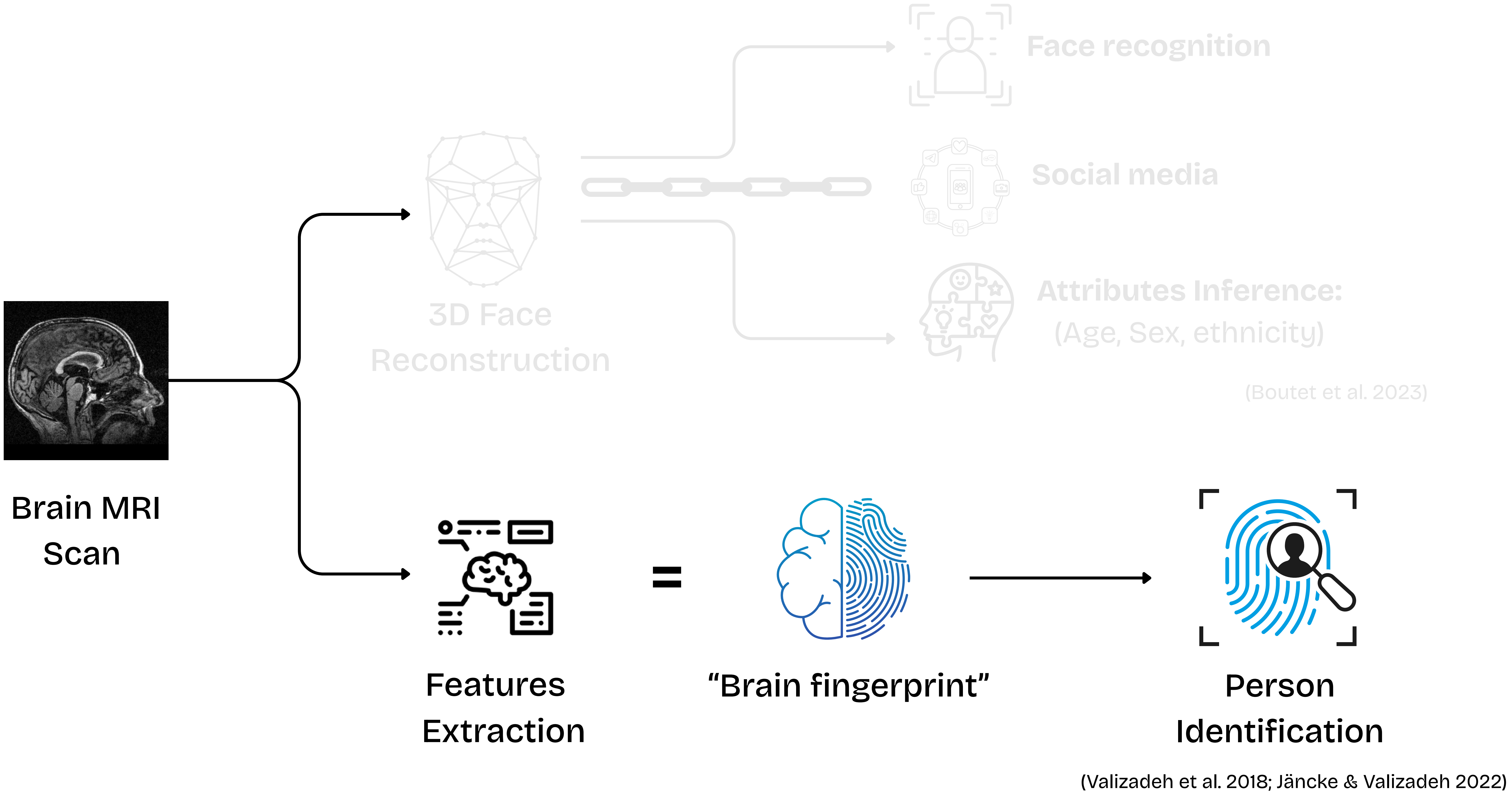
- Automated removal , blurring or masking of facial features from brain MRI scans
- Goal: Prevent face-based re-identification attacks on MRIs



A sagittal slice of an MRI is displayed over the reconstruction of the whole-head MRI.(Gao et al., 2023)

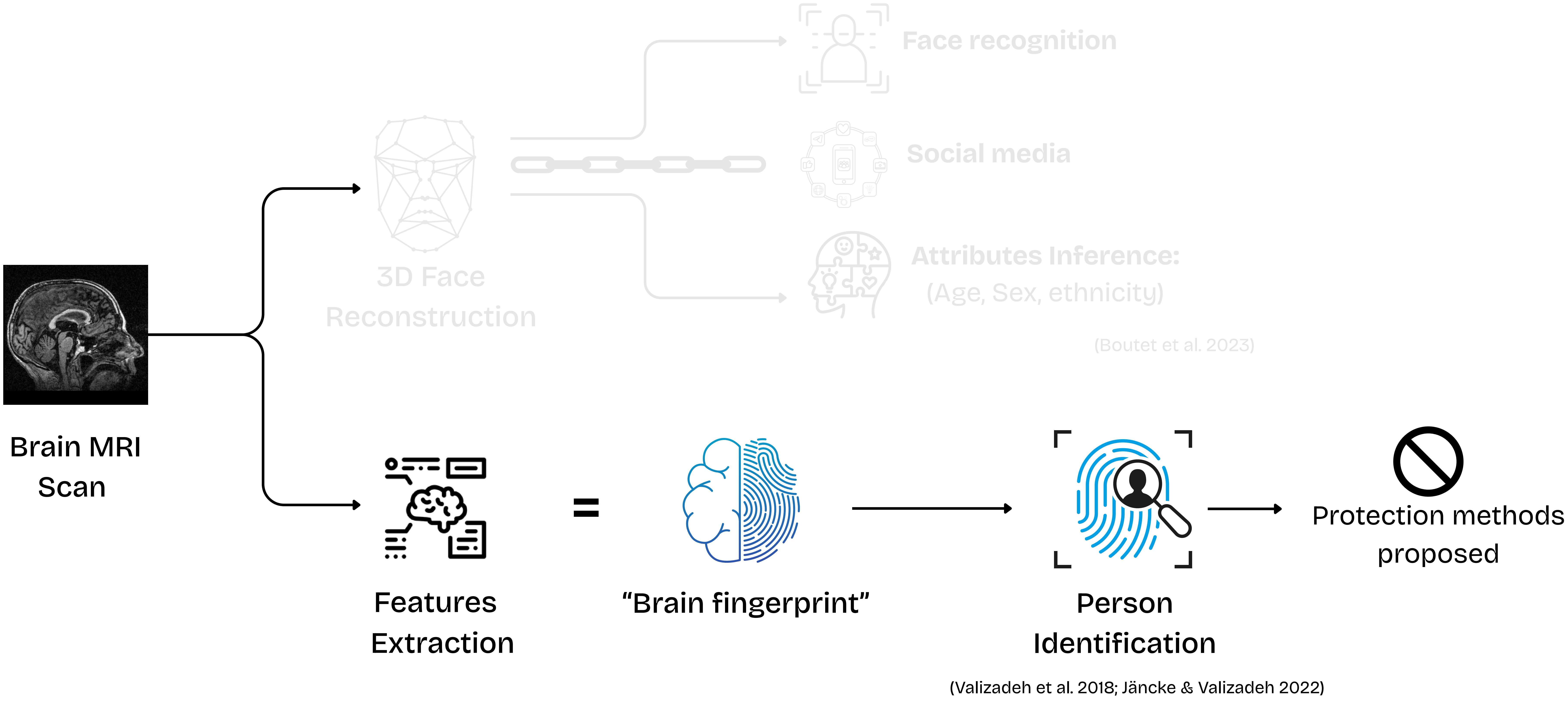
Gao, C., Landman, B. A., Prince, J. L., & Carass, A. (2023). A reproducibility evaluation of the effects of MRI defacing on brain segmentation. *Journal of Medical Imaging*. https://doi.org/10.1117/1.JMI.10.6.064001

Context



Valizadeh SA, Liem F, Mérillat S. Identification of individual subjects on the basis of their brain anatomical features. Scientific Reports. 2018
Jäncke L, Valizadeh SA. Identification of individual subjects based on neuroanatomical measures obtained 7 years earlier. European Journal of Neuroscience. 2022

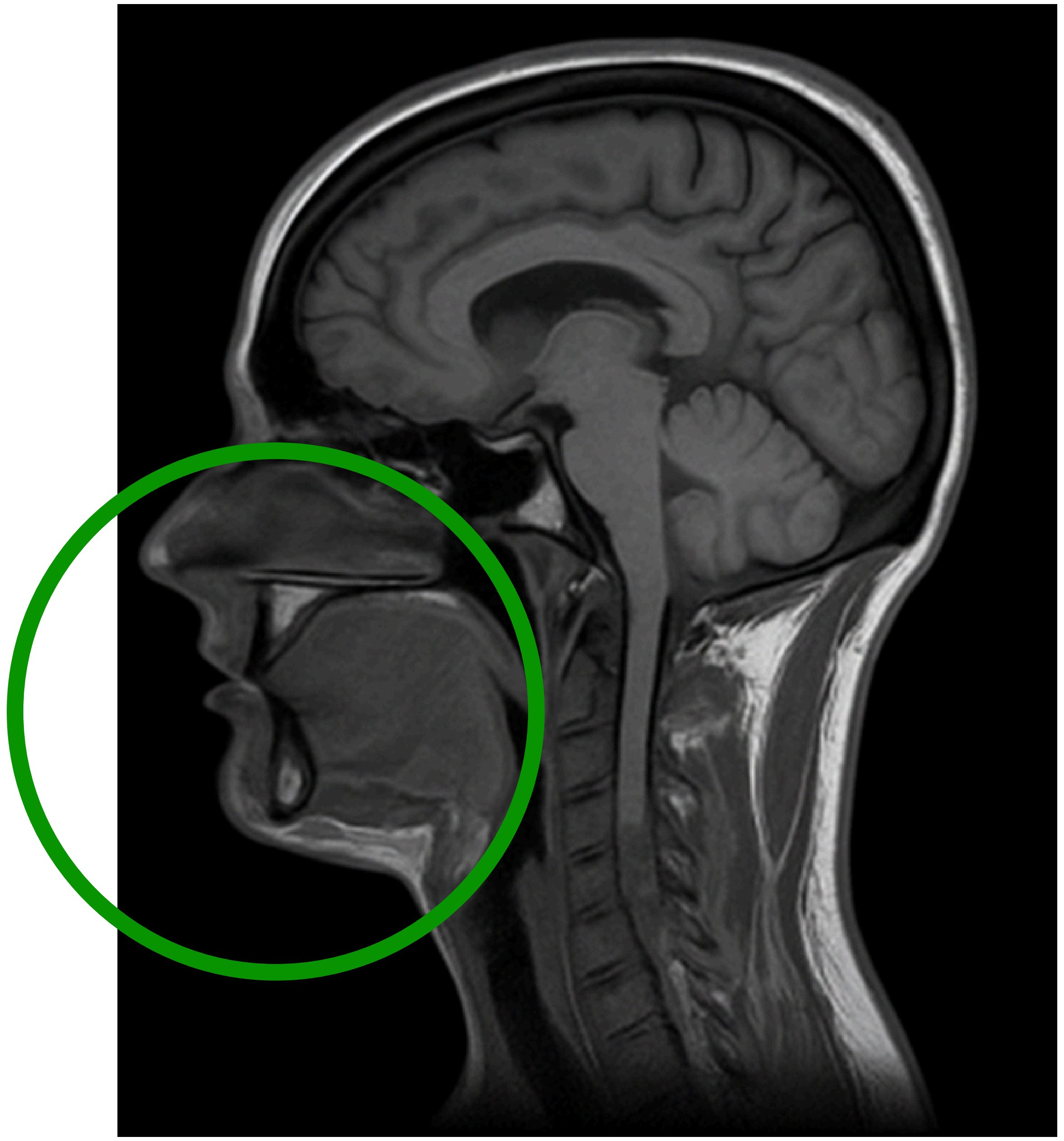
Context



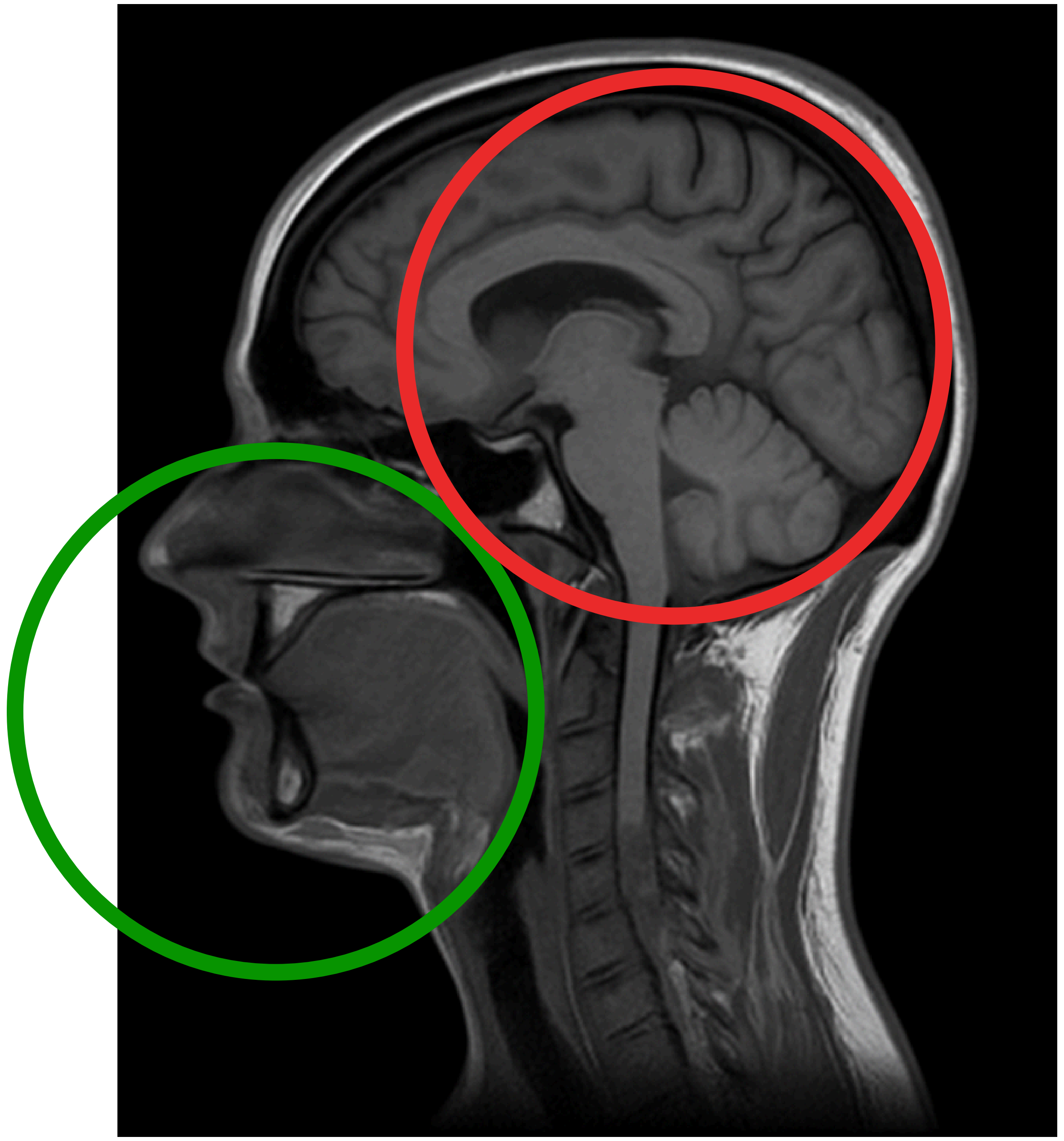
Valizadeh SA, Liem F, Mérillat S. Identification of individual subjects on the basis of their brain anatomical features. Scientific Reports. 2018
Jäncke L, Valizadeh SA. Identification of individual subjects based on neuroanatomical measures obtained 7 years earlier. European Journal of Neuroscience. 2022

Beyond Facial Features: Is Defacing Enough?

"Protected "
by defacing



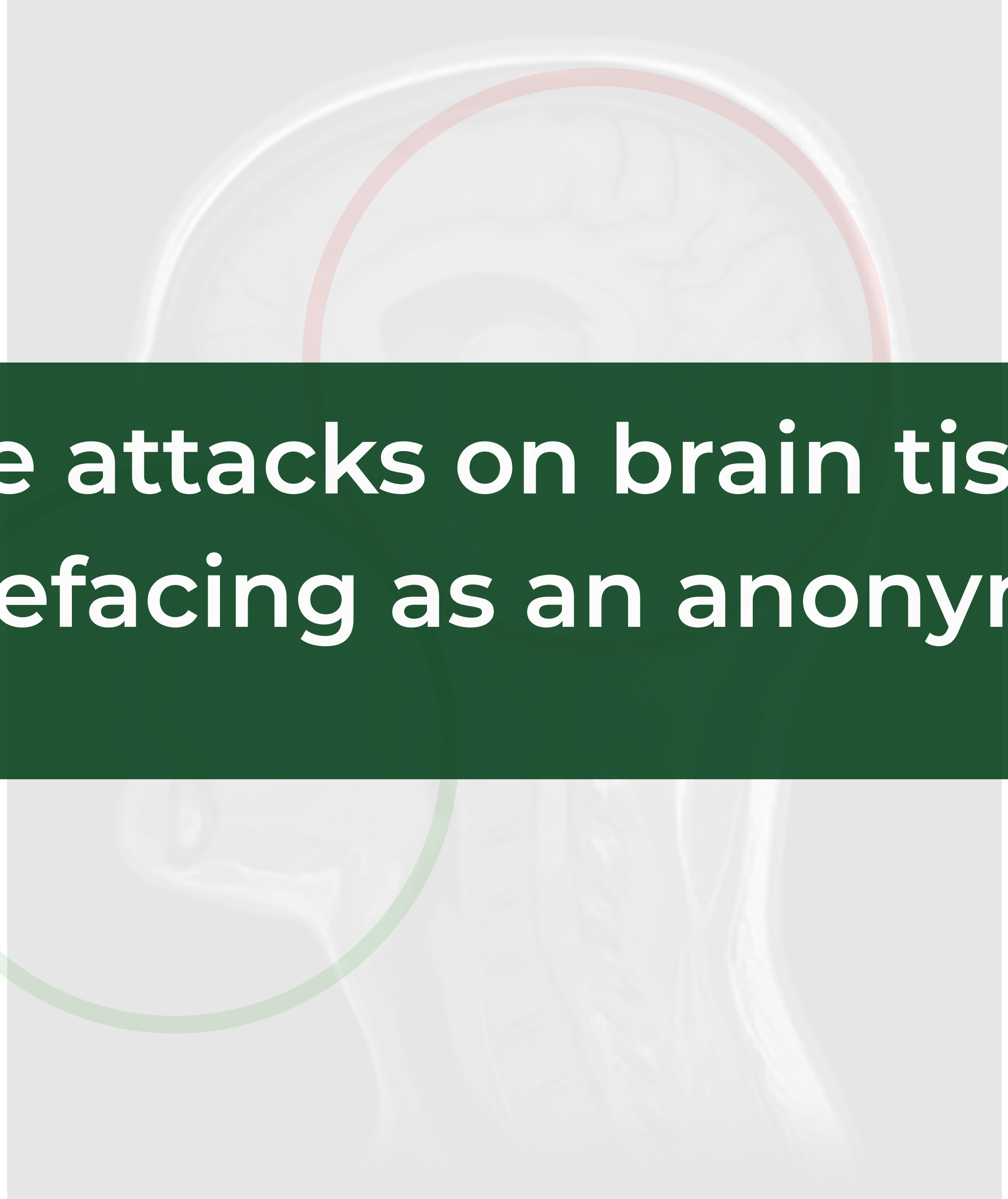
Beyond Facial Features: Is Defacing Enough?



Security flaw ?

“Protected “
by defacing

Beyond Facial Features: Is Defacing Enough?



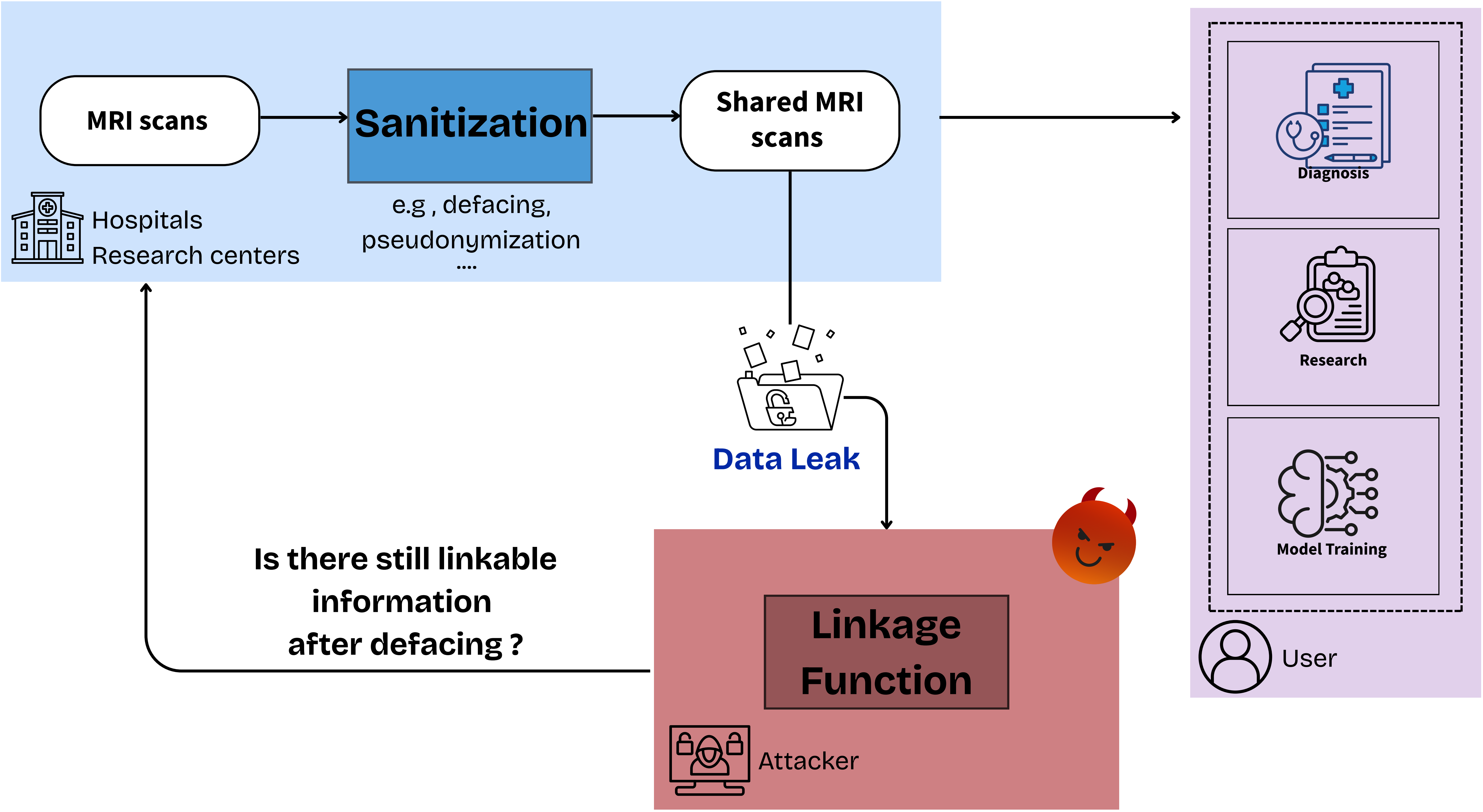
Security flaw ?

“Protected” by defacing

Explore attacks on brain tissue patterns to evaluate defacing as an anonymization technique.

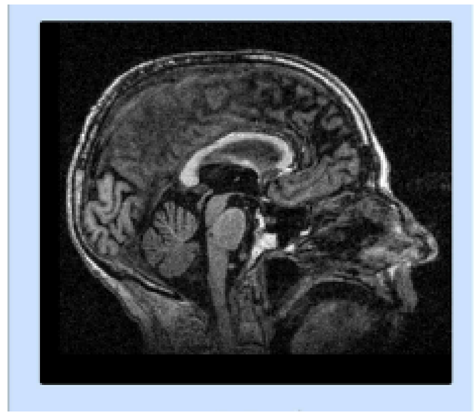
Methodology

System Model

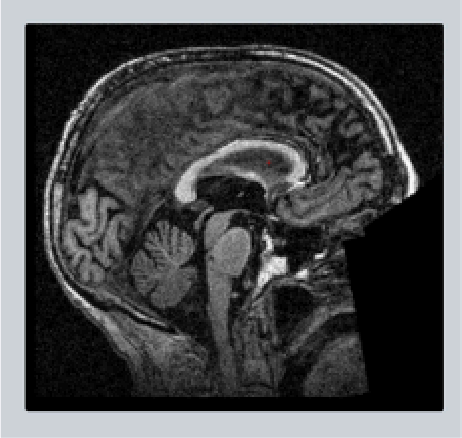
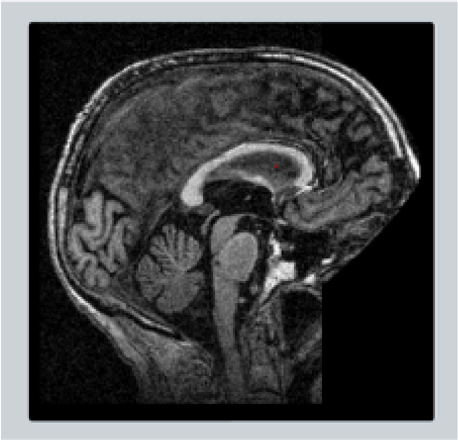


Method

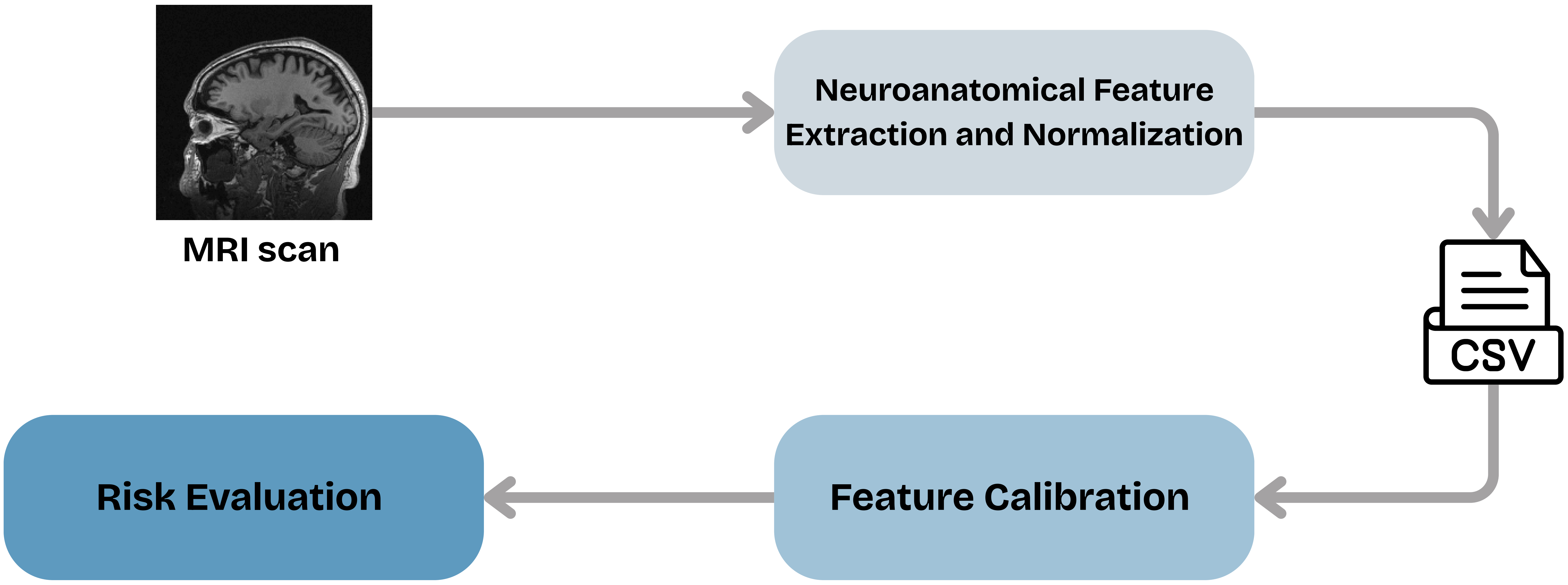
Defacing techniques



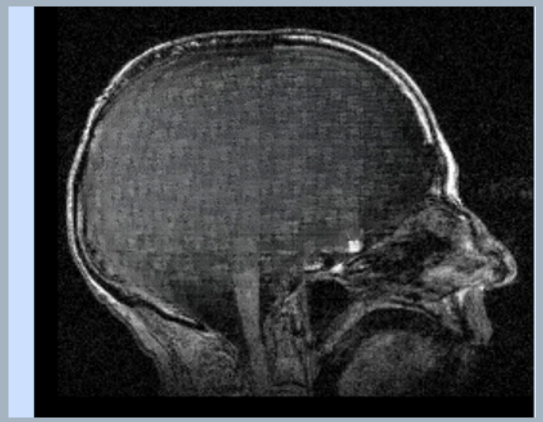
Original

	PyDeface (v2.0.2)	PyFaceWipe (v0.1.1)	mri_reface (v0.3.5)	fsl_deface (v6.0.7.16)
Outputs				
Description	Applies a predefined mask to remove facial details.	Uses a skull-stripping technique to erase facial features.	Replaces facial regions with an average face.	Removes the facial features by aligning the MRI to a standard model.

Method


MRI scan
Neuroanatomical Feature Extraction and Normalization
CSV
Feature Calibration
Risk Evaluation

Neuroanatomical Feature Extraction and Normalization



For each scan:

- **Original**

&

- **4 defaced versions**

Features Extraction and Normalization Pipeline

Segmentation & Parcellation by FreeSurfer

15 Extracted Features (volumes):

- Global measures:
  - **Intracranial Volume (ICV)**
  - Total Cortical Thickness
  - Total Cortical Surface Area
  - Cortical Gray Matter Volume
- Subcortical volumes:
  - Thalamus
  - Putamen
  - Pallidum
  - Caudate
  - Hippocampus
  - Amygdala
  - Accumbens
  - Brainstem
- Other tissue volumes:
  - Cerebral White Matter
  - Cerebellar Gray Matter
  - Cerebellar White Matter

Normalization by ICV

Feature Calibration

PLDA Training

Training Set :

- Normalized Features of 20 subjects

- Extracted from original scans

Learned Transformation

Apply to the remaining subjects

PLDA :Probabilistic Linear Discriminant Analysis

- Improve the separability between subjects.
- Emphasize differences between individuals while reducing variations within the same subject.
- Make the identification task more accurate.

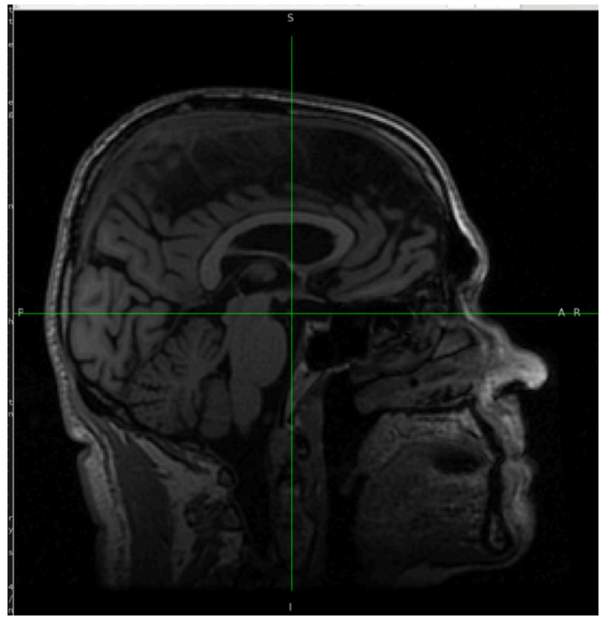
Risk Evaluation

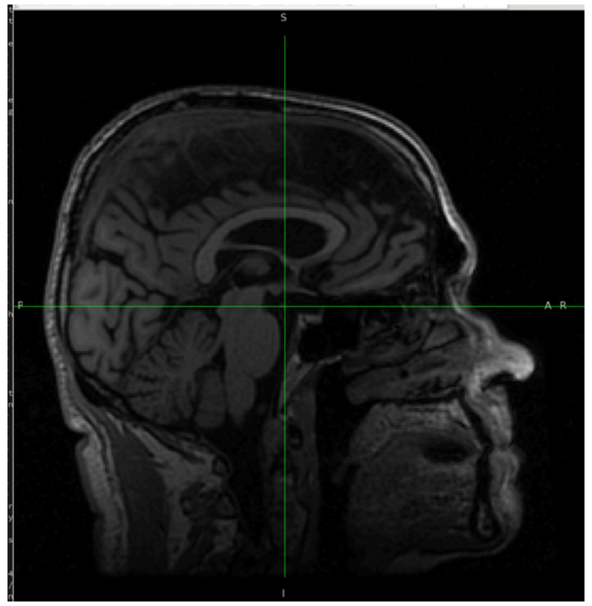
Threat Scenarios

- Original vs Original (Baseline)
- Original vs Defaced
- Defaced vs Defaced

Attack Scenario

Original **vs** **Original**





$S_{1,1}$

H

/H

$S_{1,2}$

$S_{2,1}$

$S_{2,2}$

...

Notation:

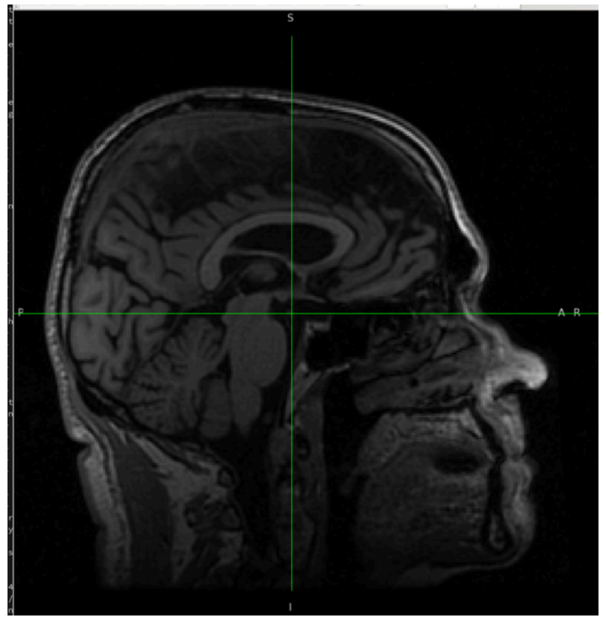
$S_{i,j}$: Subject i, Scan j

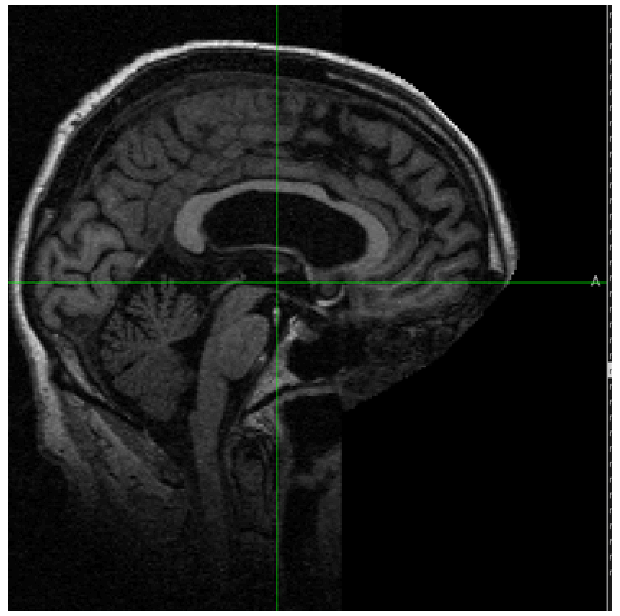
H: Mated pair (same subject)

/H: Non-mated pair (different subjects)

Attack Scenario

Original **vs** **Defaced**





$S_{1,1}$

H

$S_{1,2}$

$S_{2,1}$

$S_{2,2}$

...

/H

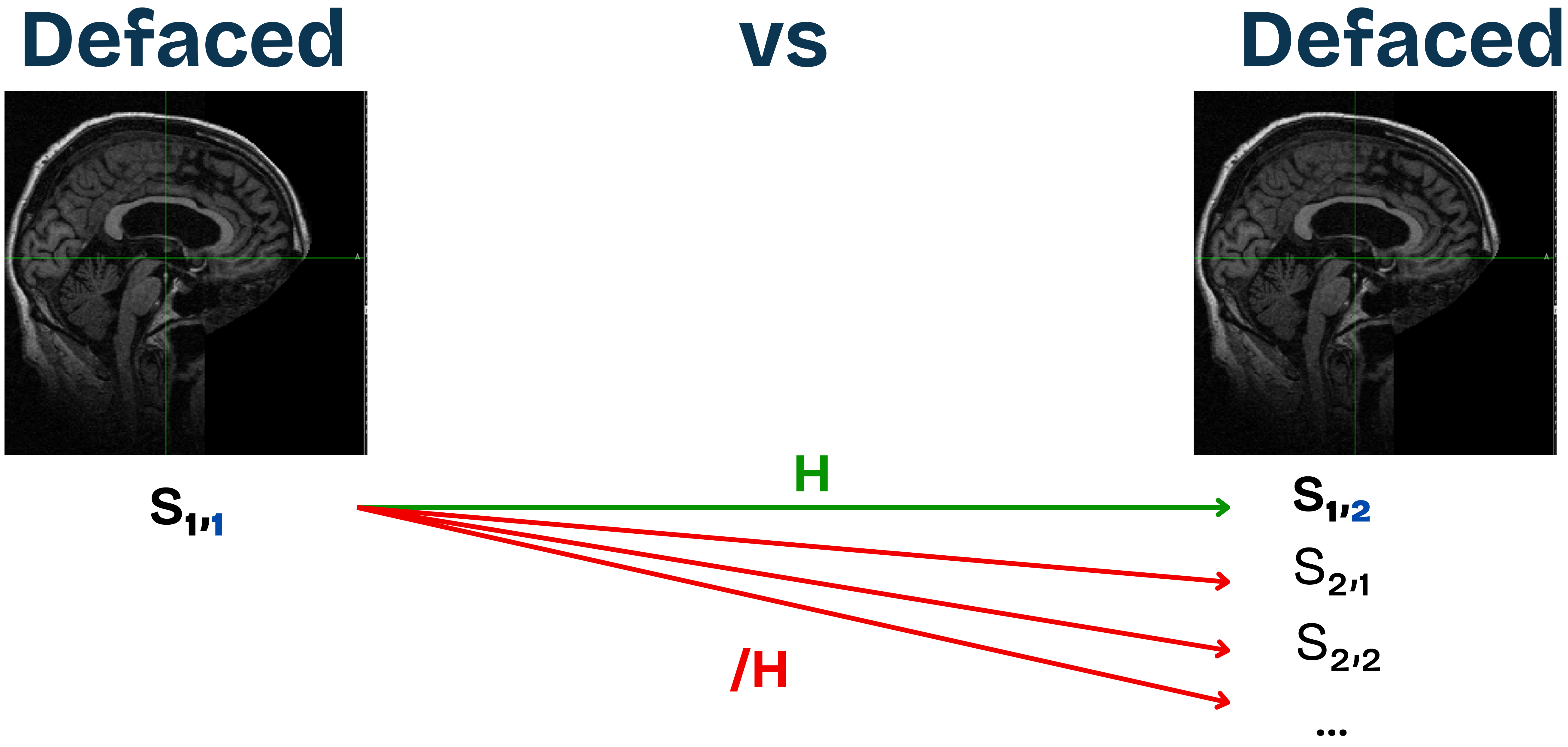
Notation:

$S_{i,j}$: Subject i, Scan j

H : Mated pair (same subject)

/H : Non-mated pair (different subjects)

Attack Scenario



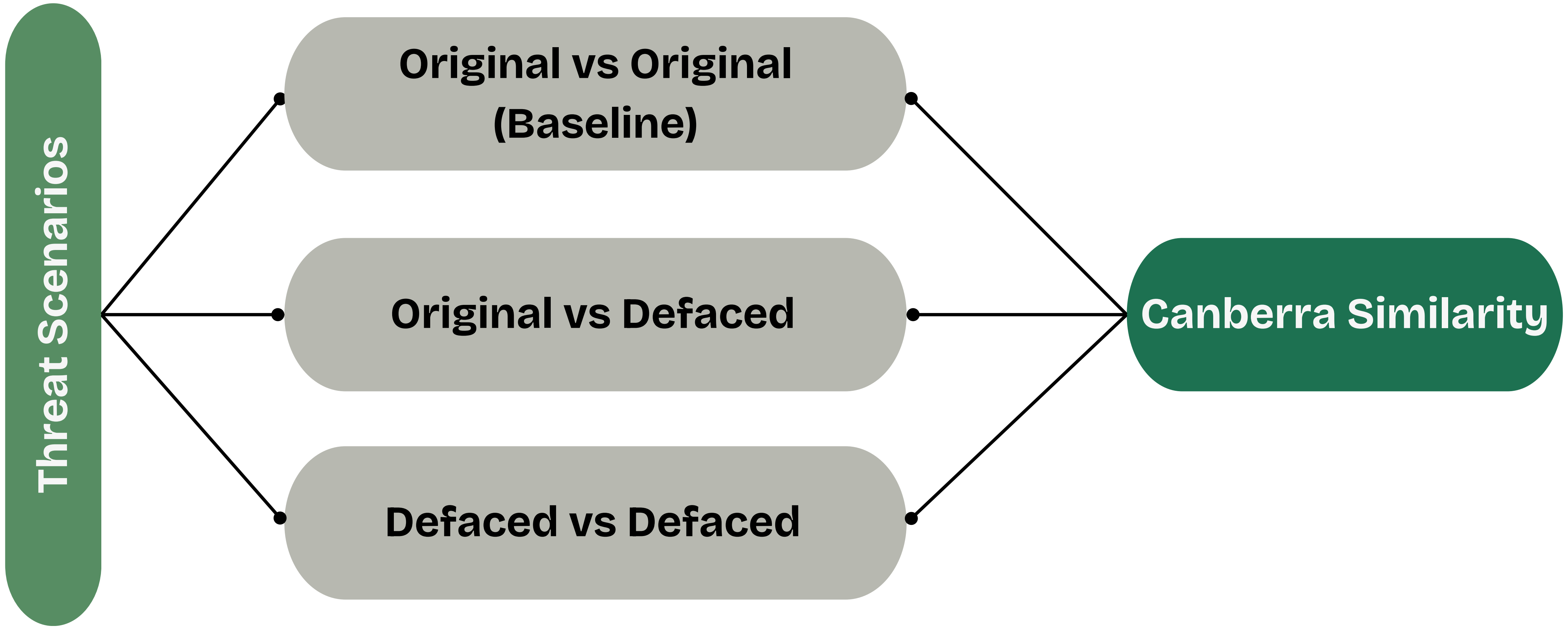
Notation:

$S_{i,j}$: Subject i, Scan j

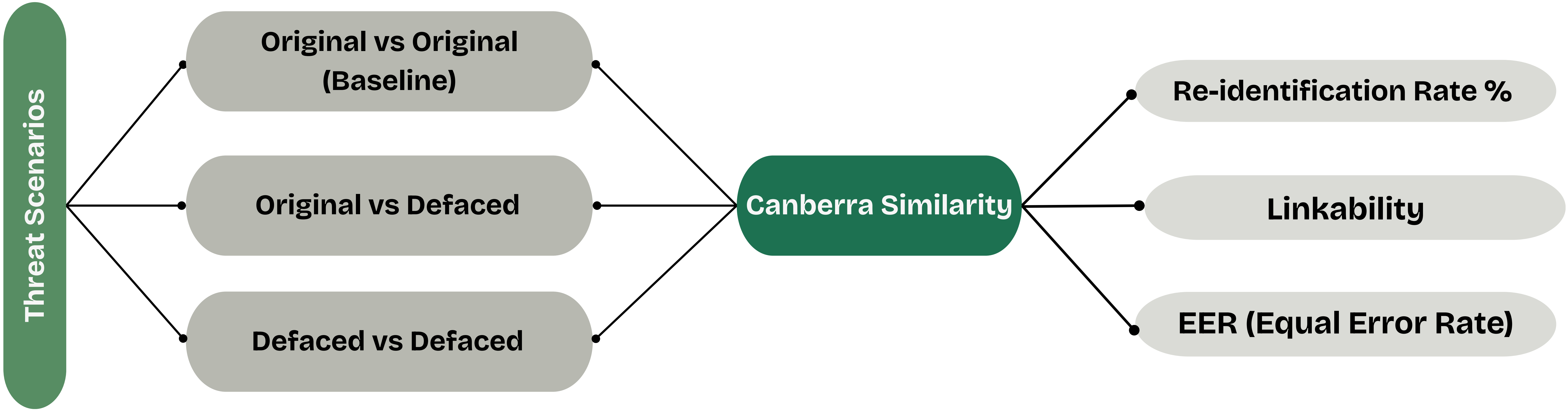
H : Mated pair (same subject)

/H : Non-mated pair (different subjects)

Risk Evaluation



Risk Evaluation

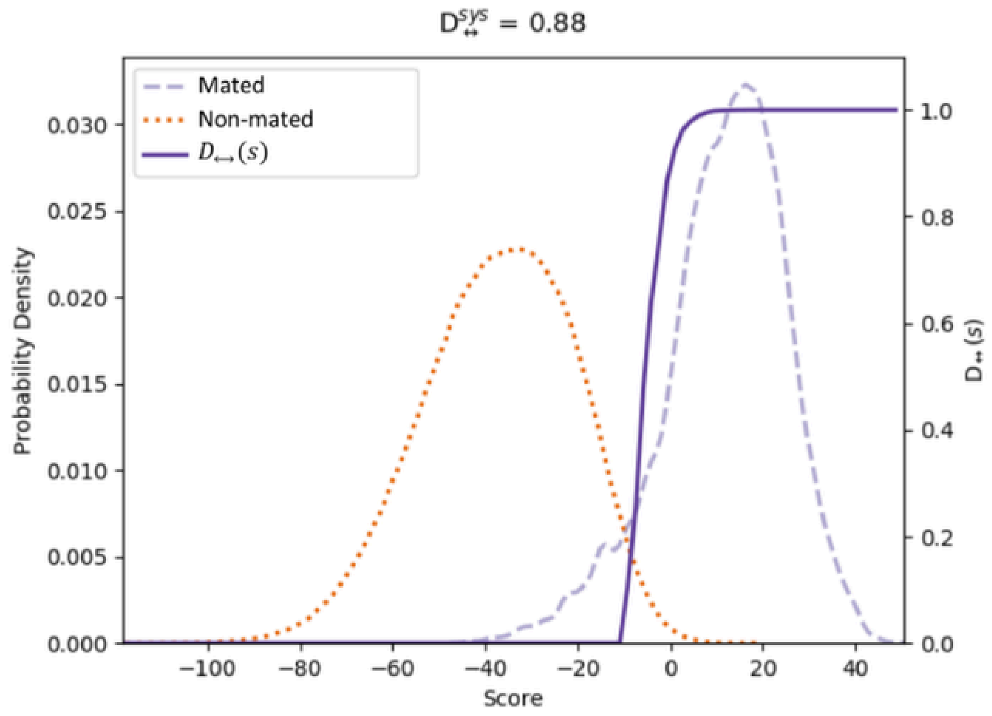


Metrics

Re-identification Rate %

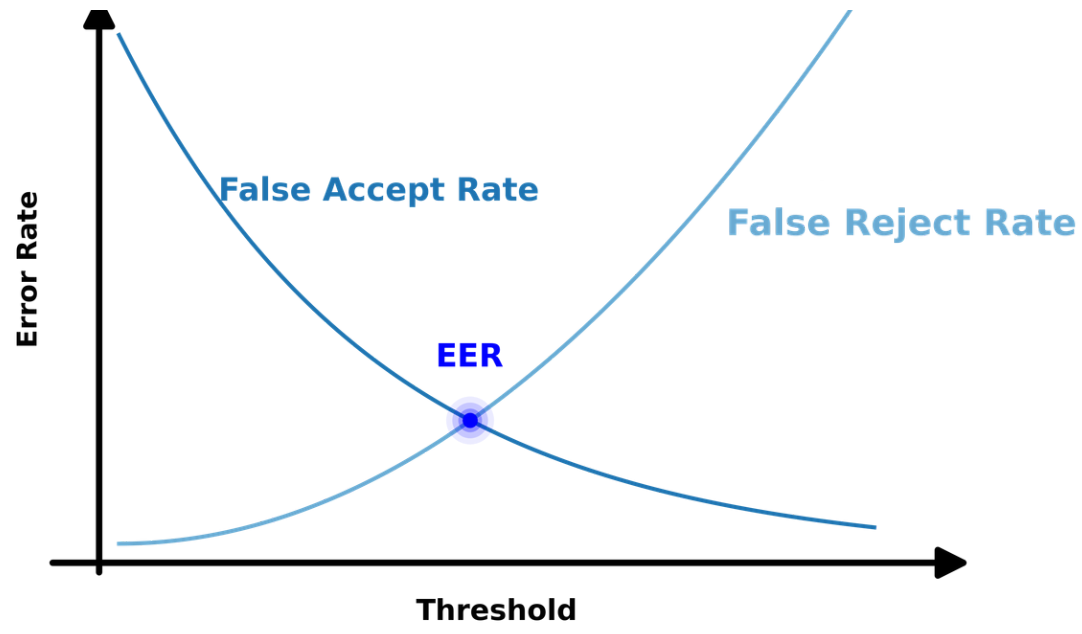
- **Best** match = **highest** similarity
- % of correctly re-identified images
- Rate ↑ ⇒ Privacy ↓

Linkability



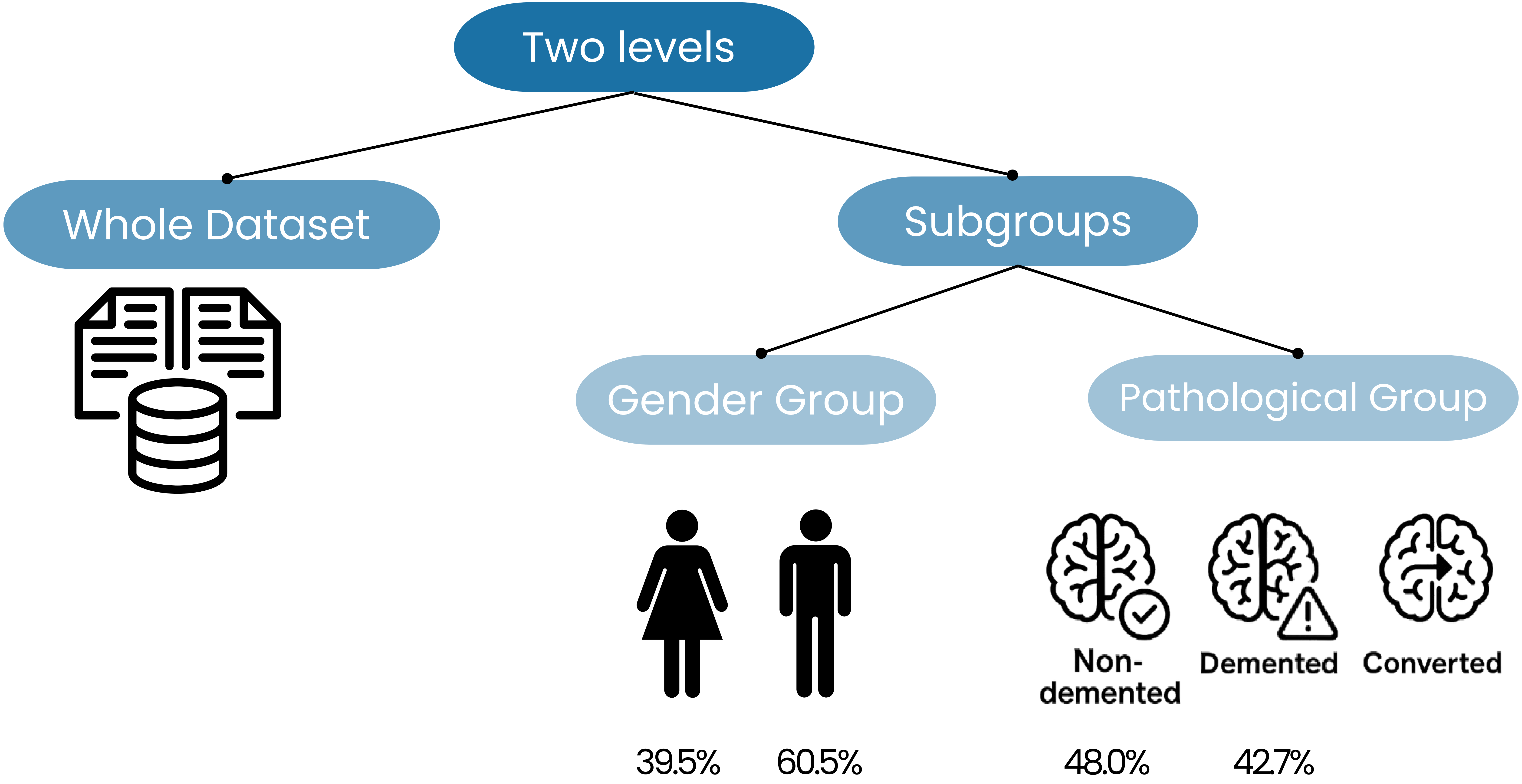
- ∈ [0, 1]
- Linkability ↓ ⇒ Privacy ↑

EER (Equal Error Rate)

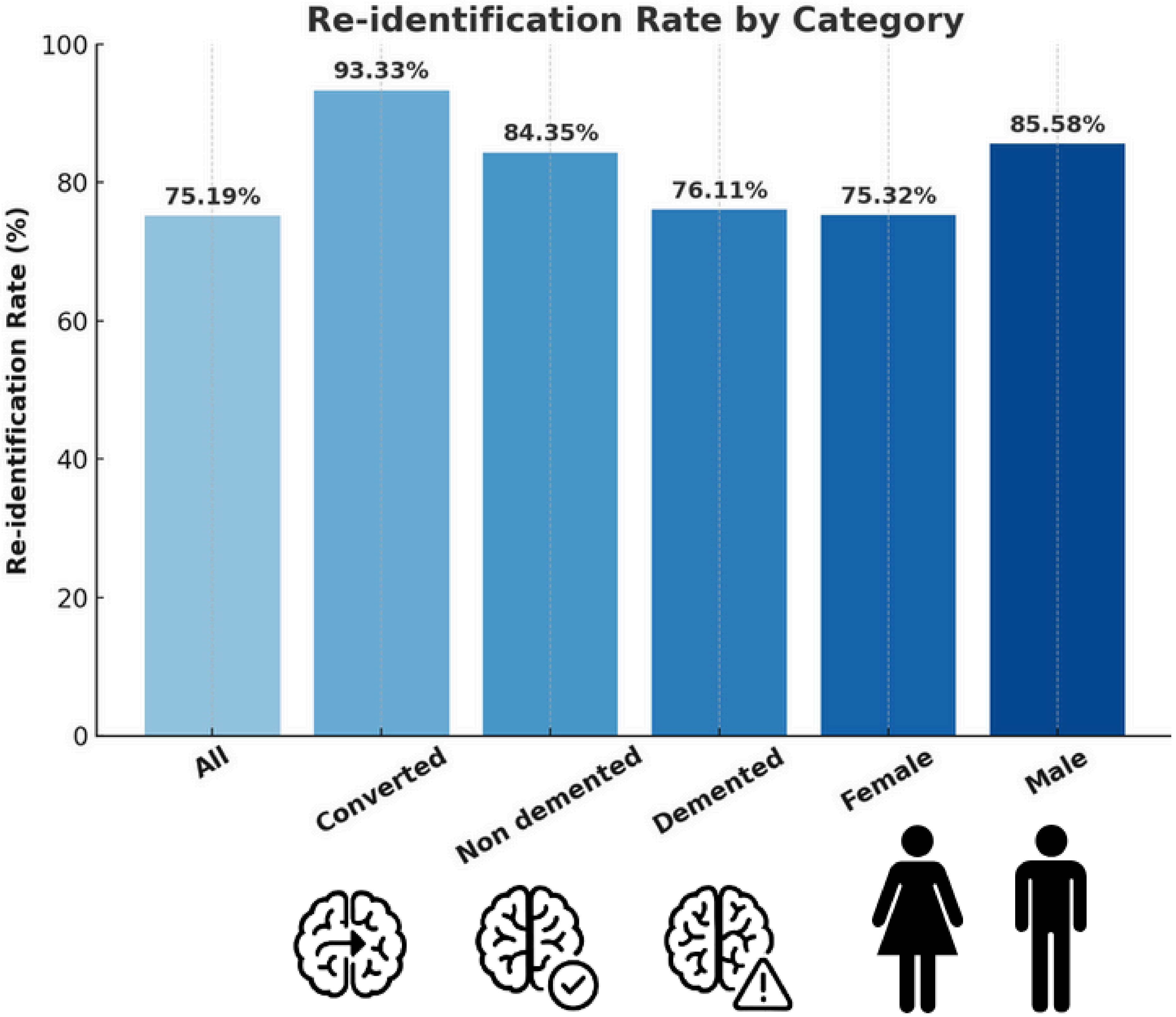


- ∈ [0, 05]
- Threshold on distance
- EER ↑ ⇒ Privacy ↑

Evaluation


Two levels
Whole Dataset
Subgroups
Gender Group
Pathological Group
Non-demented
Demented
Converted
39.5%
60.5%
48.0%
42.7%

The baseline (original-original)

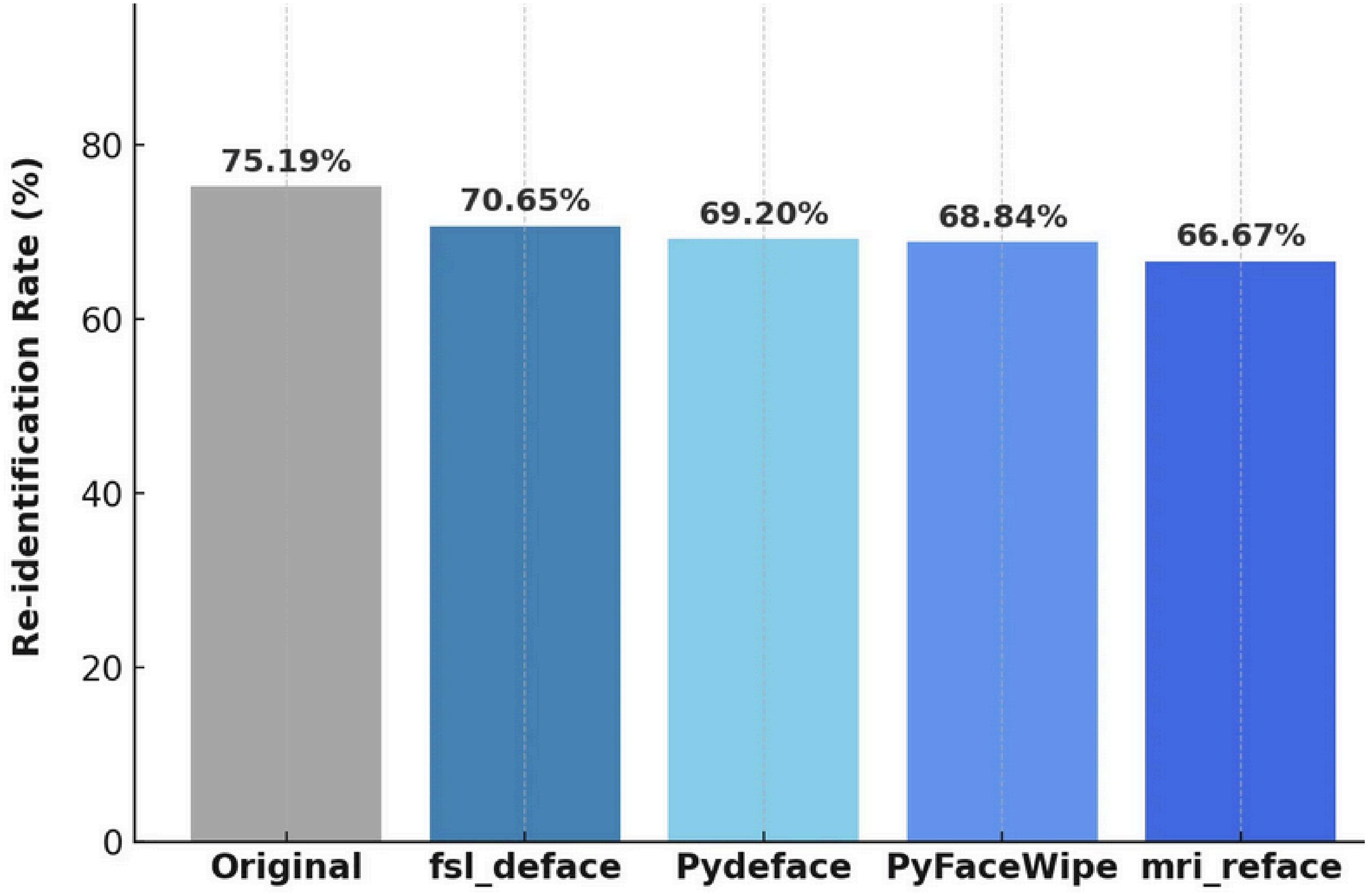


Re-identification rate on the whole dataset : **75.19% (random guess: 0.0068%)**

⇒ Re-identification is feasible from original MRI scans using neuroanatomical features, confirming their **subject-specific uniqueness**.

Impact of defacing (original-vs-defaced)

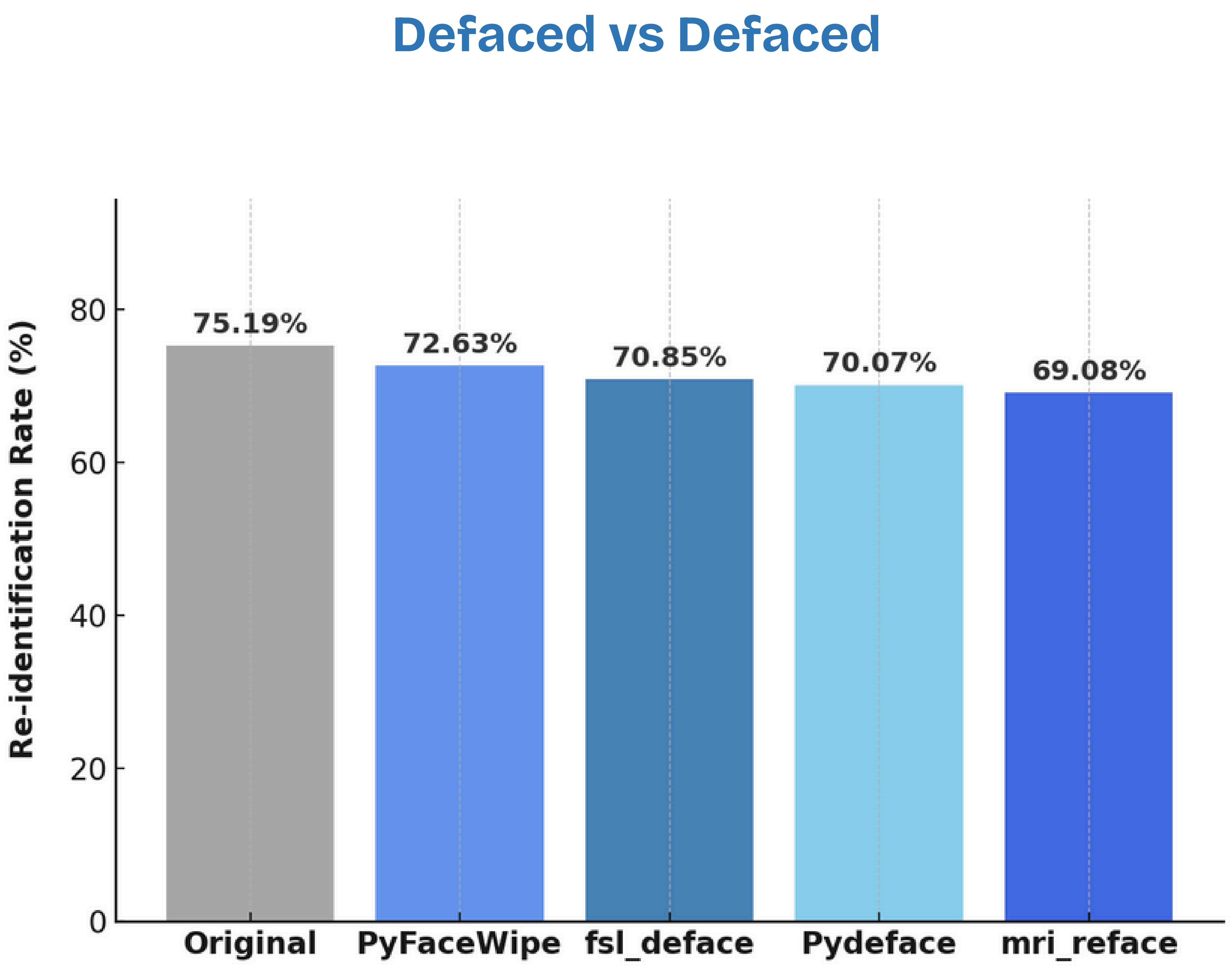
Original vs Defaced



⇒ Re-identification remains possible even after defacing.

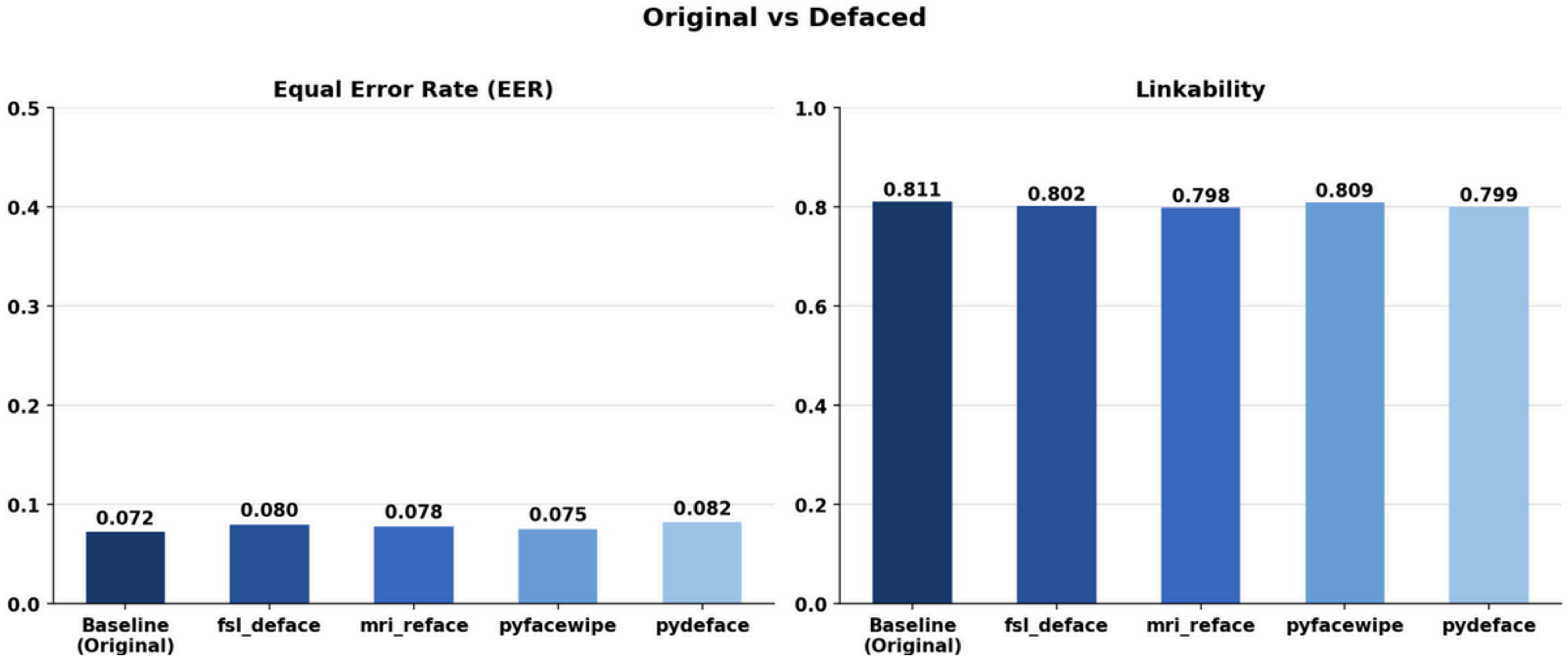
⇒ As expected, no method provides protection

Residual linkability (defaced-vs-defaced)



⇒ Identity is still linkable between defaced scans.

All the metrics agree



Results

- Confirmation of **subject-specific uniqueness** of MRI scans
  - **Baseline re-identification rate** of **75.19%** and an **EER** of **0.072**

- **None** of the **defacing techniques fully prevent re-identification**
  - The best method, mri_reface, only brings the rate by 9%

- **Identity remains linkable** even between defaced scans, with **rates still** around **69–72%** and **EERs** between **0.073 and 0.088**

Results

- Confirmation of **subject-specific uniqueness** of MRI scans
  - **Baseline re-identification rate** of **75.19%** and an **EER** of **0.072**

- **None** of the **defacing techniques fully prevent re-identification**
  - The best method, mri_reface, only brings the rate by 9%

- **Identity remains linkable** even between defaced scans, with **rates still** around **69–72%** and **EERs** between **0.073 and 0.088**

Key Takeaway: brain morphology remains **highly distinctive** even after defacing.

Key Takeaway: **defacing** methods are **insufficient** to eliminate linkage attacks.

Results

- Confirmation of **subject-specific uniqueness** of MRI scans
  - Baseline **re-identification rate** of **75.19%** and an **EER** of **0.072**

- **None** of the **defacing techniques fully prevent re-identification**
  - The best method, mri_reface, only brings the rate by 9%

- **Identity remains linkable** even between defaced scans, with **rates still** around **69–72%** and **EERs** between **0.073 and 0.088**

Key Takeaway: brain morphology remains **highly distinctive** even after defacing.

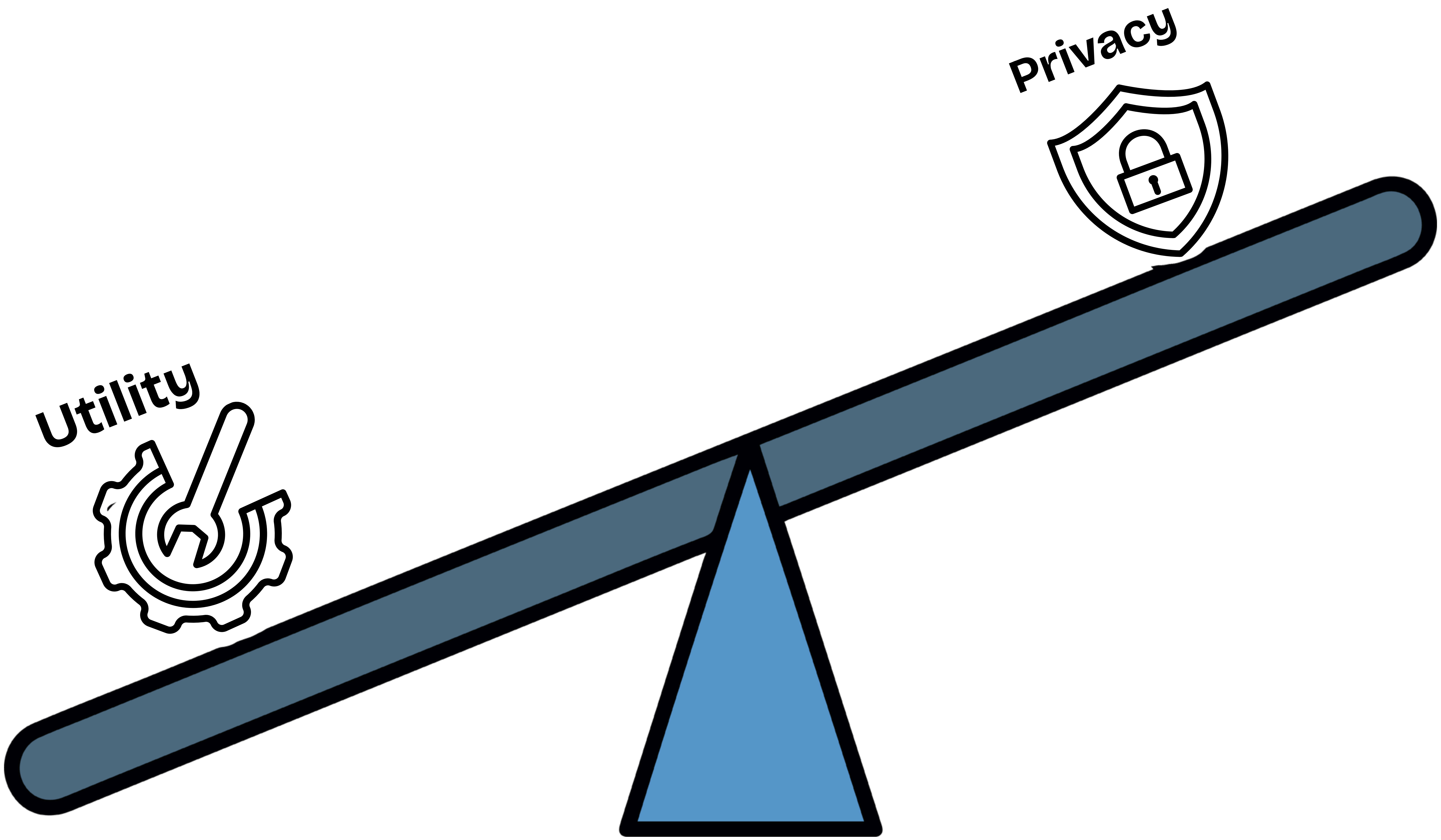
Key Takeaway: **defacing** methods are **insufficient** to eliminate linkage attacks.

New protective methods are needed , going beyond defacing while preserving clinical utility

Future Work

1. Develop privacy-preserving models
2. Evaluate data quality and utility on real clinical use cases
3. Deliver deployable solutions for clinical settings

Tradeoff

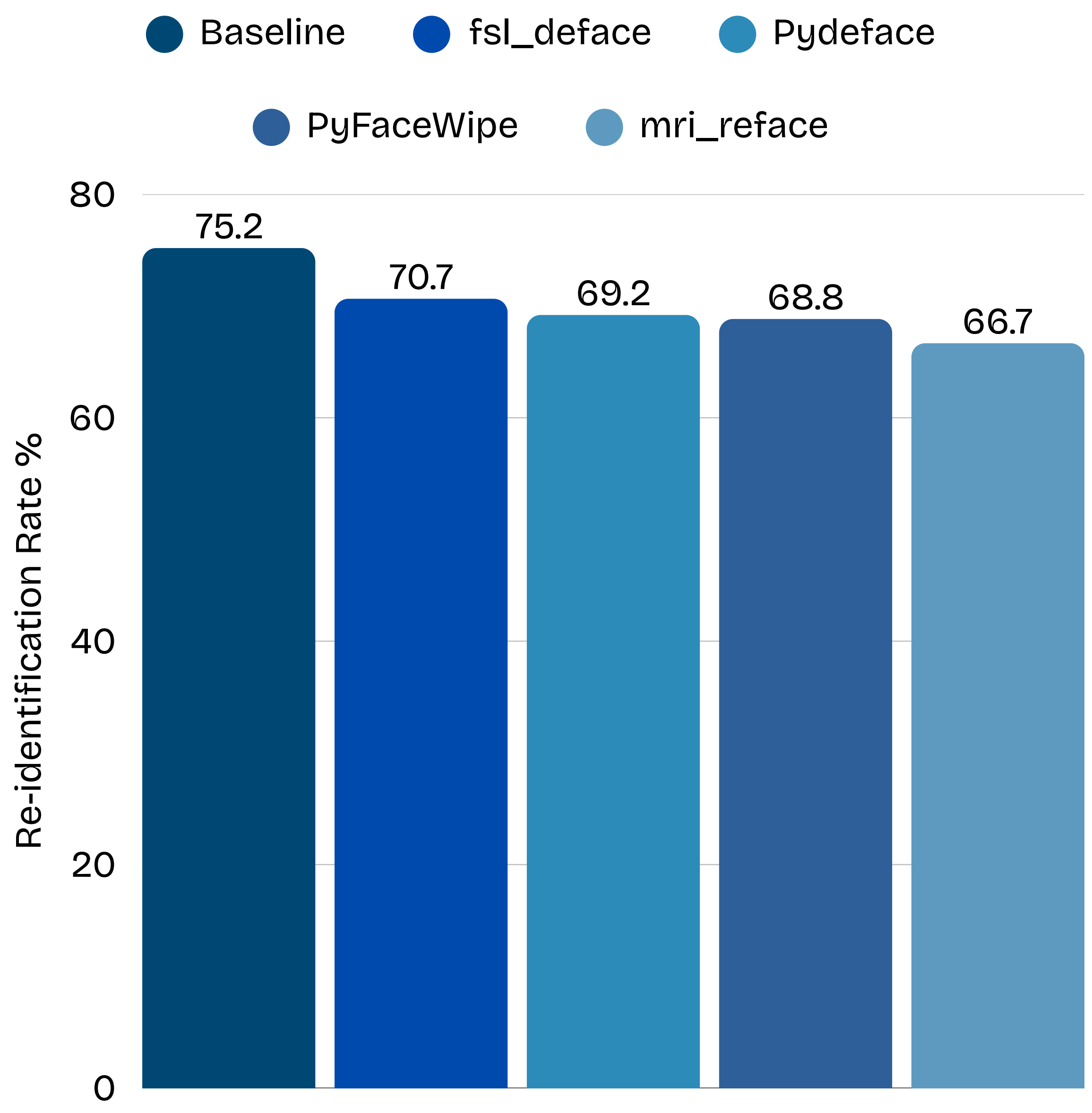


Thank you

Imen Bajar
imen.bajar@inria.fr

Results

Original vs Defaced



- **Original vs Original (Baseline)**
  - Re-identification rate: 75.19%
  - Confirms subject-specific uniqueness of MRI scans.
- **Original vs Defaced**
  - Re-identification rate: 66% – 71%, depending on method.
  - None of the defacing techniques fully prevent re-identification.
- **Defaced vs Defaced**
  - Re-identification rate: ~70%
  - Identity remains linkable even between defaced scans.

Metrics

Re-identification Rate %

Best match = highest similarity
% of correctly re-identified images

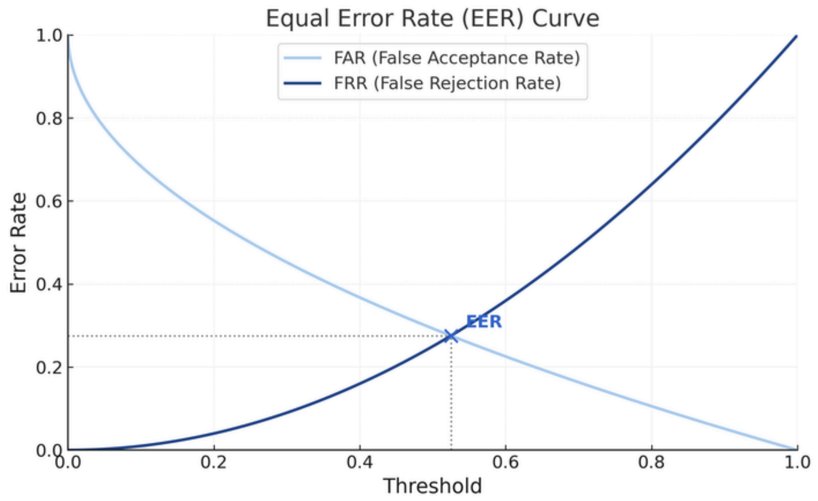
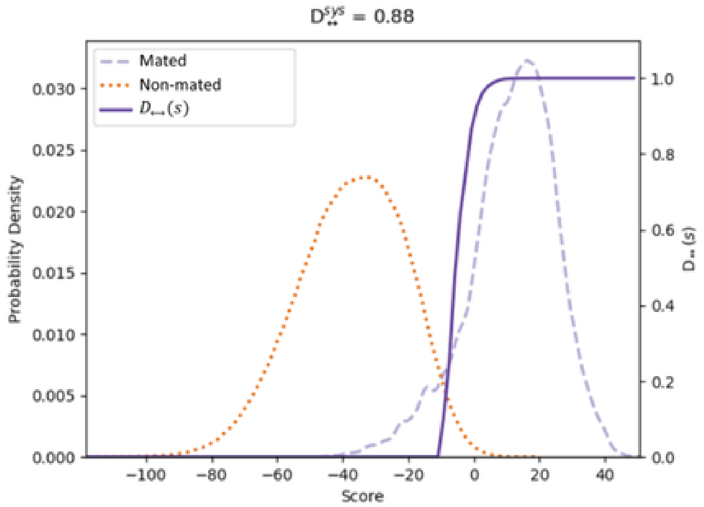
Privacy Metrics

	EER (Equal Error Rate)	Linkability
📏 **Range**	[0,0.5]	[0,1]
👍 **More Privacy**	• Higher EER • 0.5 = random guessing	•Lower linkability ⇒ better anonymization • 0 = perfect anonymization
👎 **Less Privacy**	• Lower EER ⇒ Weaker anonymization (more identifiable)	• High linkability ⇒ Worse anonymization (perfect distinguishability)

Metrics

Canberra Similarity Score

Privacy Metrics

	EER (Equal Error Rate)	Linkability
📖 **Definition**		
📏 **Range**	[0,0.5]	[0,1]
👍 **More Privacy**	• Higher EER • 0.5 = random guessing (max confusion)	•Lower linkability ⇒ better anonymization • 0 = perfect anonymization
👎 **Less Privacy**	• Lower EER ⇒ Weaker anonymization (more identifiable)	• High linkability ⇒ worse anonymization (perfect distinguishability)
🎯 **Goal**	**Maximize EER** → harder to verify identity	**Minimize linkability** → harder to link identities

Re-identification Rate %

Best match = highest similarity
Percentage of correctly re-identified images

Canberra Similarity

It is a weighted version of L_1 (Manhattan) distance :

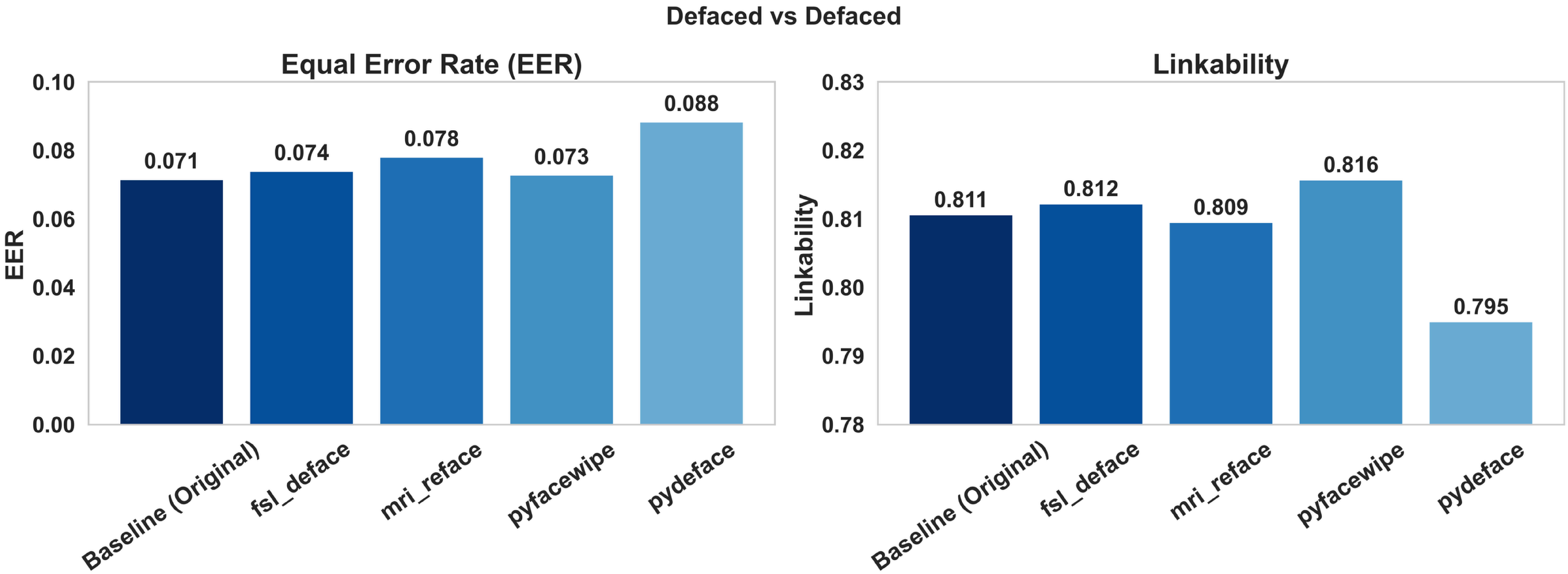
$$d(\mathbf{p}, \mathbf{q}) = \sum_{i=1}^{n} \frac{|\mathbf{p}_i - \mathbf{q}_i|}{|\mathbf{p}_i| + |\mathbf{q}_i|}$$

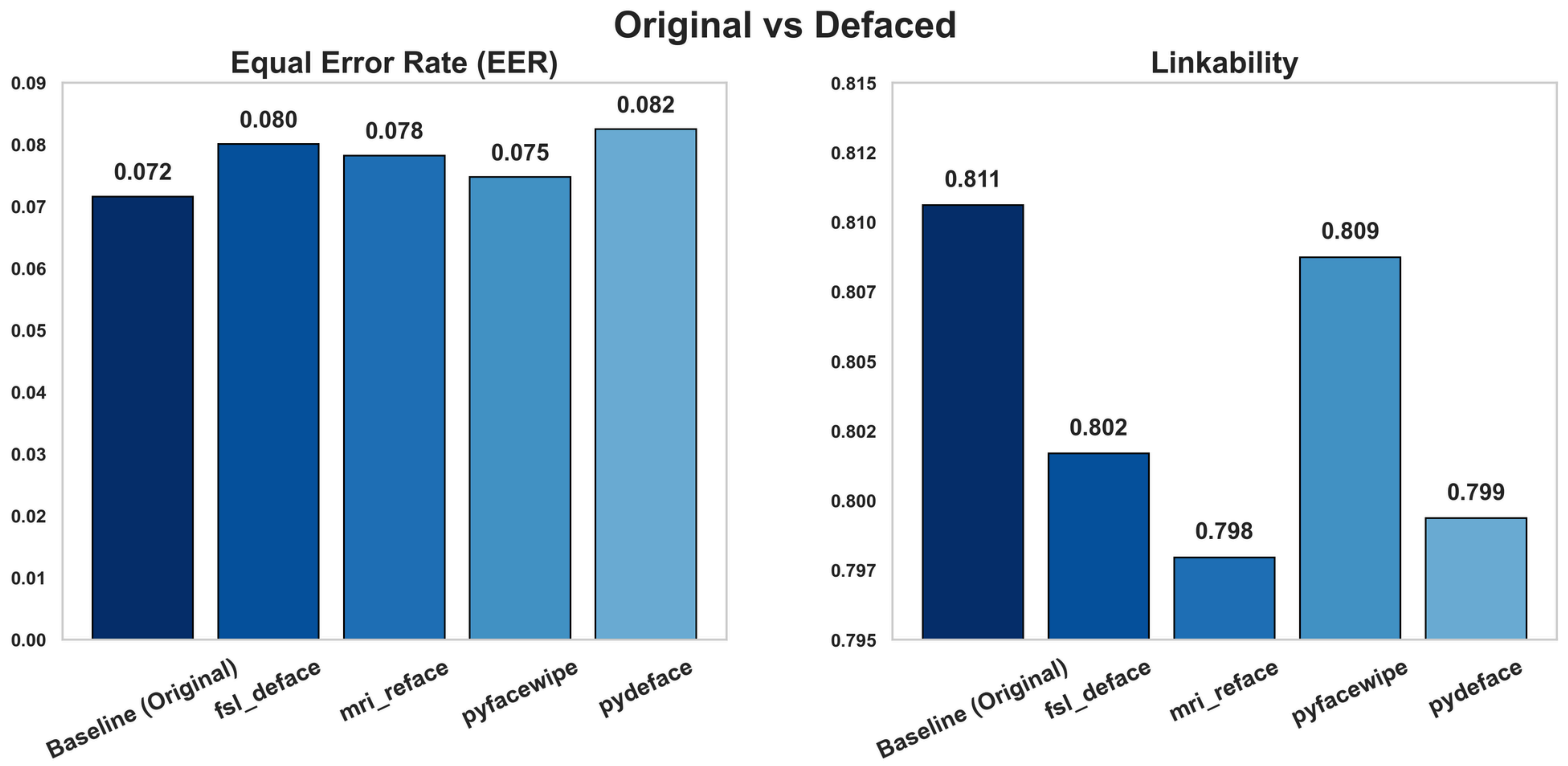
Where $\mathbf{p} = (p_1, p_2, \ldots, p_n), \quad \mathbf{q} = (q_1, q_2, \ldots, q_n)$

are vectors

From distance to similarity :

$$\text{Similarity} = \frac{1}{1 + d(\mathbf{p}, \mathbf{q})}$$


Defaced vs Defaced
Equal Error Rate (EER)
EER
0.10
0.08
0.06
0.04
0.02
0.00
0.071
0.074
0.078
0.073
0.088
Baseline (Original)
fsl_deface
mri_reface
pyfacewipe
pydeface
Linkability
Linkability
0.83
0.82
0.81
0.80
0.79
0.78
0.811
0.812
0.809
0.816
0.795
Baseline (Original)
fsl_deface
mri_reface
pyfacewipe
pydeface


Original vs Defaced
Equal Error Rate (EER)
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
0.072
0.080
0.078
0.075
0.082
Baseline (Original)
fsl_deface
mri_reface
pyfacewipe
pydeface
Linkability
0.815
0.812
0.810
0.807
0.805
0.802
0.800
0.797
0.795
0.811
0.802
0.798
0.809
0.799
Baseline (Original)
fsl_deface
mri_reface
pyfacewipe
pydeface