QUANTIFYING PRIVACY RISKS IN FOUNDATION MODEL REPRESENTATIONS OF BRAIN MRI

Anonymous authors
Paper under double-blind review

ABSTRACT

Pretrained representations are increasingly used to share and reuse information from medical images without exposing the original scans, but it is unclear how much patient-specific information these representations retain. We investigate this question for brain magnetic resonance imaging (MRI) by evaluating five frozen pretrained encoders spanning different architectures, pretraining objectives, and training domains on two longitudinal T1-weighted datasets, OASIS-2 and MIRIAD. Under an attacker model with access to released embeddings, the corresponding public encoder, and auxiliary MRI scans, we evaluate patient linkage and sensitive-attribute inference. Across the evaluated representations, we observe substantial subject-specific information: rank-1 patient linkage reaches up to 99% on MIRIAD and exceeds 90% on OASIS-2 for the highest-risk representation, despite the absence of an identification objective during encoder training. We further show that the released embeddings retain information about a demographic attribute, with recorded-sex prediction reaching balanced accuracy of 90.9% on OASIS-2 and 95.5% on MIRIAD. Overall, our results demonstrate that replacing raw brain MRI with learned embeddings does not, by itself, eliminate patient-linkage or attribute-inference risk, and is therefore insufficient on its own to establish anonymity.

1 INTRODUCTION

Magnetic resonance imaging (MRI) is an imaging modality that generates detailed images of the internal structures of the body, including the brain's anatomy. Given this level of anatomical detail, brain MRI is a routine tool in the diagnosis and monitoring of neurological and psychiatric conditions such as Alzheimer's disease, stroke, and brain tumors (Frisoni et al., 2010; Ahmed et al., 2021; Martucci et al., 2023). Many AI models have been trained to detect and classify these conditions from brain MRI scans (Korolev et al., 2017; Mohsen et al., 2018).

At the same time, studies have shown that brain morphology can serve as a biometric signature, a "brain fingerprint", which enables subject re-identification through comparison with reference data (Wachinger et al., 2014; Takao et al., 2015; Valizadeh et al., 2018). Therefore, brain MRI has a dual nature, at once a diagnostic tool and a carrier of identifying biometric information.

In recent years, foundation models, large neural networks trained on broad and diverse datasets, have emerged. Many of these models use self-supervised learning to learn transferable representations that can be adapted to a range of downstream tasks (Bommasani et al., 2022). This approach is especially useful for medical imaging, where labeled data can be scarce and costly to obtain. Instead of training a model from scratch, a pretrained model can be adapted to the specific task at hand with limited task-specific labeled data (Liu et al., 2025; Tak et al., 2026). For brain MRI specifically, a growing body of work has explored foundation models for representation learning and downstream analysis (Barbano et al., 2026; Ghamizi et al., 2025; Mazher et al., 2025). Throughout this paper, we refer to the vector representation produced by a pretrained encoder as an *embedding*.

While the utility and generalization capabilities of brain foundation models have received growing attention (Liu et al., 2025; Tak et al., 2026), these evaluations do not establish how much patient-specific information their representations retain. Prior work has investigated patient re-identification from foundation-model features in other imaging modalities (Nebbia et al., 2025), motivating a

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

closer examination of brain MRI. Indeed, brain MRI is a particularly relevant setting because longitudinal scans preserve subject-specific internal anatomy across repeated acquisitions, potentially causing representations from the same patient to remain close in embedding space and thereby providing a persistent signal for linkage. In particular, when MRI embeddings are shared or leaked, access to the same publicly available encoder and auxiliary scans may allow an adversary to link those embeddings to reference records from the same individuals.

This question is also relevant to how anonymization is assessed in practice. Recent guidelines from the European Data Protection Board (EDPB) identify three criteria for evaluating whether data can be considered anonymous: *No Record Isolation, No Linkage, and No Inference* (European Data Protection Board, 2026). Our study examines two corresponding technical risks for released brain MRI representations: linkage across records belonging to the same individual and inference of withheld patient attributes.

As pretrained encoders are increasingly used to produce reusable representations of brain anatomy for downstream tasks, a central privacy question arises: ***when a learned representation is released instead of the raw brain MRI, can it still enable patient linkage or reveal additional patient-specific information?*** Across the evaluated models and datasets, we observe substantial variation in privacy risk, with rank-1 patient matching exceeding 90% for some representations and remaining well above chance across several model–dataset combinations. We address this question through four contributions:

- Define explicit threat models for patient linkage, re-identification, and sensitive-attribute inference from released brain MRI representations, and systematically evaluate five frozen pretrained encoders across two longitudinal MRI datasets.
- Quantify identity-related information through a score-based linkability audit and an explicit rank-1 matching attack (exploiting the proximity of embeddings derived from the same patient), showing that several representations preserve substantial subject-specific information despite the absence of an identification objective during encoder training.
- Evaluate sensitive-attribute inference from released brain MRI representations, showing that the embeddings can reveal the subject's recorded sex in addition to supporting patient linkage.
- Demonstrate that releasing embeddings instead of raw brain MRI does not, by itself, eliminate linkage or attribute-inference risk and therefore is insufficient, on its own, to establish anonymity.

2 Related Work

Privacy risks in brain MRI: Historically, safeguarding patient privacy in medical imaging, as required by privacy regulations like the General Data Protection Regulation (GDPR) European Parliament and Council of the European Union (2016), Personal Information Protection and Electronic Documents Act (PIPEDA) Government of Canada (2000), and California Consumer Privacy Act (CCPA) California State Legislature (2026), has involved removing direct identifiers from medical data before secondary use or disclosure. In medical imaging, this typically includes stripping metadata such as names, birth dates, and medical record numbers from image headers. However, medical images themselves still contain detailed clinical and anatomical information, and patient-specific traces may remain even after obvious identifiers are removed (RSNA).

Brain MRIs, for instance, capture detailed facial anatomy, which makes 3D facial reconstruction and therefore identification via face-recognition deep learning achievable. (Boutet et al., 2023) demonstrated that faces reconstructed in 3D from brain MRI scans could be matched against public photos using deep learning, effectively exposing a person's identity. To guard against this, techniques like defacing are commonly used (Gulban et al., 2020; Mitew et al., 2024). Yet defacing does not fully eliminate the risk either: deep learning can still reconstruct a plausible face from a defaced scan (Jeong et al., 2020; Abramian & Eklund, 2019).

The problem is not limited to the face. Research has shown that the shape of the brain itself can serve as an identifying signature, a "brain fingerprint". Wachinger et al. (2014) developed BrainPrint, a technique that uses brain shape to identify individuals with near-perfect accuracy. Valizadeh

et al. (2018) found that brain features remain distinctive and stable even across years. DeepBrainPrint (Puglisi et al., 2023) explicitly trained a contrastive representation for longitudinal brain MRI re-identification, showing that learned representations can also encode stable subject-specific information across scans. More recently, Bajar et al. (2026) reinforced this finding by testing four defacing methods on brain scans; even after defacing, subjects were still re-identified 66–73% of the time using brain-feature-based similarity attacks. Taken together, these findings show that brain MRI can reveal identity in more than one way through brain features even after defacing methods have been applied.

Medical foundation models: Brain MRI foundation models can be broadly grouped according to their pretraining objective. Task-agnostic models are pretrained without a specific downstream task in mind; they rely on self-supervised objectives to learn general-purpose representations that can later be fine-tuned for various tasks (Barbano et al., 2026; Kaczmarek et al., 2025; Huang et al., 2025; Caron et al., 2021).

Task-specific foundation models, in contrast, are pretrained around a particular downstream objective, most often segmentation (Wang et al., 2024; Cox et al., 2024). Their encoders are therefore shaped by this task rather than by the goal of producing general-purpose representations. We evaluate representative models from these different paradigms; their architectures and pretraining objectives are described in Section 3.1.

Privacy risks for medical foundation models: Nebbia et al. (2025) showed that representations extracted from frozen medical foundation models can support patient retrieval in fundus photography, OCT, and chest X-ray, and can also expose demographic attributes. However, it does not report privacy metrics or define a threat model. Also, subsequent work questioned whether some of these results reflect properties specific to foundation models or simpler acquisition and image-level similarity, highlighting the importance of carefully specifying the attacker and evaluation protocol (Engelmann et al., 2026). More broadly, learned representations are known to expose sensitive information beyond their intended downstream objective (Song & Shmatikov, 2019), and recent work has emphasized the importance of explicit attacker knowledge and disclosure assumptions when evaluating clinical foundation models (Tonekaboni et al., 2026).

Summary: Our work connects two previously separate observations: brain anatomy is individually distinctive, acting as a biometric trait, and reusable learned representations can leak information beyond their intended purpose. Focusing on longitudinal brain MRI, we evaluate whether frozen pretrained encoders enable both patient linkage and sensitive-attribute inference, explicitly separating linkability from re-identification threat models. By quantifying how much these representations expose patients to re-identification, we highlight the need for stronger safeguards and greater awareness as such models are released as open-source resources and made widely accessible.

3 Threat Model and Experimental Setup

This section describes the experimental setting used to evaluate privacy risks in pretrained brain MRI representations. We first introduce the evaluated encoders and longitudinal datasets, then formalize the attacker capabilities and privacy threats, and finally describe the embedding extraction, linkage, attribute-inference, and evaluation protocols.

3.1 Models and datasets

Models. We evaluate five pretrained encoders that differ in pretraining domain, input dimensionality, architecture, and objective (Table 1): AnatCL (Barbano et al., 2026), 3D-Neuro-SimCLR (Kaczmarek et al., 2025), Brain-MRI-SSL (Huang et al., 2025), DINO-ViT (Caron et al., 2021), and SAM-Med3D (Wang et al., 2024). The first three were pretrained specifically on brain MRI, whereas DINO-ViT was pretrained on natural images and SAM-Med3D on heterogeneous medical images. AnatCL, 3D-Neuro-SimCLR, and SAM-Med3D process full 3D volumes, while Brain-MRI-SSL and DINO-ViT operate on 2D images. Most models are trained with self-supervised objectives to learn general-purpose representations, whereas SAM-Med3D is trained specifically for segmentation in a supervised way. The models also use different loss objectives for their training. For instance, AnatCL is trained with weakly contrastive loss that combines two terms: an anatomical loss that brings closer subjects with similar anatomical descriptors, and a y-Aware loss that does

the same based on age. 3D-Neuro-SimCLR and Brain-MRI-SSL both use contrastive learning. 3D-Neuro-SimCLR uses SimCLR's NT-Xent loss, which pulls together two views of the same scan and pushes apart different scans in a batch. Brain-MRI-SSL uses MoCov3's symmetrized InfoNCE loss: each augmented view, encoded by a gradient-updated encoder, is contrasted against the other view produced by a momentum-updated encoder, with the other samples of the batch serving as negatives. In contrast, DINO-ViT relies on self-distillation. It minimizes the cross-entropy between a student network and a momentum teacher, without using negatives or labels. Finally, SAM-Med3D uses a supervised segmentation loss (Dice and cross-entropy). This diversity allows us to assess whether privacy risks persist across substantially different pretraining domains, architectures, and objectives.

Model	Pretraining objective	Input	Backbone	Pretraining domain	Dim.
AnatCL (Barbano et al., 2026)	Weakly sup. contrastive	3D	ResNet18	Brain MRI	512
3D-Neuro-SimCLR (Kaczmarek et al., 2025)	Contrastive (SimCLR)	3D	ResNet18	Brain MRI	512
Brain-MRI-SSL (Huang et al., 2025)	Contrastive (MoCov3)	2D	ResNet50	Brain MRI	2048
DINO-ViT (Caron et al., 2021)	Self-distillation	2D	ViT-B/8	Natural images	768
SAM-Med3D (Wang et al., 2024)	Sup. segmentation	3D	ViT	Multi-organ medical	384

Table 1: Overview of the pretrained encoders evaluated in this work.

Datasets. We use two T1-weighted brain MRI datasets that were not included in the training data of any of our tested models: OASIS-2 (Marcus et al., 2010) and MIRIAD (Malone et al., 2013), summarized in Table 2. Both are longitudinal datasets including Alzheimer's patients and healthy controls, but they differ substantially in scale and acquisition design. OASIS-2 includes more subjects (150) but fewer scans per subject (2–5 sessions), whereas MIRIAD includes fewer subjects (69) but with more repeated scans per subject (up to 12 scans over 9 visits, 708 scans in total). All MIRIAD scans are available as raw images that were acquired on the same scanner by the same radiographer. This acquisition consistency is advantageous for our privacy analysis because scanner/site membership and acquisition setting do not vary systematically across subjects, reducing the possibility that subject matching is driven by between-subject acquisition-condition differences. OASIS-2 scans are provided in a preprocessed form (motion-corrected, atlas-registered, skull-stripped, and intensity-corrected). Importantly, these are longitudinal acquisitions rather than near-duplicate acquisitions of the same scan: MIRIAD follow-ups are separated by approximately two weeks to two years, while OASIS-2 follow-ups extend over multiple years. The attack therefore evaluates whether subject-specific information persists across distinct acquisitions over time.

Characteristics	**OASIS-2**	**MIRIAD**
Subjects	150	69
Sex	58.7% F, 41.3% M	55.1% F, 44.9% M
Classes	72 NC, 64 AD, 14 Converted	46 AD, 23 CN
Format	NIfTI	DICOM (converted to NIfTI)
Resolution	256×256×128, 1.25 mm	256×256×124, 1.5 mm
Preprocessing	Yes	No
Scans per subject	at least 2 scans variable follow-ups (1–6+ years)	4–12 scans (up to 9 visits) intervals of 2 weeks–2 years

Abbreviations: AD = Alzheimer's Disease, MCI = Mild Cognitive Impairment, CN = Cognitively Normal, NC = Nondemented Controls. F = Female, M=Male.

Table 2: Characteristics of the OASIS-2 and MIRIAD datasets used in this study.

3.2 Threat Model

Figure 1 summarizes three threats associated with released brain MRI embeddings: patient linkability, re-identification, and attribute inference.

Released data and attacker access. For each evaluated model, let f denote its scan-to-embedding pipeline, comprising a publicly available frozen encoder and the preprocessing and aggregation operations described in Section 3.3. Each scan m is mapped to an embedding $f(m) \in \mathbb{R}^d$, where d is the representation dimension for that model. A data holder has a private MRI collection $\mathcal{D}_H =$

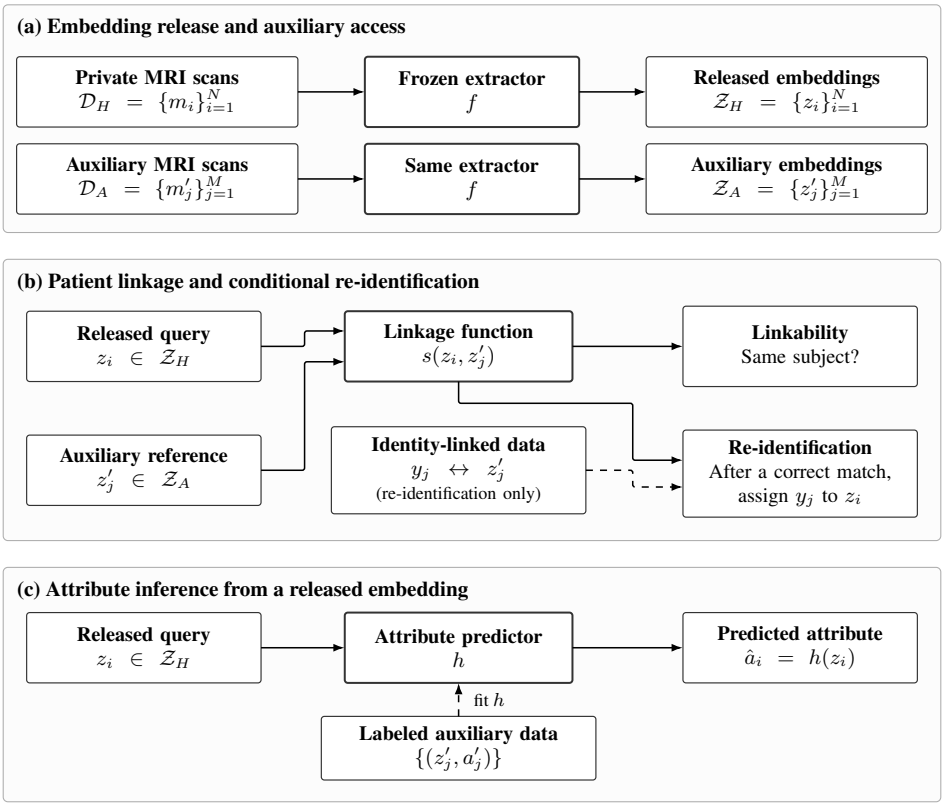


Figure 1: Threat models for released brain MRI representations. (a) The data holder and attacker apply the same frozen feature extractor f to private and auxiliary scans, respectively; either collection may contain multiple acquisitions per subject. The hospital's original scans remain unavailable to the attacker. (b) A linkage function assesses same-subject correspondence; re-identification additionally requires identity-linked reference information. (c) A predictor fitted to attribute-labeled auxiliary embeddings estimates a withheld attribute from a released embedding, without requiring successful linkage.

$\{m_i\}_{i=1}^N$ and releases only its embeddings $\mathcal{Z}_H = \{z_i\}_{i=1}^N$, with $z_i = f(m_i)$. The attacker obtains $\mathcal{Z}_H$ through an intended release or a data leak, but does not receive the underlying scans or their associated identities and attribute labels. The attacker knows f and has auxiliary scans $\mathcal{D}_A = \{m'_j\}_{j=1}^M$, from which they compute $\mathcal{Z}_A = \{z'_j\}_{j=1}^M$, with $z'_j = f(m'_j)$. Throughout the paper, we use the terms *embedding* and *representation* interchangeably to refer to $z_i = f(m_i)$.

Both collections may contain multiple acquisitions per individual: N and M count scans, not distinct subjects. For linkage and re-identification, the collections may overlap in subjects through different acquisitions, without giving the attacker access to the hospital's original scans. Such overlap could arise, for example, when a patient undergoes MRI examinations at multiple clinics and an attacker with access to scans from one clinic attempts to link them to embeddings released by another. Let u_i and u'_j denote the ground-truth subject labels of m_i and m'_j in a common label space. These labels establish matching correctness during evaluation; they are not supplied for released queries. Encoder weights remain fixed, although attack models may be fitted using labeled auxiliary data.

Linkability. Given a released embedding z_i and an auxiliary embedding z'_j, the attacker computes a similarity score $s(z_i, z'_j)$ to assess whether the corresponding scans originate from the same individual. The goal is to establish a correspondence between records belonging to the same subject, without requiring knowledge of that subject's real-world identity.

Re-identification. Re-identification builds on the same linkage procedure but assumes that the auxiliary reference records are additionally associated with identity-linked information y_j. The attacker ranks the auxiliary references according to their similarity to a released query z_i and assigns

the identity associated with the highest-scoring reference. Thus, the underlying matching operation is the same in both settings: linkability establishes that two records correspond to the same subject, whereas re-identification uses identity-linked auxiliary information to associate the released record with an identified individual.

Attribute inference. The attacker fits a predictor h using attribute-labeled auxiliary embeddings $\{(z'_j, a'_j)\}$ and estimates the withheld attribute a_i of a released scan as $\hat{a}_i = h(z_i)$. We evaluate inference of the subject's recorded sex. This attack requires auxiliary attribute labels, but neither identity-linked references nor a successful match to another acquisition from the query subject. These threats concern information retained in released representations and do not require the corresponding scans to have been used to pretrain the encoder.

3.3 Evaluation Protocol

Embedding extraction. We apply model-specific preprocessing as described in its paper and extract one embedding per MRI scan. For Brain-MRI-SSL, we select 20 axial slices at fixed relative positions within the 25%–75% depth range of each volume. For DINO-ViT, we select 10 slices in each of the axial, coronal, and sagittal orientations, yielding 30 slices per volume within the same relative range along each axis. In both pipelines, slice embeddings are averaged within each volume. Thus, $f(m_i)$ is a scan-level representation, not an aggregation of different acquisitions from the same subject.

Similarity scoring. Let $\mathcal{Q} \subseteq \{1, \ldots, N\}$ denote the query indices and $\mathcal{G}_i \subseteq \{1, \ldots, M\}$ the eligible reference indices for query i. For each query embedding z_i, we compute a Canberra distance-based similarity score $s_{ij} = s(z_i, z'_j) = 1/\big(1 + \sum_k \frac{|z_{ik} - z'_{jk}|}{|z_{ik}| + |z'_{jk}|}\big)$ against each eligible reference embedding z'_j, with larger scores indicating greater similarity. These scores are used to compute rank-1 linkage accuracy and estimate the same-subject and different-subject score distributions for the linkability measure. An ablation comparing Canberra and cosine similarities shows that Canberra gives the highest linkability in the evaluated setting (Appendix A.1); we therefore use it in the main analysis. To conduct the experiment, each scan is treated in turn as the released query. Its reference gallery contains every other scan in the dataset; the query scan itself is excluded. Scans from the same subject constitute genuine references, while scans from all other subjects constitute impostor references. Since every subject has multiple longitudinal acquisitions, each query has at least one genuine reference.

Attribute inference. For each evaluated encoder and dataset, we train a random forest predictor h to infer the subject's recorded sex from the extracted embeddings. We use subject-disjoint stratified cross-validation within each dataset: all scans from a given subject are assigned exclusively to either the training or held-out test, ensuring that no subject appears in both partitions. In each split, the training partition serves as the attribute-labeled auxiliary collection $\{(z'_j, a'_j)\}$, while the held-out partition serves as the collection of released query embeddings. The predictor is fitted to the auxiliary embedding–label pairs and applied to each held-out embedding z_i to obtain $\hat{a}_i = h(z_i)$. The held-out attribute labels a_i are used only to evaluate the predictions, and subject identifiers are not predictor inputs. No query–reference matching or Canberra scoring is used in this attack: attribute inference operates directly on the released embeddings through the fitted predictor (see Figure 1).

Metrics. We report $D^{\text{sys}}_{\leftrightarrow}$ as a score-based audit of linkage risk, rank-1 accuracy as the success rate of the explicit matching attack, and held-out classification performance for attribute inference.

Linkability score. We compute the system-level measure $D^{\text{sys}}_{\leftrightarrow}$ (Gomez-Barrero et al., 2018; Maouche et al., 2020). Let S denote the similarity score, H the same-subject hypothesis, and $\overline{H}$ the different-subject hypothesis, with prior $\pi = P(H)$. The posterior $P(H \mid S = t)$ is obtained from $p(t \mid H)$ and $p(t \mid \overline{H})$ by Bayes' rule. The linkability measure is

$$D^{\text{sys}}_{\leftrightarrow} = \int_{\mathbb{R}} p(t \mid H) \left[P(H \mid S = t) - P(\overline{H} \mid S = t)\right]_+ \mathrm{d}t, \tag{1}$$

where $[v]_+ = \max(v, 0)$. It lies in $[0, 1]$, with larger values indicating stronger same-subject separability under the specified scoring procedure. A data holder with repeated acquisitions and known subject correspondence can compute this measure before releasing embeddings; we therefore use it as an audit metric rather than as the outcome of the explicit matching attack. Its value remains

dependent on the scoring function, prior, and density estimation. We use all available same-subject and different-subject pairs to estimate the corresponding similarity-score distributions. Using the full set of available pairs provides the most complete empirical estimate of these distributions for each dataset, which are then used to compute the system-level linkability measure $D_{\leftrightarrow}^{\mathrm{sys}}$.

Rank-1 linkage accuracy (re-identification rate). We evaluate the explicit matching attack by ranking the eligible auxiliary references for each query and selecting the highest-scoring reference, $\hat{j}(i) \in \arg\max_{j \in \mathcal{G}_i} s_{ij}$. We report

$$\mathrm{ReID@1} = \frac{1}{|\mathcal{Q}|} \sum_{i \in \mathcal{Q}} \mathbf{1}\Big\{ u_i = u'_{\hat{j}(i)} \Big\}, \tag{2}$$

This scan-averaged metric measures successful subject linkage; its interpretation as re-identification additionally requires the identity-linked reference setting of Section 3.2. In the experiments, each scan is treated once as the released query and compared with every other scan in the dataset. The reported ReID@1is therefore the proportion of individual scans for which the highest-scoring reference belongs to the same subject, averaged over all scans in the dataset.

Attribute-inference performance. We report balanced accuracy and precision on held-out embeddings. Balanced accuracy gives equal weight to both recorded sex categories, with a performance of 0.5 for random guess for this binary task. Higher performance indicates that the released representations retain information predictive of the withheld attribute.

4 Privacy risk evaluation of pretrained brain representations

We now evaluate the privacy risks defined in Section 3.2. We first assess patient linkage and re-identification across the five pretrained encoders, then evaluate sensitive-attribute inference from the released embeddings. Finally, we discuss how differences in pretraining objectives may relate to the observed privacy risks.

4.1 Patient Linkage and Re-identification

	OASIS-2		**MIRIAD**	
Model	Linkability	ReID@1	Linkability	ReID@1
AnatCL	0.9312	0.9205	0.9714	0.9971
3D-Neuro-SimCLR	0.6571	0.6647	0.7938	0.9787
Brain-MRI-SSL	0.5647	0.5764	0.8460	0.9816
DINO-ViT	0.7796	0.8235	0.8761	0.9870
SAM-Med3D	0.2627	0.1394	0.5139	0.8958
FreeSurfer features (Baseline)	0.8134	0.7500	0.9797	0.9993
Subject-level Random guess	–	*0.0068*	–	*0.0145*

Table 3: Re-identification risk evaluation of the pretrained encoders on OASIS-2 and MIRIAD

Table 3 reports the results and some examples of the linkability plots are displayed in Appendix A.4. For MIRIAD, the risk of re-identifying patients is high across all models, with both a high re-identification rate and high linkability. SAM-Med3D shows the lowest risk, with a re-identification rate above 89% and a linkability of 0.5139, which remains substantial. The other four models achieve a re-identification rate of at least 98% and a linkability of at least 0.79, with AnatCL and Brain-MRI-SSL showing the highest values and therefore being the least private. On OASIS-2, results vary more across models: AnatCL and DINO-ViT are the riskiest, with linkabilities of 0.9312 and 0.7796 and re-identification rates of 92.05% and 82.35%, respectively, while SAM-Med3D is again the safest, with a linkability of 0.2627 and a re-identification rate of only 13.94%.

Although none of these models was trained for re-identification, their embeddings enable it, and most of them are therefore not privacy-safe. The lower risk of SAM-Med3D may be related to its training objective: it is the only model trained in a supervised manner for organ segmentation, so

its embeddings are optimized for a different purpose. All the other models are self-supervised and not trained with a specific downstream task in mind. DINO-ViT and SAM-Med3D are the only two models not specific to brain MRI, and DINO-ViT is not even trained on medical images. Surprisingly, it still poses a considerable re-identification risk, higher on OASIS-2 than brain-specific models such as Brain-MRI-SSL and 3D-Neuro-SimCLR.

Preprocessing may also contribute to these differences. Among the 3D models, AnatCL uses the most detailed pipeline: after nonlinear registration, it operates on gray matter maps, which describe the amount of gray matter at each location of the brain and therefore capture individual structural variations. 3D-Neuro-SimCLR uses skull-stripped and linearly registered volumes, and SAM-Med3D applies no brain-specific preprocessing. On OASIS-2, the risk follows this order, with a re-identification rate of 92.05% for AnatCL, 66.47% for 3D-Neuro-SimCLR, and 13.94% for SAM-Med3D. On MIRIAD, AnatCL and 3D-Neuro-SimCLR both reach near-complete re-identification, while SAM-Med3D remains the least risky.

AnatCL is among the riskiest models on both datasets (re-identification rate above 0.9). Performing re-identification using only the three descriptors, guiding its weakly contrastive training, yields results close to those obtained with the full AnatCL embeddings, suggesting that these anatomical features drive much of its risk. Overall, these results show that pretrained brain MRI encoders are not privacy-safe: their embeddings can be used to re-identify patients, even though none of these models was trained for this purpose. For a fair comparison with OASIS-2, we repeated the MIRIAD analysis with 2–5 scans per subject (Appendix A.3).

This linkability comparison can serve as a practical audit, allowing a hospital to decide whether a model requires additional protection before using it or sharing its embeddings. Risk varies across models: critical for AnatCL (above 0.93 on both datasets), consistently high for DINO-ViT (0.78–0.88), dataset-dependent for 3D-Neuro-SimCLR and Brain-MRI-SSL (0.79 and 0.85 on MIRIAD vs. 0.66 and 0.56 on OASIS-2), and lowest for SAM-Med3D, which still reaches 0.51 on MIRIAD and cannot be considered safe.

We also evaluate an ad-hoc baseline representation following Bajar et al. (2026), which consists of 15 features derived from FreeSurfer rather than a learned embedding. On MIRIAD, this representation reaches a linkability of 0.9797 and a re-identification rate of 99.93%. On OASIS-2, it remains highly identifying, with a linkability of 0.8134 and a re-identification rate of 75.00%. These 15 features are manually extracted anatomical measurements known to be highly distinctive. Yet deep learning models can extract more discriminative features unintentionally and reach similar and sometimes higher re-identification risk.

4.2 Sensitive Attribute Inference

Beyond re-identification, we investigate whether these model embeddings also leak demographic attributes, specifically sex. We train a Random Forest classifier on top of the extracted embeddings to predict the subject's sex. The classifier is evaluated using subject-disjoint stratified cross-validation, and we report balanced accuracy and precision as our metrics (Table 4). High classification performance indicates that sex-discriminative information is embedded in the representation, even though the foundation model was never explicitly trained for this task. This supports the claim that these embeddings encode more identity-linked information than intended, raising a further type of privacy risk beyond re-identification alone. Importantly, linkage and attribute inference do not induce the same ordering across representations. This indicates that privacy risk is multidimensional: reducing subject linkability does not necessarily reduce other sensitive information retained by the representation. Appendix A.2 reports the results when the classifier is trained on one dataset and tested on the other.

	OASIS-2		MIRIAD	
Model	Bal. Acc.	Precision	Bal. Acc.	Precision
AnatCL	0.6772	0.6965	0.6946	0.6983
3D-Neuro-SimCLR	0.7011	0.7075	0.7789	0.7806
Brain-MRI-SSL	0.9093	0.9153	0.9546	0.9568
DINO-ViT	0.7894	0.8100	0.8716	0.8704
SAM-Med3D	0.8632	0.8735	0.8850	0.8887

Table 4: Sex classification performance on OASIS-2 and MIRIAD.

4.3 Implications of pretraining objectives for privacy

The evaluated encoders differ substantially in how their representations are learned, yet none is explicitly optimized to match longitudinal acquisitions from the same patient. This contrasts with biometric representation learning, where identity consistency is part of the objective. For example, Fallahi et al. (2026) train EEG encoders using supervised metric learning that brings recordings from the same subject across sessions closer while separating different subjects. In our setting, patient identity is never provided as supervision.

The contrastive models also differ in their training mechanism. Both 3D-Neuro-SimCLR and Brain-MRI-SSL treat augmented views of the same scan as positives and other scans as negatives, but 3D-Neuro-SimCLR optimizes an NT-Xent objective with a single shared encoder (Kaczmarek et al., 2025), whereas Brain-MRI-SSL follows MoCov3, in which keys are produced by a momentum encoder (Huang et al., 2025). These objectives are instance-based rather than identity-aware and do not explicitly encourage different acquisitions of the same subject to be close. If such acquisitions occur among the negatives, the objective may even discourage their similarity, depending on the pretraining sampling strategy. AnatCL instead guides contrastive learning using anatomical descriptors and age (Barbano et al., 2026); DINO-ViT uses self-distillation across augmented views (Caron et al., 2021); and SAM-Med3D is optimized for supervised segmentation (Wang et al., 2024).

Despite these different learning signals, all evaluated representations retain subject-specific information supporting longitudinal linkage, although the magnitude varies across models and datasets. This suggests that such leakage is not specific to a particular contrastive formulation or the use of explicit negatives, but can arise as a by-product of preserving stable anatomical, structural, or instance-specific characteristics. AnatCL provides a direct example: although not designed for identity matching, its anatomical descriptors alone already support strong re-identification. However, because the encoders also differ in architecture, pretraining data, and preprocessing, differences in linkability cannot be attributed to the training objective alone.

These findings have implications for both privacy-oriented representation learning and AI governance. An identity-agnostic objective should not be assumed to produce an identity-invariant representation, motivating privacy-aware training or post-hoc sanitization of released embeddings. For high-risk AI systems, Article 9 of the EU AI Act requires an iterative process for identifying and evaluating reasonably foreseeable risks to health, safety, and fundamental rights and adopting appropriate risk-management measures (European Union, 2024). Privacy audits such as the linkage and attribute-inference evaluations considered here can provide technical evidence of residual risks before deployment or release, complementing the data-protection assessment of linkage and inference risks discussed earlier (European Data Protection Board, 2026), without by themselves establishing regulatory compliance or anonymity.

5 Conclusion

We audited whether released representations from frozen pretrained encoders retain patient-specific information in longitudinal brain MRI. Across five encoders and two T1-weighted datasets, we find that several representations support strong patient linkage and re-identification under the evaluated threat model, despite the absence of an identification objective during encoder training. We also show that released embeddings retain information predictive of the subject's recorded sex, extending the privacy risk beyond identity linkage alone. These findings indicate that releasing learned representations instead of raw brain MRI does not, by itself, eliminate linkage or attribute-inference risk and is therefore insufficient on its own to establish anonymity.

Our study is limited to two longitudinal T1-weighted datasets, five pretrained encoders, and one inferred attribute. Moreover, the evaluated models differ simultaneously in architecture, pretraining domain, and objective, so our comparisons should not be interpreted as isolating the causal effect of any single design choice. Future work should move from measuring privacy risk to reducing it through privacy-aware training objectives and post-hoc sanitization methods that suppress identity-bearing information in released embeddings. An important next step is to evaluate how effectively these approaches reduce privacy risks while preserving downstream utility.

AI USE STATEMENT

Generative AI tools were used in this work to support language refinement and improve the clarity of the manuscript, as well as to assist with selected parts of the experimental implementation and figure generation. These tools were not used to define the research questions, perform the literature review, or establish the scientific conclusions of the study. All AI-assisted text and code were manually reviewed by the authors, and the resulting experiments and visualizations were independently checked for correctness and consistency. The authors remain fully responsible for the content of the manuscript, the reported results, and all material developed with the assistance of generative AI.

ETHICS STATEMENT

This work studies privacy risks in released brain MRI representations using existing research datasets and does not attempt to identify participants in the real world. Our re-identification evaluation is performed at the dataset-subject level: the experiments measure whether different scans can be linked to the same subject, while attribution to a real-world identity is considered only as an explicit assumption in the threat model when identity-linked auxiliary information is available. We do not recover names or other civil identifiers, nor do we release patient-level linkage results. The purpose of the study is to characterize potential privacy risks and inform safer sharing of medical-image representations; all reported results are presented in aggregate.

REPRODUCIBILITY STATEMENT

We provide the code required to reproduce the feature-extraction, linkage, re-identification, attribute-inference, and evaluation pipelines in an anonymous repository: `https://anonymous.4open.science/r/mri-reid-privacy-86C3/README.md`. The repository includes the model-specific embedding-extraction procedures, evaluation scripts, metric implementations, and plotting code used in the experiments. The datasets evaluated in this work are publicly available from their original providers and are not redistributed. Additional implementation details, including model configurations, data-processing choices, and ablation settings, are documented in the appendix and repository.

REFERENCES

D. Abramian and A. Eklund. Refacing: reconstructing anonymized facial features using GANs. *Medical Image Analysis*, 54:322–332, 2019.

Adnan Ahmed, Memona Nazir, Ajmaal Jami, Muhammad Tahir, Mahnoor Khalil, and Tahir Baig. Magnetic resonance imaging and computed tomography in emergency assessment of patients with suspected acute stroke: A prospective comparison. *Pakistan Journal of Medical and Health Sciences*, 15:1898–1902, 06 2021. doi: 10.53350/pjmhs211561898.

Imen Bajar, Mohamed Maouche, and Carole Frindel. Beyond the mask: The illusion of privacy in defaced brain mri. In *2026 IEEE 23rd International Symposium on Biomedical Imaging (ISBI)*, pp. 1–5, 2026. doi: 10.1109/ISBI61048.2026.11516026.

Carlo Alberto Barbano et al. Anatomical foundation models for brain MRIs. *Pattern Recognition Letters*, 199:178–184, 2026. ISSN 0167-8655. doi: 10.1016/j.patrec.2025.11.028. URL `https://www.sciencedirect.com/science/article/pii/S0167865525003848`.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent,

Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. On the opportunities and risks of foundation models, 2022. URL https://arxiv.org/abs/2108.07258.

A. Boutet, C. Frindel, and M. Maouche. Towards an evolution in the characterization of the risk of re-identification of medical images. In *Proc. IEEE Int. Conf. on Big Data*, pp. 1–6, Sorrento, Italy, 2023.

California State Legislature. California civil code, section 1798.140. California Legislative Information, 2026. URL https://leginfo.legislature.ca.gov/faces/codes_displaySection.xhtml?lawCode=CIV&sectionNum=1798.140.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9650–9660, 2021.

Joseph Cox, Peng Liu, Skylar E. Stolte, Yunchao Yang, Kang Liu, Kyle B. See, Huiwen Ju, and Ruogu Fang. Brainsegfounder: Towards 3d foundation models for neuroimage segmentation, 2024. URL https://arxiv.org/abs/2406.10395.

Justin Engelmann, Siegfried K. Wagner, Pearse A. Keane, et al. Matters arising: Near-identical images, not foundation models, explain purported re-identification of patients from medical imaging. *npj Digital Medicine*, 9:159, 2026. doi: 10.1038/s41746-026-02440-9.

European Data Protection Board. Guidelines 02/2026 on anonymisation. Technical Report 02/2026, European Data Protection Board, July 2026. URL https://www.edpb.europa.eu/public-consultations/guidelines-022026-on-anonymisation_en.

European Parliament and Council of the European Union. Regulation (EU) 2016/679 of the european parliament and of the council. Official Journal of the European Union, 2016. URL https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32016R0679. CELEX: 32016R0679.

European Union. Regulation (eu) 2024/1689 of the european parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act), 2024. URL https://eur-lex.europa.eu/eli/reg/2024/1689/oj. Official Journal of the European Union.

Matin Fallahi, Patricia Arias-Cabarcos, and Thorsten Strufe. Advancing brainwave-based biometrics: a large-scale, multi-session evaluation. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2026.

Giovanni B. Frisoni, Nick C. Fox, Clifford R. Jack, Philip Scheltens, and Paul M. Thompson. The clinical use of structural MRI in Alzheimer disease. *Nature Reviews Neurology*, 6(2):67–77, 2010. doi: 10.1038/nrneurol.2009.215.

Salah Ghamizi, Georgia Kanli, Yu Deng, Magali Perquin, and Olivier Keunen. Brain imaging foundation models, are we there yet? a systematic review of foundation models for brain imaging and biomedical research, 2025. URL https://arxiv.org/abs/2506.13306.

Marta Gomez-Barrero, Javier Galbally, Christian Rathgeb, and Christoph Busch. General framework to evaluate unlinkability in biometric template protection systems. *IEEE Transactions on Information Forensics and Security*, 13(6):1406–1420, 2018. doi: 10.1109/TIFS.2017.2788000.

Government of Canada. Personal information protection and electronic documents act. S.C. 2000, c. 5, 2000. URL https://laws-lois.justice.gc.ca/eng/acts/P-8.6/. Personal Information Protection and Electronic Documents Act (PIPEDA).

O. F. Gulban, D. Nielson, R. Poldrack, J. Lee, C. Gorgolewski, Vanessasaurus, and S. Ghosh. poldracklab/pydeface: v2.0.0. *Zenodo*, 2020. doi: 10.5281/zenodo.3524401.

Xinhao Huang, Zihao Wang, Weichen Zhou, Kexin Yang, Kaihua Wen, Haiguang Liu, Shoujin Huang, and Mengye Lyu. Tailored self-supervised pretraining improves brain MRI diagnostic models. *Computerized Medical Imaging and Graphics*, 123:102560, 2025. ISSN 0895-6111. doi: 10.1016/j.compmedimag.2025.102560.

Y. Jeong, H. Kim, and Y. Kwak. Reconstruction of facial features from defaced MRI using deep learning. In *Proc. IEEE Int. Symp. Biomed. Imaging (ISBI)*, pp. 1–4, 2020.

Emily Kaczmarek, Justin Szeto, Brennan Nichyporuk, and Tal Arbel. Building a general SimCLR self-supervised foundation model across neurological diseases to advance 3D brain MRI diagnoses. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pp. 1321–1330, 2025.

Sergey Korolev, Amir Safiullin, Mikhail Belyaev, and Yulia Dodonova. Residual and plain convolutional neural networks for 3d brain mri classification. In *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*, pp. 835–838, 2017. doi: 10.1109/ISBI.2017.7950647.

Peirong Liu, Oula Puonti, Xiaoling Hu, Karthik Gopinath, Annabel Sorby-Adams, Daniel C. Alexander, W. Taylor Kimberly, and Juan E. Iglesias. A modality-agnostic multi-task foundation model for human brain imaging, 2025. URL https://arxiv.org/abs/2509.00549.

Ian B. Malone, David Cash, Gerard R. Ridgway, David G. Macmanus, Sebastien Ourselin, Nick C. Fox, and Jonathan M. Schott. MIRIAD: Public release of a multiple time point Alzheimer's MR imaging dataset. *NeuroImage*, 70:33–36, 2013. doi: 10.1016/j.neuroimage.2012.12.044.

Mohamed Maouche, Brij Mohan Lal Srivastava, Nathalie Vauquier, Aurélien Bellet, Marc Tommasi, and Emmanuel Vincent. A comparative study of speech anonymization metrics. In *Proc. Interspeech 2020*, pp. 1708–1712, 2020.

D. S. Marcus, A. F. Fotenos, J. G. Csernansky, J. C. Morris, and R. L. Buckner. Open access series of imaging studies: longitudinal MRI data in nondemented and demented older adults. *Journal of Cognitive Neuroscience*, 22(12):2677–2684, 2010. doi: 10.1162/jocn.2009.21407. URL https://doi.org/10.1162/jocn.2009.21407.

Matia Martucci, Rosellina Russo, Francesco Schimperna, et al. Magnetic resonance imaging of primary adult brain tumors: State of the art and future perspectives. *Biomedicines*, 11(2):364, 2023. doi: 10.3390/biomedicines11020364.

Moona Mazher, Geoff J. M. Parker, and Daniel C. Alexander. Towards generalisable foundation models for brain mri, 2025. URL https://arxiv.org/abs/2510.23415.

S. Mitew, T. Ma, D. Smith, and C. M. Wiggins. pyfacewipe: a new defacing tool for almost any MRI contrast. *Magnetic Resonance Materials in Physics, Biology and Medicine (MAGMA)*, 2024.

Heba Mohsen, El-Sayed A. El-Dahshan, El-Sayed M. El-Horbaty, and Abdel-Badeeh M. Salem. Classification using deep learning neural networks for brain tumors. *Future Computing and Informatics Journal*, 3(1):68–71, 2018. ISSN 2314-7288. doi: 10.1016/j.fcij.2017.12.001.

Giacomo Nebbia, Sourav Kumar, Stephen McNamara, Christopher Bridge, John Campbell, Michael Chiang, Naresh Mandava, Praveer Singh, and Jayashree Kalpathy-Cramer. Re-identification of patients from imaging features extracted by foundation models. *npj Digital Medicine*, 8, 07 2025. doi: 10.1038/s41746-025-01801-0.

L. Puglisi, F. Barkhof, D. C. Alexander, G. J. M. Parker, A. Eshaghi, and D. Ravì. Deepbrainprint: A novel contrastive framework for brain MRI re-identification. In *Alzheimer's Disease Neuroimaging Initiative (ADNI)*, 2023.

RSNA. Privacy risks in radiology ai. https://www.rsna.org/news/2026/may/privacy-risks-in-radiology-ai/. Accessed: 13/08/2026.

Congzheng Song and Vitaly Shmatikov. Overlearning reveals sensitive attributes. *arXiv preprint arXiv:1905.11742*, 2019.

Divyanshu Tak, B.A. Gormosa, A. Zapaishchykova, et al. A generalizable foundation model for analysis of human brain mri. *Nature Neuroscience*, 2026. doi: 10.1038/s41593-026-02202-6. URL https://doi.org/10.1038/s41593-026-02202-6.

Hidemasa Takao, Naoto Hayashi, and Kuni Ohtomo. Brain morphology is individual-specific information. *Magnetic resonance imaging*, 33(6):816–821, 2015.

Sana Tonekaboni, Lena Stempfle, Sasha Ronaghi, Corinna Coupette, I. Glenn Cohen, Emily Alsentzer, and Marzyeh Ghassemi. Protecting patient privacy in clinical foundation models: Technical and legal perspectives. *arXiv preprint arXiv:2608.07705*, 2026.

S. A. Valizadeh, P. H. Haghparast, J. Ye, and T. M. Palmer. Identification of individuals using brain anatomical features. *Scientific Reports*, 8:5611, 2018.

C. Wachinger, P. Golland, and M. Reuter. Brainprint: identifying subjects by their brain. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 8675, pp. 41–48. Springer, 2014.

Haoyu Wang, Sizheng Guo, Jin Ye, Zhongying Deng, Junlong Cheng, Tianbin Li, Jianpin Chen, Yanzhou Su, Ziyan Huang, Yiqing Shen, Bin Fu, Shaoting Zhang, Junjun He, and Yu Qiao. SAM-Med3D: Towards general-purpose segmentation models for volumetric medical images. *arXiv preprint arXiv:2310.15161*, 2024.

A Additional Results

A.1 Ablation: Distance Metric

To assess the sensitivity of our results to the similarity function, we compare Canberra-based and cosine similarity in Table 5. Canberra yields higher linkability for every encoder on both datasets, indicating greater separation between the same-subject and different-subject score distributions under this metric. Rank-1 matching is less sensitive to this choice: neither similarity function consistently dominates across all model–dataset combinations, although Canberra achieves slightly higher performance overall. We therefore retain Canberra for the main analysis because it consistently provides the stronger linkability audit while supporting comparable rank-1 matching results.

	OASIS-2				MIRIAD			
	Linkability		ReID@1		Linkability		ReID@1	
Model	Canberra	Cosine	Canberra	Cosine	Canberra	Cosine	Canberra	Cosine
AnatCL	**0.9312**	0.5707	**0.9205**	0.8716	**0.9714**	0.6017	**0.9971**	**0.9971**
3D-Neuro-SimCLR	**0.6571**	0.3502	**0.6647**	0.4794	**0.7938**	0.3888	**0.9787**	0.9176
Brain-MRI-SSL	**0.5647**	0.4204	0.5764	**0.6461**	**0.8460**	0.6674	**0.9816**	0.9788
DINO-ViT	**0.7796**	0.5694	0.8235	**0.8412**	**0.8761**	0.8677	0.9870	**0.9899**
SAM-Med3D	**0.2627**	0.2177	0.1394	**0.1528**	**0.5139**	0.4691	0.8958	**0.9085**

Table 5: Linkability and rank-1 matching under Canberra-based and cosine similarity. Bold marks the higher (i.e., stronger attack) value for each pair.

A.2 Cross Dataset Attribute Inference

In Section 4.2, we showed that the evaluated embeddings retain information predictive of recorded sex when the classifier is trained and tested within the same dataset. We now examine whether this information generalizes across cohorts by training on one dataset and testing on the other. Table 6 reports results in both directions, where A→B denotes training on dataset A and testing on dataset B. Balanced accuracy decreases for every encoder under cross-dataset evaluation, with most models approaching chance level (0.5) in at least one direction. SAM-Med3D is the main exception, retaining balanced accuracy of 0.81 for OASIS-2→MIRIAD and 0.79 for MIRIAD→OASIS-2, compared with 0.86–0.89 within datasets. Interestingly, this differs from the linkage results, where SAM-Med3D exhibited the lowest linkability and rank-1 matching among the evaluated encoders, further showing that different privacy risks need not induce the same ordering across representations. Cross-dataset performance is also asymmetric: training on MIRIAD and testing on OASIS-2 performs better than the reverse for all encoders except SAM-Med3D. This asymmetry suggests sensitivity to dataset-specific factors such as preprocessing, acquisition conditions, and cohort characteristics; however, our experiments do not isolate which of these factors drives the observed domain shift.

A.3 Ablation: Impact of Number of Scans

MIRIAD contains more repeated scans per subject than OASIS-2, which may increase the success of a retrieval-based matching attack by providing more genuine references in the gallery. To assess the effect of this difference, we repeat the MIRIAD analysis after randomly subsampling each subject to between 2 and 5 scans, matching the range observed in OASIS-2. We repeat the subsampling ten times and report the mean and standard deviation for each model.

Table 7 shows that capping the number of scans lowers the re-identification rate for all models, but to very different extents. AnatCL, DINO-ViT, and Brain-MRI-SSL are only mildly affected, with drops of 0.9, 2.0, and 4.7 percentage points, respectively. 3D-Neuro-SimCLR drops by 6.3 points and SAM-Med3D, already the least risky model, drops by far the most, by 27.1% (from 89.58% to 62.46%). Linkability, in contrast, remains almost unchanged for all models except 3D-Neuro-SimCLR, varying by less than 0.01. Although SAM-Med3D loses 27.1 points in re-identification rate, its linkability remains essentially stable. Unlike rank-1 matching, linkability measures the separation between empirical same-subject and different-subject score distributions rather than retrieval

	Within-dataset				Cross-dataset			
	OASIS-2		MIRIAD		OASIS-2 → MIRIAD		MIRIAD → OASIS-2	
Model	Bal. Acc.	Prec.	Bal. Acc.	Prec.	Bal. Acc.	Prec.	Bal. Acc.	Prec.
AnatCL	0.6772	0.6965	0.6946	0.6983	0.4784	0.4769	0.5130	0.5132
3D-Neuro-SimCLR	0.7011	0.7075	0.7789	0.7806	0.5000	0.2763	0.6170	0.6198
Brain-MRI-SSL	**0.9093**	**0.9153**	**0.9546**	**0.9568**	0.5032	0.7783	0.6351	0.7759
DINO-ViT	0.7894	0.8100	0.8716	0.8704	0.5151	0.6039	0.5464	0.7682
SAM-Med3D	0.8632	0.8735	0.8850	0.8887	**0.8098**	**0.8221**	**0.7859**	**0.8168**

Table 6: Sex inference from frozen embeddings. Within-dataset results use a patient-disjoint split; cross-dataset results train on one cohort and test on the other. Best value per column in bold.

	MIRIAD (full)		MIRIAD (capped 2–5)	
Model	Linkability	ReID@1	Linkability	ReID@1
AnatCL	0.9714	0.9971	0.9653 ± 0.0143	0.9883 ± 0.0064
3D-Neuro-SimCLR	0.7938	0.9787	0.7829 ± 0.0214	0.9160 ± 0.0138
Brain-MRI-SSL	0.8460	0.9816	0.8420 ± 0.0166	0.9343 ± 0.0157
DINO-ViT	0.8761	0.9870	0.8778 ± 0.0182	0.9675 ± 0.0085
SAM-Med3D	0.5139	0.8958	0.5204 ± 0.0185	0.6246 ± 0.0298

Table 7: Re-identification risk on MIRIAD: full dataset vs. capped (2–5 scans per subject). Capped results are reported as mean $\pm$ std over 10 random subsamplings.

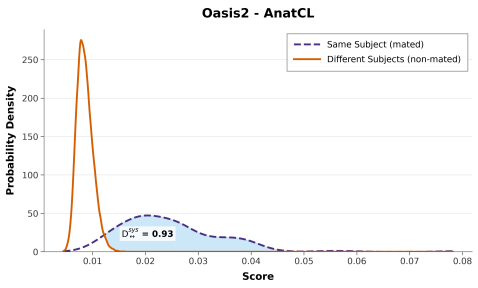
success from a gallery. Subsampling can nevertheless alter these distributions by changing which acquisitions and pairwise comparisons are retained.

Two factors may explain why MIRIAD remains riskier than OASIS-2 even after capping the number of scans. First, MIRIAD scans of the same subject are acquired close together in time and are therefore very similar, whereas OASIS-2 scans are spaced at least one year apart. Second, all MIRIAD scans are released without preprocessing. OASIS-2 is released already preprocessed (motion-corrected, registered, and skull-stripped). This prior preprocessing may remove identifying information that is still present in the raw MIRIAD scans.

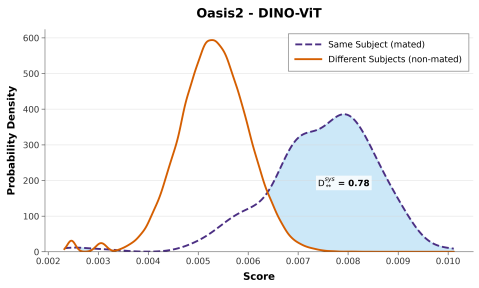
The number of scans per subject therefore contributes to re-identification, but other dataset-specific factors also shape the results. This does not change our overall conclusion: re-identification remains high for all models, confirming our finding that the embeddings of these pretrained encoders are not privacy-safe.

A.4 Score Distributions and Linkability

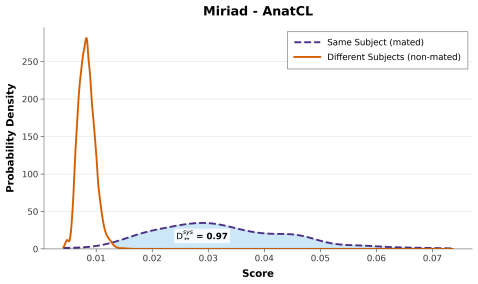
Figure 2 shows representative distributions of Canberra similarity scores for mated (same-subject) and non-mated (different-subject) pairs for AnatCL and DINO-ViT on both datasets. The shaded region indicates where the mated density exceeds the non-mated density, and the corresponding system-level linkability $D_{\leftrightarrow}^{\text{sys}}$ is reported in each panel.



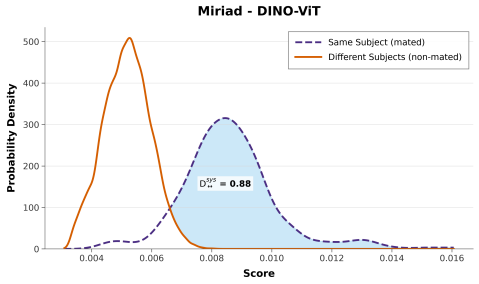
(a) AnatCL, OASIS-2



(b) DINO-ViT, OASIS-2



(c) AnatCL, MIRIAD



(d) DINO-ViT, MIRIAD

Figure 2: Representative mated and non-mated score distributions (KDE) for AnatCL and DINO-ViT on OASIS-2 (top) and MIRIAD (bottom), with the linkability $D_{\leftrightarrow}^{\mathrm{sys}}$.