
White Rose Research Online

eprints@whiterose.ac.uk
https://eprints.whiterose.ac.uk
Universities of Leeds, Sheffield and York

Deposited via The University of York.

White Rose Research Online URL for this paper:
https://eprints.whiterose.ac.uk/id/eprint/238949/

Preprint:
Deo, Yash, Jia, Yan, Lassila, Toni et al. (2026) A Calibrated Memorization Index (MI) for Detecting Training Data Leakage in Generative MRI Models. [Preprint]

https://doi.org/10.48550/arXiv.2602.13066

Reuse
Items deposited in White Rose Research Online are protected by copyright, with all rights reserved unless indicated otherwise. They may be downloaded and/or printed for private study, or other acts as permitted by national copyright laws. The publisher or other rights holders may allow further reproduction and re-use of the full text version. This is indicated by the licence information on the White Rose Research Online record for the item.

Takedown
If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.



A CALIBRATED MEMORIZATION INDEX (MI) FOR DETECTING TRAINING DATA LEAKAGE IN GENERATIVE MRI MODELS

Yash Deo[1], *Yan Jia*[1], *Toni Lassila*[2], *Victoria J Hodge*[1],
Alejandro F Frangi[5,6], *Chenghao Qian*[2], *Siyuan Kang*[4], *Ibrahim Habli*[1]

[1] Department of Computer Science, University of York, York, UK
[2] School of Computer Science, University of Leeds, Leeds, UK
[4] Department of Computing and Mathematics, Manchester Metropolitan University, Manchester, UK
[5] Department of Computer Science, University of Manchester, Manchester, UK
[6] Department of Cardiovascular Sciences, KU Leuven, Leuven, Belgium

arXiv:2602.13066v1 [cs.CV] 13 Feb 2026

© 2026 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

ABSTRACT

Image generative models are known to duplicate images from the training data as part of their outputs, which can lead to privacy concerns when used for medical image generation. We propose a calibrated per-sample metric for detecting memorization and duplication of training data. Our metric uses image features extracted using an MRI foundation model, aggregates multi-layer whitened nearest-neighbor similarities, and maps them to a bounded *Overfit/Novelty Index* (ONI) and *Memorization Index* (MI) scores. Across three MRI datasets with controlled duplication percentages and typical image augmentations, our metric robustly detects duplication and provides more consistent metric values across datasets. At the sample level, our metric achieves near-perfect detection of duplicates.

1 Introduction

Deep learning models for synthetic image generation are increasingly used to address data scarcity, class imbalance, and patient privacy concerns [1, 2]. However, these models can exhibit problematic behaviors, such as memorizing training data (data leakage) [3]. This risk is especially concerning in medical applications, as it can compromise patient privacy by reproducing personally identifiable data.

Validating the safety of these generative models requires the use of no-reference image-quality metrics (NRIQMs), since paired ground-truth images are typically unavailable for direct comparison. Distributional fidelity metrics such as FID and MMD [4], while useful for assessing visual quality, are often *counter-diagnostic* for leakage. When duplicated training examples contaminate an evaluation set, the empirical test distribution moves closer to the training distribution, causing these scores to *improve* even as memorization increases [5].

Metrics designed specifically for duplication detection, such as the CT-score [6], authenticity metric [7], or Vendi score [8], also prove inadequate. These metrics are typically developed using features from models (like InceptionV3) trained on *natural images*. This feature space transfers poorly to medical images, which are characterized by high anatomical regularity and limited appearance diversity. In medical images, clinically relevant variation is often subtle and localized. As a result, metrics based on natural-image features tend to misinterpret acquisition noise as "diversity" or are easily fooled by simple augmentations, failing to expose near-duplicates [5].

Furthermore, comparing different no-reference image quality metrics (NRIQMs) is complicated by arbitrary scaling effects that vary across datasets. An ideal metric for distribution-level comparisons should be robust to small perturbations (e.g., noise, small rotations) and provide a calibrated, consistent range of values to be interpretable (e.g., a score near 0 indicating novelty and a score near 1 indicating an exact copy) across different datasets. To our knowledge, no existing NRIQM for medical data memorization satisfies all these criteria.

We propose a novel, calibrated, multi-scale metric for detecting data memorization in MRI datasets. Our method uses features extracted from a domain-specific MRI foundation model (MRI-CORE [9]) at multiple scales, capturing both fine-grained textures and gross anatomical structures. We validate our metric on three public MRI datasets (brain, knee, and spine) [10, 11, 12] and show that it detects data leakage more consistently than generic metrics and is significantly more robust to common data augmentations. While tested on MRI, our methodology is general and applicable to other medical imaging modalities, provided an appropriate domain-

specific feature space is used. The implementation code is available at :- https://github.com/YashDeo-York/Mem-index

2 Methodology

We introduce a calibrated, multi-scale metric for slice-level memorisation detection in MRI datasets. The pipeline comprises: (i) multi-layer transformer feature extraction, (ii) per-layer whitening and similarity computation, (iii) multi-scale aggregation and (iv) calibration via empirical null distribution. A high level overview of the Methodology is shown in Figure 1.

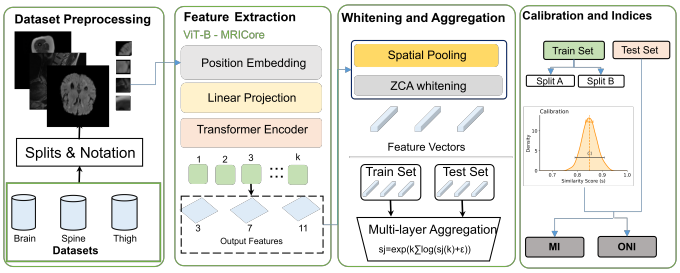

Fig. 1. Overview of our methodology to calculate the Memorisation Index (MI) and the Overfit-Novelty Index (ONI)

2.1 Feature Extraction

We employ the MRI-CORE [9] foundation model (SAM ViT-B encoder) pretrained on diverse MRI datasets. In ViT, early layers tend to exhibit localized attention similar to early convolutions in CNNs, while deeper layers progressively increase their 'attention distance' to integrate information more globally [13]; Hence , we extract features from transformer blocks $k \in \{3, 7, 11\}$, (from 11 total blocks) capturing representations from early texture patterns to high-level anatomical structures. For block k, spatial token outputs are reshaped to $\mathbf{F}^{(k)} \in \mathbb{R}^{C_k \times h_k \times w_k}$ and spatially pooled via adaptive average pooling:

$$\mathbf{f}^{(k)} = \frac{1}{h_k w_k} \sum_{u,v} \mathbf{F}^{(k)}[:, u, v] \in \mathbb{R}^{C_k} \tag{1}$$

We denote train/test pooled features by $\mathcal{F}^{(k)}_{\text{train}} = \{\mathbf{f}^{(k)}_i\}$ and $\mathcal{F}^{(k)}_{\text{test}} = \{\mathbf{g}^{(k)}_j\}$.

2.2 Whitening and Layer-wise Similarity

To render cross-layer similarities comparable, we apply ZCA whitening [14] estimated on train features per layer. ZCA whitening decorrelates features while maintaining the original feature space, unlike PCA which rotates to principal components. With mean $\boldsymbol{\mu}^{(k)}$ and covariance $\mathbf{C}^{(k)}$:

$$\hat{\mathbf{f}}^{(k)} = (\mathbf{f}^{(k)} - \boldsymbol{\mu}^{(k)})\mathbf{W}^{(k)}, \quad \mathbf{W}^{(k)} = (\mathbf{C}^{(k)} + \epsilon \mathbf{I})^{-1/2} \tag{2}$$

where $\epsilon = 10^{-6}$ ensures numerical stability. Whitened vectors are ℓ_2-normalized. For each test sample j at layer k, we compute the maximum cosine similarity to any training sample:

$$s^{(k)}_j = \max_i \left\langle \frac{\hat{\mathbf{g}}^{(k)}_j}{\|\hat{\mathbf{g}}^{(k)}_j\|_2}, \frac{\hat{\mathbf{f}}^{(k)}_i}{\|\hat{\mathbf{f}}^{(k)}_i\|_2} \right\rangle \tag{3}$$

2.3 Multi-Scale Aggregation

Given the per-layer nearest–neighbour similarities $s^{(k)}_j \in [0, 1]$ computed above for layers $k \in K$ (here $K = \{3, 7, 11\}$), we aggregate them into a single sample-level score via a geometric mean:

$$s_j = \left(\prod_{k \in K} (s^{(k)}_j + \epsilon) \right)^{1/|K|} = \exp\left(\frac{1}{|K|} \sum_{k \in K} \log\left(s^{(k)}_j + \epsilon\right) \right), \tag{4}$$

and define the distance $d_j = 1 - s_j$ and set $\epsilon = 10^{-6}$.

This log–average acts as a product-of-experts, rewarding *cross-scale consensus*: a sample is scored as highly similar only if it is simultaneously close across early, mid, and late features. Relative to an arithmetic mean, it suppresses single-layer outliers and penalizes cases where any scale fails to find a close neighbor. For diagnostics, we also retain per-layer neighbor indices and a consensus count (number of layers selecting the same neighbor).

2.4 Calibration and Memorization Index

To establish significance thresholds, we construct an empirical null distribution by bootstrapping the training set. We randomly sample independent subsets A, B (50% of training data, $n = 10$ iterations) and compute aggregated similarities $s_{A,B}$ using the same whitening and aggregation pipeline. The null parameters are:

$$\mu_{\text{null}} = \mathbb{E}[s_{A,B}], \quad \sigma_{\text{null}} = \sqrt{\text{Var}(s_{A,B}) + 10^{-8}} \tag{5}$$

Test similarities are standardized into a *Memorization Index* (MI) and mapped to a bounded *Overfit/Novelty Index* (ONI):

$$\text{MI}_j = \frac{s_j - \mu_{\text{null}}}{\sigma_{\text{null}}}, \quad \text{ONI}_j = -\tanh(\text{MI}_j) \tag{6}$$

Values of ONI close to -1 indicate high similarity (potential memorisation), values of ONI close to 0 indicate similarity consistent with the null distribution (i.e., the typical, expected similarity between two unrelated samples from the training set) and values close to $+1$ indicate novelty.

3 Results

Experimental Setup: We evaluated our metric on three MRI datasets: BRATS [15] brain tumour images (axial),

Table 1. Robustness to augmentations: standard deviation across 8 augmentation types at each duplication level. Our metric is 6–20× more stable than CT Score.

Dup	BRATS		Spine		Knee	
	CT	Ours	CT	Ours	CT	Ours
5%	0.32	**0.02**	0.18	**0.02**	0.14	**0.02**
15%	0.87	**0.05**	0.47	**0.05**	0.39	**0.05**
30%	1.75	**0.09**	0.70	**0.10**	0.69	**0.09**
45%	2.73	**0.14**	0.85	**0.14**	0.92	**0.14**
Ratio	**18×**		**9×**		**7×**	

knee/thigh images [12] (sagittal), and spine images (sagittal) [10], each with 500 randomly selected slices divided into disjoint train and test sets. We introduced controlled duplication by replacing test samples with randomly selected training slices at 5%, 15%, 30%, and 45% duplication levels. To assess robustness, the duplicates were perturbed with eight augmentations: Gaussian noise ($\sigma \in \{0.01, 0.02\}$), rotations ($\pm 3, \pm 5$), horizontal/vertical flips, and intensity scaling ([0.9, 1.1]). We tested the ability of our metrics (MI and ONI) to detect this duplication and compared them against CT Score [6], FID, MMD, AuthPct [7], and Vendi Score [8].

Metric responses to duplication under augmentations: Fig. 2 shows the overall trends of each metric as the duplication level is increased for the different types of augmentations applied. In general, every metric responded in some way to increased duplication, with most metric exhibiting lower sensitivity as more noise or small rotations were applied.

Interpretation of observed trends: As the percentage of duplication increases, MI increases nearly linearly even in the presence of augmentations (Fig. 2a). CT score also increases, but is not always monotonic and exhibits a stronger spread when augmentations are applied (Fig. 2b). FID/MMD decrease with higher duplication (Fig. 2c–d), which could falsely imply better image fidelity. The Vendi score provides little/no usable signal (Fig. 2f), while AuthPct detects duplication well but is highly sensitive to the augmentation applied (Fig. 2e).

Robustness over different augmentations. To measure how consistent the metrics are to different image augmentations, we computed statistics for the same score over different augmentations. Table 1 quantifies the augmentation spread (stdev across the 8 augmentations at each duplication level). MI is 6–20× more stable than CT on all three datasets (e.g., at 45% leak on BRATS: 2.73 vs. 0.14; Spine: 0.85 vs. 0.14; Knee: 0.92 vs. 0.14), making it easier to interpret consistently.

Cross-dataset consistency: A metric whose scale shifts from dataset to dataset is hard to interpret. We performed the duplicate detection experiments in three different MRI datasets to test how consistent the metric values were. Table 2 reports the coefficient of variation (CV) across datasets at each duplication level. MI achieves 5.5× lower CV on average

Table 2. Cross-dataset consistency measured by coefficient of variation (CV). Our metric achieves 5.5× better consistency, enabling universal thresholds.

Duplication	CT Score	Our Metric	Improvement
5%	0.683	**0.163**	4.2×
15%	0.441	**0.069**	6.4×
30%	0.225	**0.035**	6.4×
45%	0.229	**0.019**	12.0×
Mean	0.395	**0.072**	**5.5**×

Table 3. ONI scores for clean samples: stable across all conditions (CV < 0.013), enabling universal threshold (ONI $<$ 0.68).

Dataset	Mean	Std	CV
BRATS	0.700	0.004	0.006
Knee	0.710	0.009	0.013
Spine	0.735	0.007	0.010
All	**0.715**	**0.017**	**0.023**

(0.072 vs. 0.395), with the largest gap at high levels of duplication (12× at 45%). This potentially enables universal thresholds for duplication to be set, e.g., a CT value of "-0.3" is dataset-dependent, whereas MI or ONI has a stable range of interpretation. Table 3 showcases clean ONI having stability across multiple datasets. This consistency / stability of MI/ONI across datasets justifies their usage as early-leak detectors across datasets.

Detection of duplicate samples (MI only): Unlike most other metrics, our metric provides *per-sample* scores and can be used identify (near) duplicate images. Detection performance was quantified via ROC-AUC and average precision (AP) using known duplication labels from synthetic contamination. Table 4 shows AUC by augmentation: detection is perfect (≈1.0) for clean, noise, and intensity; it remains good under geometric transforms (e.g., rot$\pm 3°$: mean 0.871; rot$\pm 5°$: 0.758; flips ≈0.73). This directly supports curation (flagging/removing a small set of memorized slices), which set-level metrics cannot address.

4 Discussion

We introduced a metric for detecting memorization that: (i) is monotonic and augmentation-stable at the set-level, (ii) avoids the misleading directionality of fidelity metrics under duplication, (iii) maintains cross-dataset scale consistency that permits universal thresholds, and (iv) provides per-sample scores to identify and remove leaked training images. Together, these properties turn memorization assessment from a passive diagnostic into a practical curation tool that complements fidelity metrics and improves the reliability of image generative models. Our experiments establish three

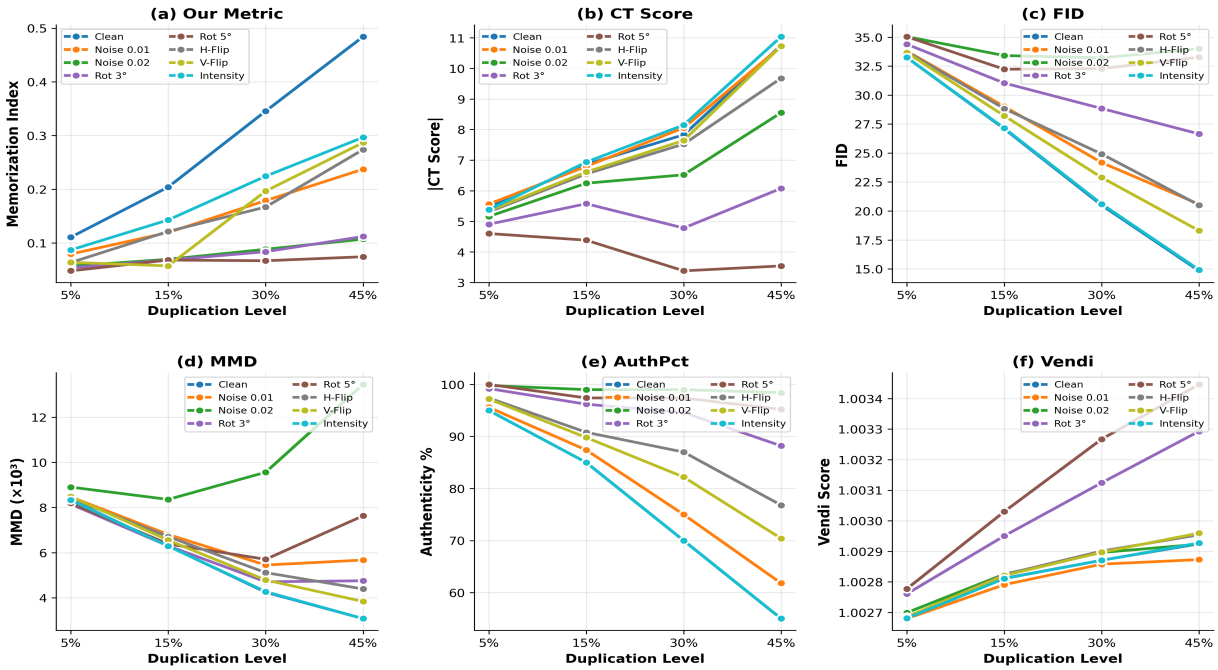

Fig. 2. **Metric response to duplication under augmentations.** (a) MI increases near-linearly and remains tight across augmentations. (b) CT spreads across augmentations, especially at higher duplication percentages. (c–d) FID/MMD decrease as duplication increases, which can be misleading if used as the only quality metric. (e) AuthPct is highly augmentation-sensitive. (f) Vendi shows little/no signal.

Table 4. Sample-level detection (AUC) by augmentation type. Perfect detection (1.0) for clean/noise/intensity; good detection (>0.72) for geometric transforms.

Augmentation	Mean AUC	Min AUC
Clean	1.000	1.000
Noise (0.01)	1.000	1.000
Noise (0.02)	1.000	0.999
Intensity Scale	1.000	1.000
Rotation (±3°)	0.871	0.805
Rotation (±5°)	0.758	0.711
H-Flip	0.733	0.626
V-Flip	0.727	0.635
Overall	**0.886**	**0.626**

practical properties of the proposed memorization detector. First, at the *set level*, our distance increases monotonically with duplication and remains tight across common augmentations, matching the trend-signal of CT while being substantially more stable (Table 1) and more consistent across datasets (Table 2). This matters because a metric whose scale drifts by dataset is hard to interpret; CT's starting points vary widely, whereas our score (MI and ONI) exhibit small cross-dataset CV, enabling universal thresholds.

Second, *fidelity* metrics (FID/MMD) decrease as duplication increases, which by construction suggests "better quality" even when the generator is copying. This confirms that fidelity should be paired with a memorization metric. Vendi shows negligible sensitivity and AuthPct is highly augmentation-dependent, limiting their diagnostic value.

Third, unlike set-level metrics, our method yields *per-sample* scores and can recover the actual near-duplicate pairs. Across datasets, ONI achieves near-perfect AUC for clean/noise/intensity cases and remains effective under small geometric changes (Table 4), making data curation *actionable* (flag/remove a tiny subset) rather than merely observational.

Limitations and failure modes. Sensitivity under small rotations and flips is lower than for appearance-preserving perturbations, especially at very low leak (5%). This is unsurprising given the local-invariance profile of the embeddings and the calibration's null. Two lightweight remedies are straightforward: (i) rotation/translation pooling of features before nearest-neighbor aggregation, and (ii) incorporating mild geometric jitter into the empirical null for MI→ONI calibration.

Acknowledgements: This work was supported by the Centre for Assuring Autonomy, a partnership between Lloyd's Register Foundation and the University of York. AFF acknowledges support from the Royal Academy of Engineering under the RAEng Chair in Emerging Technologies (INSILEX CiET1919/19), ERC Advanced Grant– UKRI Frontier Re-

search Guarantee (INSILICO EP/Y030494/1), the UK Centre of Excellence on in-silico Regulatory Science and Innovation (UK CEiRSI) (10139527), the National Institute for Health and Care Research (NIHR) Manchester Biomedical Research Centre (BRC) (NIHR203308), the BHF Manchester Centre of Research Excellence (RE/24/130017), and the CRUK RadNet Manchester (C1994/A28701).

Compliance with Ethical Standards: The experiments in this study were based on publicly available data collected with informed written consent and anonymised before distribution. The authors report no financial conflicts of interest.

5 References

[1] Bardia Khosravi, Frank Li, Theo Dapamede, Pouria Rouzrokh, Cooper U Gamble, Hari M Trivedi, Cody C Wyles, Andrew B Sellergren, Saptarshi Purkayastha, Bradley J Erickson, et al., "Synthetically enhanced: unveiling synthetic data's potential in medical imaging research," *EBioMedicine*, vol. 104, 2024.

[2] Chenghao Qian, Mahdi Rezaei, Saeed Anwar, Wenjing Li, Tanveer Hussain, Mohsen Azarmi, and Wei Wang, "Allweather-net: Unified image enhancement for autonomous driving under adverse weather and low-light conditions," in *International Conference on Pattern Recognition*. Springer, 2024, pp. 151–166.

[3] Muhammad Usman Akbar, Wuhao Wang, and Anders Eklund, "Beware of diffusion models for synthesizing medical images—a comparison with GANs in terms of memorizing brain MRI and chest x-ray images," *Mach Learn Sci Technol*, vol. 6, no. 1, pp. 015022, 2025.

[4] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter, "GANs trained by a two time-scale update rule converge to a local Nash equilibrium," *Adv Neural Inf Process Syst*, vol. 30, 2017.

[5] Yash Deo, Yan Jia, Toni Lassila, William AP Smith, Tom Lawton, Siyuan Kang, Alejandro F Frangi, and Ibrahim Habli, "Metrics that matter: Evaluating image quality metrics for medical image generation," *arXiv preprint arXiv:2505.07175*, 2025.

[6] Casey Meehan, Kamalika Chaudhuri, and Sanjoy Dasgupta, "A non-parametric test to detect data-copying in generative models," *arXiv preprint arXiv:2004.05675*, 2020.

[7] Ahmed Alaa, Boris Van Breugel, Evgeny S Saveliev, and Mihaela Van Der Schaar, "How faithful is your synthetic data? Sample-level metrics for evaluating and auditing generative models," in *International Conference on Machine Learning*. PMLR, 2022, pp. 290–306.

[8] Dan Friedman and Adji Bousso Dieng, "The Vendi Score: A diversity evaluation metric for machine learning," *arXiv preprint arXiv:2210.02410*, 2022.

[9] Haoyu Dong, Yuwen Chen, Hanxue Gu, Nicholas Konz, Yaqian Chen, Qihang Li, and Maciej A Mazurowski, "MRI-CORE: A foundation model for magnetic resonance imaging," *arXiv preprint arXiv:2506.12186*, 2025.

[10] L. Zhou, W. Wiggins, J. Zhang, et al., "The duke university cervical spine mri segmentation dataset (cspineseg)," *Scientific Data*, vol. 12, pp. 1695, 2025.

[11] Spyridon Bakas, Mauricio Reyes, Andras Jakab, et al., "Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge," 2019.

[12] Huahua Zhang et al., "Multimodal, multiethnic thigh-muscle MRI analysis," 2023, Accessed on 2025-11-12. URL: https://github.com/Hirrirririir/.

[13] Alexey Dosovitskiy, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[14] Agnan Kessy, Alex Lewin, and Korbinian Strimmer, "Optimal whitening and decorrelation," *The American Statistician*, vol. 72, no. 4, pp. 309–314, Jan. 2018.

[15] Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al., "The multimodal brain tumor image segmentation benchmark (BraTS)," *IEEE Trans Med Imaging*, vol. 34, no. 10, pp. 1993–2024, 2014.