

This is a repository copy of *Metrics that matter:Evaluating image quality metrics for medical image generation*.

White Rose Research Online URL for this paper:

<https://eprints.whiterose.ac.uk/226595/>

---

**Preprint:**

Deo, Yash, Jia, Yan, Lassila, Toni et al. (5 more authors) (2025) Metrics that matter:Evaluating image quality metrics for medical image generation. [Preprint]

<https://doi.org/10.48550/arXiv.2505.07175>

---

**Reuse**

This article is distributed under the terms of the Creative Commons Attribution (CC BY) licence. This licence allows you to distribute, remix, tweak, and build upon the work, even commercially, as long as you credit the authors for the original work. More information and the full terms of the licence here:

<https://creativecommons.org/licenses/>

**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](mailto:eprints@whiterose.ac.uk) including the URL of the record and the reason for the withdrawal request.

# Metrics that Matter: Evaluating Image Quality Metrics for Medical Image Generation

Yash Deo<sup>1,\*</sup>, Yan Jia<sup>1</sup>, Toni Lassila<sup>2</sup>, William A. P. Smith<sup>1</sup>, Tom Lawton<sup>1,8</sup>,  
Siyuan Kang<sup>3</sup>, Alejandro F. Frangi<sup>4,5,6,7</sup>, Ibrahim Habli<sup>1</sup>

<sup>1</sup>Department of Computer Science, University of York, York, UK

<sup>2</sup>School of Computer Science, University of Leeds, Leeds, UK

<sup>3</sup>Department of Computing and Mathematics,  
Manchester Metropolitan University, Manchester, UK

<sup>4</sup>Department of Computer Science, School of Engineering, University of Manchester,  
Manchester, UK

<sup>5</sup>School of Health Sciences, Division of Informatics, Imaging and Data Sciences,  
University of Manchester, Manchester, UK

<sup>6</sup>Department of Cardiovascular Sciences, KU Leuven, Leuven, Belgium

<sup>7</sup>Department of Electrical Engineering (ESAT), KU Leuven, Leuven, Belgium

<sup>8</sup>Improvement Academy, Bradford Institute for Health Research, Bradford Royal Infirmary,  
Bradford, UK

## Abstract

Evaluating generative models for synthetic medical imaging is crucial yet challenging, especially given the high standards of fidelity, anatomical accuracy, and safety required for clinical applications. Standard evaluation of generated images often relies on no-reference image quality metrics when ground truth images are unavailable, but their reliability in this complex domain is not well established. This study comprehensively assesses commonly used no-reference image quality metrics using brain MRI data, including tumour and vascular images, providing a representative exemplar for the field. We systematically evaluate metric sensitivity to a range of challenges, including noise, distribution shifts, and, critically, localised morphological alterations designed to mimic clinically relevant inaccuracies. We then compare these metric scores against model performance on a relevant downstream segmentation task, analysing results across both controlled image perturbations and outputs from different generative model architectures. Our findings reveal significant limitations: many widely-used no-reference image quality metrics correlate poorly with downstream task suitability and exhibit a profound insensitivity to localised anatomical details crucial for clinical validity.

Furthermore, these metrics can yield misleading scores regarding distribution shifts, e.g. data memorisation. This reveals the risk of misjudging model readiness, potentially leading to the deployment of flawed tools that could compromise patient safety. We conclude that ensuring generative models are truly fit for clinical purpose requires a multifaceted validation framework, integrating performance on relevant downstream tasks with the cautious interpretation of carefully selected no-reference image quality metrics.

## 1 Introduction

The development of advanced medical imaging applications, particularly those leveraging artificial intelligence (AI), often requires large, diverse, and accurately labelled datasets. However, acquiring such datasets faces significant hurdles due to patient privacy regulations and the high cost and complexity associated with expert anatomical annotation. These limitations can constrain the training and validation of potentially beneficial AI tools. Generative deep learning models offer a potential solution to these challenges. Techniques such as generative adversarial networks (GANs, see [72] for a review), variational au-

toencoders (VAEs), and denoising diffusion probabilistic models (DDPMs, see [41] for a review) have demonstrated the capacity to synthesise (or generate, terms used synonymously herein) high-fidelity medical images. These models encompass various approaches, including conditional generation (e.g., image translation [50]) and unconditional synthesis, with generated images being used for tasks including data augmentation ([42]), privacy preservation ([52]), cross-modality synthesis ([24]), image super-resolution ([3]), and virtual imaging trials ([1]). Regardless of the specific generation strategy, robust evaluation of the output’s quality and reliability is essential, particularly distinguishing the nature of these synthetic outputs from direct empirical measurements [62].

While promising, the integration of generative models and the synthetic data they produce into clinical research and potentially downstream applications necessitates rigorous validation to ensure they are trustworthy and safe. Unlike directly measured medical images, outputs from generative models function more like complex computational simulations or predictions derived from learned distributions [69, 63]. Evaluating these typically requires *no-reference image quality metrics* (NRIQMs), as a direct reference often doesn’t exist. These metrics must assess not only image *fidelity* (resemblance to real images) but also *diversity* (adequate coverage of the true data distribution) [5], moving beyond simple similarity to evaluate the model’s representation of reality [47]. Generative models can suffer from failure modes such as *mode collapse* (overproducing similar images from certain modes with limited diversity across the data distribution) or *mode invention* (creating spurious, ‘hallucinated’ images with features not present in the real data). The deployment of clinical decision support tools trained or validated using synthetic data from poorly evaluated generative models poses significant risks to patient safety. For instance, mode invention could lead to downstream tools learning from hallucinated pathological features, while mode collapse might yield tools biased against rare conditions; without the right risk controls or guardrails, reliance on such compromised tools could subsequently lead to inaccurate diagnoses or inappropriate treatment selections. Therefore, the thorough evaluation of the image generation process using appropriate and validated metrics, considering their ‘adequacy-for-purpose’ [9, 64], is not merely a technical requirement but a crucial step for ensuring the safety and reliability of clinical applications ultimately impacted by synthetic data.

In this work, we perform a systematic assessment of commonly used NRIQMs intended for evaluating

generative medical imaging outputs. We subject these NRIQMs to rigorous experiments designed to stress-test both their sensitivity to common image artifacts (e.g., noise) and their ability to detect known generative model failure modes (e.g., mode collapse, mode invention, data memorisation, morphological inaccuracies). We compare results from *upstream* evaluation, where quantitative NRIQMs are applied directly to the synthetic images to assess their intrinsic properties and distributional characteristics, against performance in *downstream* evaluation. For the purpose of this study, downstream evaluation specifically involves assessing the utility of the generated images by inputting them into a relevant, pre-trained task model – in this case, a segmentation model – and measuring the accuracy or quality of that model’s output when processing the synthetic input. This comparative approach allows us to investigate which upstream metrics robustly reflect this defined task-specific utility or detect critical image flaws, and which provide potentially misleading evaluations of model performance. We focus our validation on brain magnetic resonance imaging (MRI), utilising brain tumour images from the BRaTS dataset [7, 55, 6] and vascular angiography images from the IXI dataset [35]. Although our study uses brain MRI, the methodology and findings regarding metric behaviours are potentially extensible to other medical imaging modalities. Our implementation is available for adaptation/use at the following GitHub repository: <https://github.com/YashDeo-York/GenMed>.

Our primary contribution is a critical assessment of NRIQMs for generative models in medical imaging. Specifically, we provide concrete evidence showing that popular distance metrics (e.g. FID, KID, MMD) can yield misleadingly optimistic scores under conditions such as data memorisation or mode collapse. Critically, our work highlights a profound limitation across nearly all tested upstream metrics: their insensitivity to localised, clinically relevant morphological alterations, such as distorted tumour boundaries, questioning their ability to ensure anatomical accuracy. Furthermore, we demonstrate a significant and recurring discrepancy where upstream metric rankings of generative models (VAE, GAN, DDPM) do not align with their performance on downstream segmentation tasks. Our analysis therefore underscores the urgent need for evaluation metrics specifically validated for medical imaging and provides strong evidence supporting the necessity of multifaceted evaluation frameworks that prioritise assessments of ‘adequacy-for-purpose’ [9, 64] to ensure generative models, and the applications built upon them, meet the stringent safety and reliability standards required in clinical contexts.

## 2 Background and Related Work

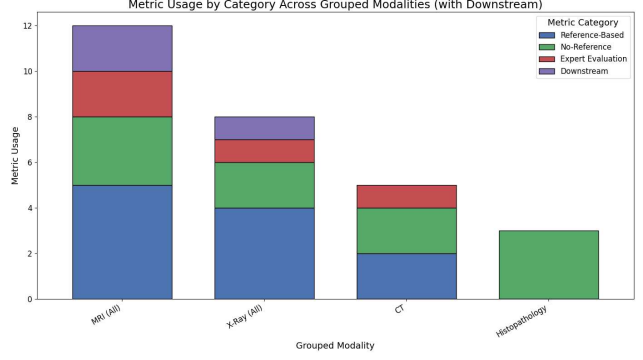
The evaluation of image-generative models requires a comprehensive approach that balances structural fidelity, perceptual realism, and task-specific performance. Generating medical images further demands metrics capable of capturing nuanced clinical and anatomical features.

Traditionally, for image quality evaluation, *full-reference* (pixel-level) metrics, such as SSIM (structural similarity index measure) and PSNR (peak signal-to-noise ratio), are widely adopted for their simplicity and ability to quantify fine-grained reconstruction accuracy. However, these metrics require a reference image to compare against, which is challenging when no ground-truth image exists. Furthermore, full-reference metrics often fail to adequately account for perceptual quality or alignment with real-world data distributions.

To address these limitations, recent research has shifted towards *NRIQMs* that evaluate image quality without requiring pixel-wise alignment to a ground truth. Among the most widely used are the Fréchet Inception Distance (FID) [33] and the Inception Score (IS) [70]. FID compares the statistical distributions of real and generated image features extracted from a pre-trained network (typically InceptionV3), providing a global measure of distributional similarity. IS, on the other hand, assesses both the quality and diversity of generated images by computing the entropy of predicted label distributions from an Inception classifier.

More recent alternatives, such as Kernel Inception Distance (KID) [8], use polynomial kernel-based Maximum Mean Discrepancy (MMD) and provide unbiased estimates that are better suited for small datasets. In parallel, several domain-agnostic metrics have emerged to capture specific generative failure modes. The **AuthPct** [5] (Authenticity Percentage) measures the proportion of generated samples falling outside the support of the real distribution, helping to identify out-of-distribution generations and address issues related to mode invention. The **CT Score** [54] (Cluster Tightness Score) quantifies the compactness of feature clusters in the generated data, potentially revealing mode collapse or insufficient diversity generations. **FD<sub>∞</sub>** [12], an extension of Fréchet distance, captures worst-case divergence and highlights rare but significant deviations in feature space.

Despite the recent surge in new metrics proposed for evaluating generative model quality and distributional characteristics, including those designed to address mode collapse and mode invention, a significant gap exists in their validation on medical imaging data.



**Figure 1. A graph of different types of metrics used in generative medical image evaluation, categorised by modality.**

Many metrics developed on natural image datasets do not adequately address the unique complexities of medical images, such as subtle pathological variations, noise sensitivity, and domain shifts arising from diverse scanner configurations. While specialised metrics tailored to medical imaging have been proposed ([74]), their adoption remains limited.

To explore current evaluation practices in the field, we conducted a literature review of papers employing generative models in medical imaging. Our search spanned publications from 2019 to 2024 across major databases including IEEE Xplore, PubMed, arXiv, and Google Scholar, using keywords operation {"medical image synthesis," or "medical imaging,"} and {"GAN," or "diffusion models" or "VAE" or "Variational auto-encoder"} and {"reference-metrics" or "no-reference metrics" or "image quality assessment"}. After filtering for relevance, We identified 28 papers. Through this review, we identify key trends in the usage of metrics for medical image generation in research, which we present in Fig 1.

Ideally, expert evaluation by clinicians should serve as the gold standard for assessing the anatomical plausibility and clinical utility of synthetic medical images. However, this approach faces significant practical challenges, including considerable time and cost requirements, inherent inter-observer variability where different experts may provide differing assessments, difficulties in scaling such evaluations to large datasets or numerous model iterations, and the need for carefully standardised rating protocols to ensure consistency. Despite these limitations, expert evaluation remains valuable and has been employed in specific contexts. For instance, [43] trained diffusion models on CT scans of different organs and relied on radiologists to rank im-

ages by realism, inter-slice consistency, and anatomical correctness. Similarly, [68] and [31] applied GAN-based approaches to mammograms and brain MRI respectively, and asked clinicians to judge whether each image was real or fake. Other works have tried to establish correlations between expert evaluations and image quality metrics ([14]; [53])

In practice, as shown in Fig 1, the most commonly used metrics for evaluating synthetic medical images are full-reference metrics, particularly PSNR and various forms of SSIM, including MS-SSIM (Multi-scale SSIM) [76] and 4GR-SSIM [49]). These metrics are especially common in settings where generative models are applied to tasks like image super-resolution [79, 77], slice synthesis or interpolation [60, 17], and image reconstruction [65, 37, 73, 15, 29], where partial or proxy ground-truth images are available. Although these applications are not strictly paired in a supervised sense, they offer 'semi-ground-truth' data that makes using full-reference metrics for comparisons feasible.

Interestingly, even in tasks involving unconditional generation — where no explicit reference image exists — researchers continue to use full-reference metrics. A common workaround involves identifying the closest real image to each generated image using nearest-neighbour search in pixel or feature space, followed by averaging metric scores over the dataset [19, 20, 11]. Other studies use MS-SSIM or 4GR-SSIM not to evaluate fidelity, but as a proxy for diversity [22, 37, 67]. The underlying assumption is that MS-SSIM values close to 1 suggest excessive similarity (potential mode collapse), while values too close to 0 indicate low-quality or noisy generations. While these techniques offer creative adaptations of full-reference metrics, they remain heuristic and lack rigorous theoretical justification.

Because full-reference metrics are inherently limited in their ability to capture perceptual quality and global distributional alignment between real and generated images, there has been a growing shift towards NRIQMs such as FID and IS [67, 38, 34], which is the second most used metric category as shown in Fig 1. These metrics evaluate image quality based on high-dimensional features extracted from deep networks, offering a more holistic assessment of realism and diversity in unpaired or distribution-based generative tasks. However, the reliability of most NRQIMs is heavily influenced by the choice of feature extractor. Most implementations rely on features derived from networks pre-trained on natural image datasets, typically using InceptionV3. Due to concerns about the suitability of such networks for medical imaging — where visual patterns differ significantly from natural images — alternatives such as Med3D [10] have

been proposed. Med3D, pre-trained on large-scale medical imaging datasets, has recently been adopted in place of Inception-based extractors in several studies [19, 67]. Conversely, subsequent research [78] has indicated that feature extractors specifically trained on medical datasets, including Med3D, may produce evaluation scores that are unstable or inconsistent when assessing generative outputs. Contrarily, the InceptionV3 network, despite its lack of domain-specific training, has demonstrated more stable results. Recently, academic focus has been directed towards self-supervised architectures such as DINOv2 [74], which possess the capability to learn broad visual representations. While these models have exhibited potential in natural image benchmarks, their efficacy and consistency within the domain of medical imaging remain substantially unvalidated.

Taken together, these contradictory findings highlight a broader uncertainty in the field: there is currently no standardised or well-validated framework for evaluating generative medical images using NRIQMs. The choice of feature extractor has a substantial impact on metric behaviour, yet the absence of consensus and the limited validation of these methods in clinical contexts continue to hinder the development of consistent and interpretable evaluation pipelines.

Alongside these developments, there is a growing trend towards using downstream tasks as a means of evaluating the quality and clinical utility of synthetic images, accounting for a proportion similar to that of expert evaluation in our literature review, as illustrated in Fig 1. This approach is particularly valuable when generative models are developed for task-specific purposes, such as simulating anatomical variations or pathological features. For example, in the context of brain vasculature generation, [19] proposed training a classifier to predict the anatomical type of generated vessels as a proxy for realism. Similarly, [20] suggested that performing downstream flow simulations on generated aneurysm-containing vessels could serve as a strong indicator of structural validity. In tumour imaging, [26] evaluated the realism of synthetic brain MRI images by passing them through a pre-trained tumour segmentation network, while [25] assessed segmentation accuracy by comparing predicted label maps from generated images against reference maps.

While downstream evaluation offers a promising framework for assessing practical utility, it does not provide a complete picture of generative performance. Specifically, downstream tasks are often insensitive to properties such as sample diversity, distributional coverage, and visual fidelity — dimensions that are critical in understanding model behaviour. Furthermore, the

absence of ground truth for generated images introduces uncertainty when interpreting downstream performance. Although recent work has proposed ground-truth-free metrics to assess segmentation accuracy on generated images [71], these approaches are still in early stages and have not yet seen widespread adoption.

Therefore, there is a clear need to systematically investigate the sensitivity and reliability of different evaluation metrics for medical image generation — both from upstream and downstream perspectives. While some prior work has attempted to assess the performance of image quality metrics in medical imaging, the scope has been limited. For instance, [21] provided a review of image quality metrics for image-to-image translation tasks but did not include NRIQMs. Other studies [13, 40, 48] have examined medical image quality metrics more broadly, yet did not focus on generative image evaluation.

In this work, we adopt a structured and objective approach to evaluate NRIQMs in a clinically relevant setting. We first conduct a series of controlled perturbation experiments — including noise injection, morphological manipulations, and domain shifts — on real data to assess how different metrics respond to well-characterised variations. This helps to quantify each metric’s sensitivity to common failure modes in generative image analysis. Subsequently, we evaluate the outputs of multiple generative models (VAE, GAN, DDPM) using the same metrics, and further assess their utility through downstream clinical tasks. This allows us to examine how well metric responses align with real-world use cases. Together, this framework offers a more nuanced understanding of the strengths and limitations of NRIQMs in the context of medical image generation.

## 2.1 Metrics used for Upstream Evaluation

This section summarises NRIQMs used in this study to assess synthetic medical images where a ground truth reference image is typically unavailable. We divide the selected metrics into three complementary categories based on the primary aspect of quality they aim to capture: *distribution alignment metrics*, *perceptual realism metrics*, and *statistical fidelity metrics*. A concise technical summary, including formulas and references, is provided in Table 1.

**(A) Distribution Alignment Metrics** These metrics quantify how well the distribution of generated image features aligns with that of real data features, primarily evaluating whether a model captures the overall

diversity and statistical moments of the true distribution.

- **FID (Fréchet Inception Distance):** Computes the Fréchet distance between Gaussian distributions fitted to feature embeddings (typically from InceptionV3) of real versus generated images, comparing their means and covariances. Lower FID values indicate closer distribution alignment. While sensitive to various artifacts, the choice of feature space can impact results; consistent use of standard InceptionV3 features is common for comparability ([33]). We follow recent findings suggesting the reliability of Inception features even for medical images [78].
- **KID (Kernel Inception Distance):** Similar to FID but uses polynomial kernel comparisons (specifically, the squared Maximum Mean Discrepancy, MMD) between feature distributions, offering potentially greater stability for smaller sample sizes often encountered in medical datasets ([8]).
- **sFID (Spatial FID):** Extends standard FID by incorporating spatial information from feature maps, aiming to provide sensitivity to structural misalignments (e.g., organ shape, lesion location) which are critical in medical imaging ([58]).
- **PRDC (Precision, Recall, Density, Coverage):** A suite of metrics evaluating distribution similarity based on feature manifolds. *Precision* measures the fraction of generated samples falling within the real data manifold (fidelity), while *Recall* measures the fraction of the real data manifold covered by generated samples (diversity). High recall is crucial for capturing rare conditions, while high precision helps avoid unrealistic artifacts ([57, 46]). Density and Coverage provide further insights into relative densities and manifold coverage.
- **Vendi Score:** Measures diversity based on the similarity between samples within a set, calculated using a kernel function over feature vectors. Higher scores indicate greater diversity. It aims to capture the spread of the distribution ([27]).

**(B) Perceptual Realism Metrics** These metrics focus on qualities related to human perception, structural plausibility, or distinguishability from real images, often assessing individual samples or local features.

- **LPIPS (Learned Perceptual Image Patch Similarity):** Primarily a reference-based metric

correlating well with human perception by evaluating patch-wise similarity between two images in deep feature space [81]. However, it can be adapted for evaluating generative models in a no-reference manner by analysing aggregate perceptual distances. Common adaptations involve calculating the average pairwise LPIPS distance among generated samples (where lower average distances may indicate lower diversity or mode collapse) or computing the average LPIPS distance between each generated image and its closest neighbour in the real dataset (assessing perceptual fidelity to the real data manifold).

- **Inception Score (IS):** Assesses image quality (sharpness/distinctiveness) and diversity using predicted class probabilities from a pre-trained classifier (Inception). Higher scores suggest high-quality, diverse images. Its applicability to medical imaging can be limited by the need for relevant, well-defined class labels ([70]).
- **Realism Score:** Often employs a discriminator or classifier trained to distinguish real from generated images. High scores (or low discriminator accuracy) suggest generated samples are visually and feature-wise difficult to distinguish from real images ([45]).

### (C) Statistical Fidelity & Manifold Metrics

These metrics target statistical properties of the generated distribution, often focusing on how well generated samples adhere to the underlying manifold of real data in feature space, complementing global distribution alignment measures.

- **FLS (Feature Likelihood Score):** Quantifies the likelihood of generated features under a distribution model (often Gaussian) fitted to real data features. Higher FLS values suggest synthetic images are statistically typical of the real data manifold ([39]).
- **ASW (Approximate Sliced Wasserstein):** Approximates the Wasserstein distance between real and generated image feature distributions using random projections (slices). Lower values indicate closer distributions. It can be interpreted in terms of feature space clustering ([56]).
- **Ct Score (Clusterability Score):** Evaluates distribution overlap by assessing how well generated image features conform to clusters established by real image features in feature space. A lower

Ct Score indicates better adherence to real data clusters ([54]).

- **AuthPct (Authenticity Percentage):** Calculates the fraction of generated samples falling within a region defined by the real data distribution in feature space, based on a learned threshold. It measures the proportion of synthetic images deemed statistically plausible or *authentic* ([5]).
- **$FD_\infty$  (Infinite Fréchet Distance):** Measures the maximum divergence between feature distributions along any projection, capturing worst-case discrepancies rather than average differences, as in FID. Potentially sensitive to outliers or rare failure modes ([12]).

Table 1 consolidates the key formulas and parameter definitions for each metric mentioned above.

## 2.2 Downstream Evaluation Technique

To assess the practical utility and clinical relevance of the generated images beyond upstream quantitative metrics, we evaluated their suitability for a critical downstream task: automated segmentation. We employed relevant, pre-trained segmentation algorithms appropriate for each dataset: brain vessel segmentation for the synthetic IXI MRA images [18] and brain tumour segmentation for the synthetic BraTS MRI images [51]. These algorithms were applied directly to the synthetic images obtained from the pre-trained VAE, GAN, and DDPM models.

### 2.2.1 Addressing the Reference Image Challenge in Downstream Evaluation

A fundamental challenge in evaluating downstream tasks on purely synthetic images generated from noise or latent variables is the absence of corresponding ground truth annotations (e.g., segmentation masks) for comparison. Direct pixel-level metrics such as Dice or IoU cannot be computed without a reference. While qualitative visual assessment of the segmentation outputs on synthetic images is possible, it introduces subjectivity. To overcome this reference-free challenge and enable quantitative assessment, we adopt the regression-based segmentation quality evaluation methodology proposed by Senbi et al. [71].

This involves training an auxiliary regression model—the quality prediction model—to estimate segmentation quality scores (specifically, Dice scores in our case) based on the input image and a generated segmentation mask, without requiring the ground truth

**Table 1. Summary of Metrics Used in the Evaluation**

Name	Description	Formula / Computation Principles
FID	Fréchet Inception Distance: Evaluates similarity between real and generated feature distributions using mean and covariance of features (typically InceptionV3). Lower is better.	$\text{FID} = \ \mu_r - \mu_g\ _2^2 + \text{Tr}(\Sigma_r + \Sigma_g - 2(\Sigma_r \Sigma_g)^{1/2})$
Spatial FID (sFID)	Spatially-aware FID extension incorporating structural context by using features from intermediate convolutional layers. Lower is better.	Computed similarly to FID, but using spatially-preserved features.
Inception Score (IS)	Assesses quality (low conditional entropy) and diversity (high marginal entropy) using an Inception classifier’s predictions on generated images. Higher is better.	$\text{IS} = \exp(\mathbb{E}_x D_{KL}(p(y x)  p(y)))$
KID	Kernel Inception Distance: Compares feature distributions using the squared Maximum Mean Discrepancy (MMD) with a polynomial kernel. Lower is better.	$\text{KID} = \text{MMD}^2(\mathcal{X}_r, \mathcal{X}_g)$ computed on feature sets $\mathcal{X}_r, \mathcal{X}_g$ .
LPIPS	Learned Perceptual Image Patch Similarity: Reference-based metric comparing patch-wise deep features; lower distance implies higher perceptual similarity.	$\text{LPIPS}(x_0, x_1) = \sum_l w_l \ \Phi_l(x_0) - \Phi_l(x_1)\ _2^2$ over layers $l$ .
ASW	Approximate Sliced Wasserstein Distance: Measures Wasserstein-1 distance between feature distributions using random 1D projections ( $\theta$ ). Lower is better.	$\text{ASW} = \mathbb{E}_{\theta \sim S^{d-1}} W_1(P_\theta(\mathcal{X}_r), P_\theta(\mathcal{X}_g))$
FLS	Feature Likelihood Score: Quantifies how statistically typical generated features are under a model (e.g., Gaussian) fitted to real features. Higher is better.	Typically involves fitting $p_{\text{real}}(f)$ to real features $f_r$ and evaluating $\mathbb{E}_{f_g \sim P_g} [\log p_{\text{real}}(f_g)]$ . (Note: Original formula matched FID, likely incorrect for likelihood).
DreamSIM	Learned perceptual similarity metric trained to align with human judgments, comparing deep features. Higher is better (assuming similarity).	$\text{DreamSIM}(x_0, x_1) = f(\Phi(x_0), \Phi(x_1))$ , where $f$ is a learned similarity function.
$\text{FD}_\infty$	Infinite Fréchet Distance (often related to MMD): Measures worst-case discrepancy between distributions via maximizing difference over functions in a unit ball. Lower is better.	$\text{FD}_\infty \approx \sup_{f \in \mathcal{F}}  \mathbb{E}_{x \sim P_r}[f(x)] - \mathbb{E}_{x \sim P_g}[f(x)] $ for function class $\mathcal{F}$ .
Ct Score	Clusterability Score: Evaluates how well generated features fit within clusters defined by real features. Lower score implies better fit/less dispersion.	Conceptually measures properties like variance or distance within/between clusters formed by real vs. generated features. (Note: Original formula seemed incomplete/unclear).
AuthPct	Authenticity Percentage: Fraction of generated samples considered ‘authentic’ or statistically plausible relative to the real data distribution. Lower indicates more divergence.	$\text{AuthPct} = \frac{1}{N_g} \sum_{i=1}^{N_g} \mathbb{1}_{\{x_{g,i} \in \mathcal{D}_{\text{real}}\}}$ . (Note: $\mathcal{D}_{\text{real}}$ typically defined via threshold $\tau$ on density $p(x)$ or distance $d(x, \text{real data})$ ).
PRDC	Precision, Recall, Density, Coverage for Distributions: Evaluates sample quality (Precision) and distribution coverage (Recall) based on nearest-neighbour analysis in feature space. Higher is better.	$\text{Precision} \approx$ Fraction of generated samples whose NN is real. $\text{Recall} \approx$ Fraction of real samples whose NN is generated. (Simplified description).
Vendi Score	Measures diversity within a set of samples based on feature similarity, calculated using a kernel matrix over features. Higher score means higher diversity.	Computation involves eigenvalue decomposition of the kernel matrix $K_{ij} = k(\Phi(x_i), \Phi(x_j))$ .
Realism Score	Measures the probability that a classifier/discriminator judges a generated image as real. Higher is better.	$\text{Realism}(x_g) = P(\text{real}   x_g)$ using a trained discriminator/classifier.



mask during the final evaluation phase. Our implementation involved the following key steps:

1. **Target Segmentation Generation:** First, we obtained the segmentations whose quality needed to be assessed. This involved applying standard, pre-trained segmentation models (specific to each dataset, see [51, 18]) to the input images (either the perturbed real images from controlled experiments or the synthetic images from generative models).
2. **Quality Predictor Training Data:** Second, we created a diverse dataset to train the quality prediction model itself. For each real image in our training set, we first generated an initial segmentation mask using the standard pre-trained segmentation model. We then created multiple synthetically perturbed versions of this mask using various augmentations (e.g., adding noise, simulating small erosions/dilations). Crucially, since we possessed the ground truth masks for these real images, we calculated the accurate Dice score between each original or perturbed mask and the corresponding ground truth mask. The training dataset for the quality predictor thus consisted of pairs of real image, potentially perturbed segmentation mask, with its precisely calculated Dice score serving as the target regression label. This process generated multiple training samples per original image, covering a spectrum of segmentation qualities. (Note: Alternative training data strategies mentioned in the literature, such as using only pairs of real image, real ground truth mask assuming a perfect score, were not the primary method employed here).
3. **Quality Predictor Training:** Third, we trained the regression model (A ResNet classifier) to predict the Dice score based on the input image and its corresponding segmentation mask from the training data created in Step 2.
4. **Downstream Quality Assessment Step:** Finally, this trained quality prediction model (from Step 3) was used in both experimental phases to assess downstream quality. It was applied to pairs of input image (perturbed real or synthetic), corresponding segmentation mask (generated in Step 1) to provide the predicted Dice score (referred to as the Evanyseg score in our results), estimating the quality of the segmentation achievable on each input image without direct access to ground truth.

We acknowledge that evaluating segmentation quality using such a predictive model introduces its own

layer of potential uncertainty – the evaluation relies on the accuracy of the trained quality predictor. However, in the absence of ground truth for generated images, this method provides a standardised, quantitative, and reproducible proxy for downstream task performance. It allows for a relative comparison of how well the different generative models produce images suitable for the pre-trained segmentation task, moving beyond purely subjective visual inspection or solely relying on upstream NRIQMs. While not a perfect substitute for ground-truth validation, it serves as a necessary tool for assessing downstream task suitability in this challenging reference-free context.

### 3 Methodology

To systematically evaluate the suitability of NRIQMs for generative medical imaging, our methodology employs a unified evaluation framework integrating both upstream quantitative metric assessment and downstream task validation across two distinct phases. Phase one involves *controlled perturbation experiments evaluation* where known alterations are introduced to real medical images (MRI and MRA) to simulate specific artifacts or distributional issues. Phase two focuses on *generative model performance evaluation*, assessing synthetic images produced by representative generative architectures (VAE, GAN, DDPM). Crucially, within **both** phases, we apply the full suite of selected upstream NRIQMs directly to the images (perturbed real images in Phase 1 and synthetic images in Phase 2) *and* perform downstream evaluations by assessing the performance of a pre-trained segmentation model when applied to these same images. This comprehensive approach allows us to analyse metric sensitivity under controlled conditions, assess performance on actual generative model outputs, and critically compare whether upstream metric scores align with downstream task utility across diverse scenarios. By integrating insights from both phases, we aim to provide robust evidence regarding the strengths and limitations of common NRIQMs for evaluating generative models in clinically relevant contexts. Fig. 2 showcases a high-level overview of this integrated methodology.

#### 3.1 Phase 1: Controlled Perturbation Experiments Evaluation

To systematically evaluate NRIQM robustness under specific conditions, this phase involves applying controlled perturbations to real medical images. We utilise subsets of real images from the BRaTS MRI (tumour-focused) and IXI MRA (vessel-focused)

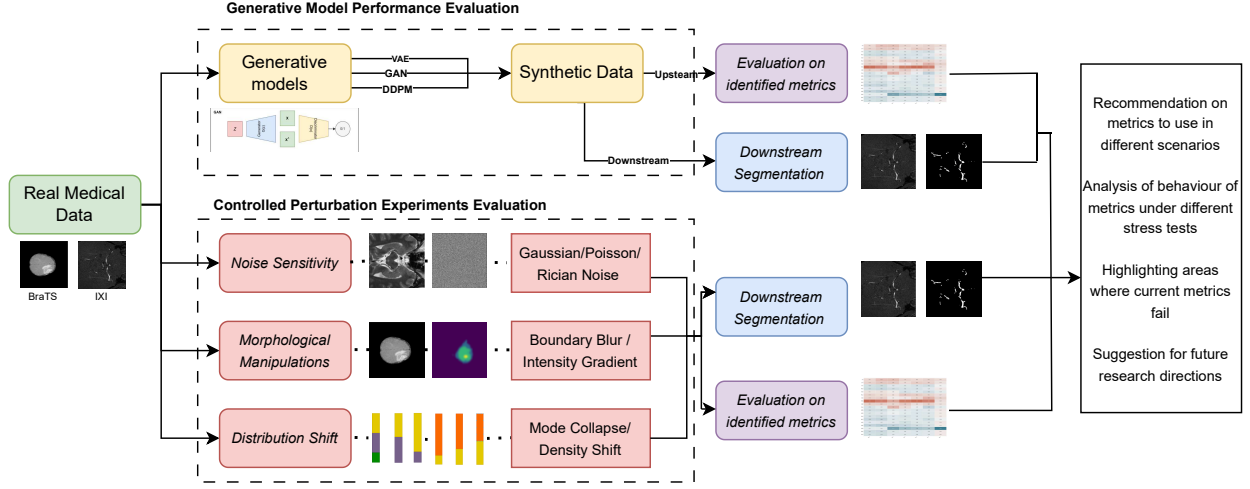


Figure 2. Overview of our methodology.

datasets as our *base image sets*. We then create multiple distinct *perturbed image sets* by applying controlled manipulations to copies of these base images. These manipulations, detailed in the following sections, simulate specific challenges including noise injection, morphological alterations, and various forms of distribution shift (e.g., data memorisation, mode changes). For metrics requiring comparison to a reference distribution (such as FID, KID, PRDC, etc.), we utilise a separate, unperturbed set of real images from the corresponding dataset, designated as the *reference distribution set*. This experimental design allows us to isolate and analyse the response of both upstream NRIQMs and downstream task performance (using the pre-trained segmentation model) to specific, well-defined image perturbations across different anatomical structures and imaging characteristics.

## Noise Sensitivity

Noise is an intrinsic characteristic of medical imaging, arising from scanner hardware limitations, acquisition protocols, and patient movement (in the form of motion artefacts). Generative models must learn to either faithfully replicate these noise patterns or appropriately compensate for them. However, different architectures might introduce distinct noise signatures, and evaluation metrics may respond inconsistently to scanner-induced versus model-generated noise. We therefore conduct structured noise-injection experiments using three clinically relevant noise models (Gaussian, Poisson, Rician) applied to our base image sets, allowing us to assess both upstream metric re-

sponses and downstream task performance under these conditions.

**Gaussian Noise Injection** Probabilistic generative models, particularly VAEs and some diffusion models, frequently introduce Gaussian-like noise artifacts or smoothing in their reconstructions due to their underlying mathematical formulation [44]. Since VAEs often optimise a likelihood function based on a Gaussian prior, their outputs can exhibit characteristic smoothness with subtle noise textures. Similarly, diffusion models may retain residual Gaussian-like noise depending on their noise schedule and de-noising process. This experiment assesses how effectively evaluation metrics and downstream tasks capture the impact of these specific noise patterns. **Experimental Design:** We introduce zero-mean Gaussian noise with progressively increasing standard deviations ( $\sigma = \{0.01, 0.05, 0.1\}$ ) to copies of the real MRA base images. These noise levels are calibrated to simulate gradual degradation. We then compute the suite of upstream NRIQMs and evaluate downstream segmentation performance on perturbed image sets. We assess the degree to which metrics and task performance change relative to the noise level, identifying potential insensitivities to this type of degradation.

**Rician and Poisson Noise Injection** Unlike Gaussian noise, magnitude MRI acquisitions typically contain Rician-distributed noise resulting from the reconstruction process of complex-valued k-space data [28]. Poisson noise represents an intensity-dependent noise process relevant to certain low-count imaging scenarios.

Generative models often struggle to accurately reproduce these specific noise distributions, sometimes overdenoising images (removing clinically relevant texture) or introducing non-physical artifacts. This experiment evaluates whether metrics and downstream tasks are sensitive to degradations reflecting more realistic noise characteristics encountered in MRI. **Experimental Design:** We apply calibrated levels of Rician and Poisson noise to copies of the MRA base images. For Rician noise, we implement the standard model:

$$I_{\text{Rician}} = \sqrt{(I_{\text{orig}} + n_1)^2 + n_2^2}$$

where  $n_1, n_2 \sim \mathcal{N}(0, \sigma^2)$  represent independent Gaussian noise components. For Poisson noise, intensity variations are scaled appropriately for the image data. We systematically vary the noise level (via  $\sigma$  or equivalent SNR) to create a spectrum of conditions. For each noise type and level, we compute the full suite of upstream NRIQMs and assess downstream segmentation performance on perturbed image sets.

**Metric and Task Response Analysis** For each noise type, we analyse the response curves of the upstream metrics (e.g., via z-score of percentage change from baseline) and the corresponding changes in downstream segmentation performance scores. We quantify the sensitivity of both metrics and the downstream task to each noise type, observing the rate of change per unit increase in noise level. Comparing sensitivities across noise types (e.g., high sensitivity to Gaussian vs. Rician for a given metric) provides insights into potential metric biases. A metric that demonstrates high sensitivity to Gaussian noise but remains relatively insensitive to Rician noise likely reflects optimisation for natural image domains rather than medical-specific degradation patterns. Conversely, metrics that appropriately penalise Rician and Poisson noise — which better represent real MRI acquisition physics — may be more suitable for evaluating medical image generation models. Critically, by comparing the magnitude of upstream metric responses to the degradation observed in downstream task performance, we can assess how well different metrics reflect the practical impact of noise on image utility for downstream tasks. These controlled noise-injection experiments establish quantitative baselines for understanding metric and task behaviour under well-characterised perturbations.

## Morphological Manipulations

Morphological alterations in medical images, particularly within pathological regions like tumours, are of

ten critical indicators for diagnosis, staging, and treatment planning. Generative models aiming to synthesise realistic medical images must accurately represent and preserve these structural details. Therefore, evaluation metrics should ideally be sensitive to even subtle variations in relevant anatomical or pathological morphology. We conduct controlled morphological alteration experiments, focusing specifically on tumour regions within the BRaTS MRI base images, using two distinct approaches: tumour boundary blurring and intensity-based tumour gradient variations. We assess the impact of these manipulations using both upstream NRIQMs and downstream segmentation performance.

**Tumour Boundary Blurring** Accurate delineation of tumour boundaries is crucial for precise surgical planning and radiotherapy target definition. Generative models that produce overly smooth or indistinct boundaries can lead to clinically significant errors in downstream applications such as volume estimation or treatment planning [16], compromising patient safety. This experiment simulates such scenarios and evaluates whether metrics or downstream tasks can detect these boundary inaccuracies. **Experimental Design:** Using the ground truth segmentation masks available in the BRaTS dataset, we identify the tumour regions in the base MRI images. To selectively blur the boundaries, we first compute the Euclidean distance transform within the tumour mask to identify boundary regions based on proximity to the edge. A standard Gaussian blurring filter was then applied specifically to the intensity values within these identified boundary regions on each slice. The strength of the blurring effect was controlled by systematically adjusting the standard deviation of the Gaussian kernel across several distinct levels, simulating varying degrees of boundary sharpness produced by generative models. Subsequently, we compute the upstream NRIQMs and evaluate downstream tumour segmentation performance on these boundary-manipulated images.

## Intensity-Based Tumour Gradient Variations

Beyond boundary shape, the internal intensity characteristics and texture gradients of tumours provide valuable diagnostic information, potentially indicating different tissue compositions or stages of development. Generative models should ideally reproduce these internal patterns accurately. Changes in tumour intensity can reflect variations in vascularity or necrosis, influencing treatment strategy. This experiment investigates whether metrics or downstream tasks can detect simulated changes in internal tumour intensity gradients.

**Experimental Design** Within the identified tumour regions of the base MRI images, we introduce artificial intensity gradients based on spatial location relative to the tumour’s centre of mass (calculated per slice). Specifically, we applied a *radial intensity gradient*, modulating pixel intensities multiplicatively based on their distance from the tumour centre using a sinusoidal function to create structured variations. The magnitude of this radial intensity variation was systematically adjusted across several levels to simulate different degrees of internal tumour heterogeneity that might be distorted or poorly represented by generative models. The upstream NRIQMs were then computed, and downstream tumour segmentation performance was evaluated on the perturbed image set with modified internal gradients to assess sensitivity.

**Metric and Task Response Analysis** For both tumour boundary blurring and intensity-based gradient variations, we analyse the response of the upstream evaluation metrics and the downstream segmentation task performance. We examine how metric scores and segmentation accuracy/quality change as the degree of morphological manipulation increases. Metrics or downstream task evaluations that show significant changes are deemed more sensitive to these clinically relevant structural variations. Conversely, metrics or downstream task evaluations that remain relatively stable despite these alterations may fail to capture these important details, highlighting potential blind spots in the evaluation process. Comparing the upstream metric responses (or lack thereof) to any observed changes in downstream performance is particularly crucial for understanding whether the metrics reflect the practical impact of these subtle morphological changes.

## Distribution Shift

Distribution shifts pose a significant challenge in medical imaging, where generative models must accurately represent the inherent variability of clinical data without introducing biases or failing to capture the full spectrum of real-world cases. While metrics have been proposed to detect distribution shifts in natural(non-medical) image generation, their efficacy and interpretation in the context of subtle medical image variations require specific investigation. Such shifts can manifest in various ways, including data memorisation, mode collapse, density disparities due to different acquisition sources, and mode invention (hallucinations). We conduct a series of controlled experiments designed to evaluate the sensitivity of both upstream NRIQMs and

downstream task performance to these specific types of distribution shifts.

**Data Memorisation** Generative models, particularly high-capacity architectures like diffusion models, have been observed to memorise or closely replicate training data examples, which can lead to overly optimistic performance evaluations if not detected [4]. We assess metrics and downstream task sensitivity to two forms of data memorisation: external data duplication - replication of training examples and internal data duplication - duplication within the generated set itself. **Experimental Design:** We perform these two types of duplication experiments using the base image sets. To simulate external duplication, representing training set replication leakage into a test set, we create perturbed sets by gradually replacing base images with exact copies of images assumed to belong to a hypothetical ‘training’ partition, using replacement rates of [5%, 15%, 30%, 45%]. To simulate internal duplication, reflecting a lack of diversity within a generated set, we create perturbed sets by introducing duplicates within the base image set at the same rates [5%, 15%, 30%, 45%]. For both scenarios, we then compute upstream NRIQMs and evaluate downstream segmentation performance to assess their sensitivity to these forms of data memorisation.

**Mode Collapse** Mode collapse, a known issue particularly for some GAN architectures, occurs when a model fails to capture the full diversity of the data distribution, instead generating samples from only a limited subset of modes [80, 23]. In medical imaging, this could mean failing to generate examples of rarer anatomies or pathologies. We investigate this by manipulating the known class distribution within the IXI MRA dataset, categorised by anatomical variations of the posterior communicating artery (PCoA) in the Circle of Willis (bilateral, unilateral, or absent). **Experimental Design:** Our base MRA dataset has an approximate distribution of 40% bilateral PCoA, 40% unilateral PCoA, and 20% absent PCoA anatomies. We create two perturbed image sets simulating mode collapse: (1) we entirely remove the ‘absent PCoA’ class (the rarest mode), resulting in a 50%-50% distribution between the remaining two classes; and (2) we further alter the distribution to an imbalanced 1:5 ratio between the bilateral and unilateral PCoA classes, still excluding the absent PCoA class. We analyse how both upstream NRIQMs and downstream vessel segmentation performance respond to these explicit shifts in class distribution, assessing their ability to detect the missing or under-represented modes.

**Density Shift** Medical imaging data is often aggregated from diverse sources (e.g., different scanners or sites), potentially leading to shifts in image intensity distributions or textures due to variations in hardware, software, or acquisition protocols. Metrics should ideally be robust to, or capable of detecting, such domain shifts. We simulate this using the IXI MRA dataset, which contains images acquired using two different Philips scanners (Intera 3T and Gyroscan Intera 1.5T) with slightly varying parameters ([35]). **Experimental Design:** We start with a base image set containing a balanced 50%-50% distribution of images from each scanner. We then create perturbed image sets with deliberately skewed distributions [10-90, 25-75, 75-25, 90-10] representing dominance of one scanner’s data over the other. We evaluate how both upstream NRIQMs and downstream vessel segmentation performance respond to these density shifts arising from scanner differences.

**Mode Invention** Mode invention occurs when a generative model produces outputs containing features or structures not present in the training data distribution, essentially creating unrealistic or spurious content (hallucinations). In medical imaging, this could manifest as anatomically implausible structures or tissue characteristics. Metrics should ideally penalise such inventions. **Experimental Design:** We simulate subtle contrast-based mode invention by manipulating T2-weighted MRI base images from the IXI dataset. We apply controlled intensity remapping specifically within grey matter and white matter regions (identified using percentile-based intensity thresholds). A small, opposing scaling factor is applied to each tissue type, slightly increasing grey matter intensity while proportionally reducing white matter intensity, thus inverting their relative contrast locally in a physically implausible manner. This introduces novel structural appearances absent from the original data. Gaussian smoothing is applied post-scaling to maintain some anatomical coherence, followed by intensity normalisation. By repeating this manipulation with varying levels of contrast distortion, we evaluate whether upstream NRIQMs and downstream segmentation performance can detect these synthetic modes that deviate subtly but systematically from the real data distribution.

These distribution shift experiments provide a comprehensive evaluation framework. Fig. 3 visually summarises the different controlled perturbation experiments conducted in this phase.

### 3.2 Phase 2: Generative Model Performance Evaluation

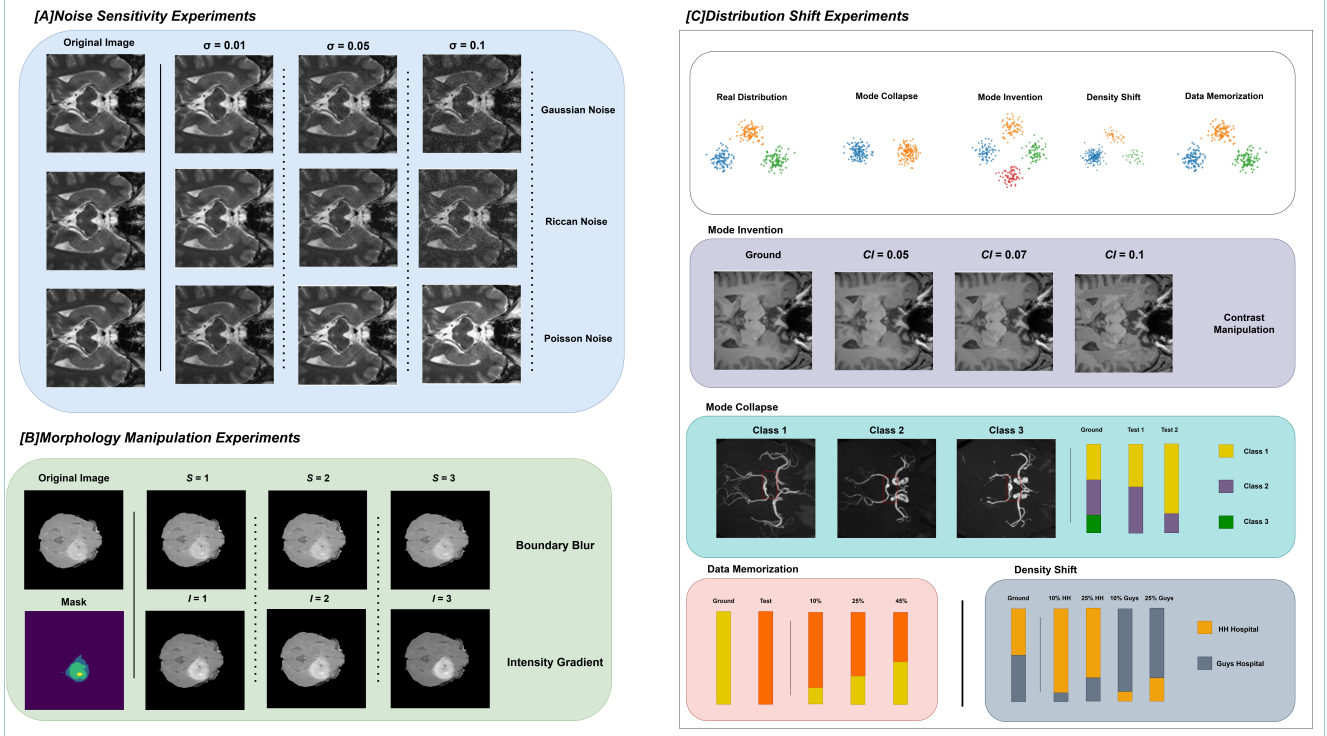
Building upon the insights from the controlled perturbation experiments, this second phase focuses on evaluating synthetic images produced by established generative model architectures. The goal is to assess how well the upstream NRIQMs and the downstream segmentation task performance align when applied to realistic outputs from different types of generative models operating on complex medical imaging data. We evaluate synthetic images corresponding to the target domains of the IXI (vascular MRA) and BraTS (brain tumour MRI) datasets.

For the generative model evaluation phase, we utilised publicly available, **pre-trained** implementations representing three prominent generative architectures. The specific models used for each dataset were:

- **For the IXI dataset:** Pre-trained model weights for all three architectures were sourced directly from the study by Deo et al. [19]. These specific implementations were:
  - **VAE:** A Vector Quantised VAE (VQ-VAE) [75].
  - **GAN:** A Wasserstein GAN (WGAN) [32].
  - **DDPM:** The Latent Diffusion Model (LDM) as implemented and detailed in the source paper [19].
- **For the BraTS dataset:**
  - **GAN:** A pre-trained Wasserstein GAN (WGAN) obtained from the Medigan library/toolkit [59]. The underlying architecture is based on [32].
  - **DDPM:** The Latent Diffusion Model (LDM) implementation described by Rombach et al. [66].

The set of synthetic images generated by each of these pre-trained models constitutes our *generated image set* for evaluation in this phase. We apply the full suite of upstream NRIQMs to these generated images, comparing their feature distributions against the corresponding real images (designated as the *reference distribution set* from Phase 1). Concurrently, we apply the pre-trained segmentation model (detailed in Section 2.2) to these same generated images to assess downstream task performance.

By systematically applying both upstream quantitative metrics and downstream task validation across controlled perturbations (Phase 1) and generative



**Figure 3. Overview of the controlled perturbation experiments in Phase 1. We apply noise, morphological, and domain-shift manipulations to real medical images to evaluate the sensitivity of NRIQMs.**

model outputs (Phase 2), this methodology enables a critical assessment of the NRIQMs. This comparative approach is designed to expose the conditions under which specific metrics succeed or fail, highlighting discrepancies between quantitative scores and practical task performance. Ultimately, this work aims to inform the selection and development of more reliable validation strategies for generative models in medical imaging, paving the way for evaluation frameworks that demonstrably correlate with clinical relevance and robustly identify model failure modes impacting potential downstream applications.

## 4 Results

In this section, we report the findings from our multifaceted evaluation framework, analysing the behaviour of NRIQMs introduced in Section 2.1. We present results from both the controlled perturbation experiments (Phase 1) and the evaluation of actual generative model outputs (Phase 2). Complete numerical results for all upstream metrics are provided in the supplementary materials (Section A.1). For the con-

trolled experiments, we compare metric scores obtained from perturbed image sets against baseline scores from real data. Upstream metric sensitivities are often using heat-maps of standardised Z-scores for intuitive comparison (as shown in Figs. 4-7). Downstream task performance is assessed quantitatively using the Evanyseg score described in Sec. 2.2.1.

### 4.1 Phase 1: Controlled Perturbation Experiments

#### 4.1.1 Noise Sensitivity Results

Our initial experiments investigated metric responses to controlled noise injections (Gaussian, Poisson, Rician). The sensitivity of upstream metrics is visualised in Fig. 4. As observed, distance-based metrics such as FID, sFID, KID, MMD, and FLS generally demonstrated increased dissimilarity (worsening scores) with rising noise levels, reacting particularly strongly to Gaussian and Rician noise. Learned perceptual metrics (LPIPS, DreamSIM) also consistently captured degradation across noise types. Conversely, several PRDC metrics (Precision, Recall, Density, Cov-

erage) and the Realism Score, while sensitive, often saturated towards worst-case scores even at moderate noise levels (especially Gaussian/Rician), potentially limiting their utility for differentiating finer degrees of degradation. Inception Score exhibited inconsistent trends across noise types: its mean value tended to decrease (worsen) with Gaussian/Rician noise but increase (improve) with Poisson noise, while its standard deviation showed the opposite pattern (improving stability with Gaussian/Rician but worsening with Poisson noise). The Vendi Score remained largely insensitive.

Downstream segmentation performance, assessed via the Evanyseg score, showed varied sensitivity depending on the noise type (Table 2). Gaussian noise injection resulted in only a moderate decrease in segmentation quality compared to the baseline (0.907). Interestingly, Poisson noise, even at the highest level tested, caused minimal degradation in the downstream task score. Rician noise, however, induced a substantial drop in segmentation performance, particularly at higher sigma values, aligning with its physical relevance in MRI and suggesting the downstream task is quite sensitive to this specific noise type. Comparing these findings, while upstream metrics such as FID reacted strongly to both Gaussian and Rician noise, the downstream task was significantly more impacted by Rician noise, highlighting a potential divergence where upstream metric sensitivity doesn’t perfectly predict downstream task robustness for all noise types.

**Table 2. Downstream Evanyseg scores for Noise Sensitivity experiments (Baseline: 0.907).**

Noise Condition	Evanyseg Score
<i>Gaussian</i>	
$\sigma = 0.01$	0.804
$\sigma = 0.05$	0.7886
$\sigma = 0.1$	0.7994
<i>Rician</i>	
$\sigma = 0.01$	0.904
$\sigma = 0.05$	0.748
$\sigma = 0.1$	0.451
<i>Poisson</i>	
$\sigma = 0.01$	0.893
$\sigma = 0.05$	0.889
$\sigma = 0.1$	0.877

#### 4.1.2 Morphological Manipulation Results

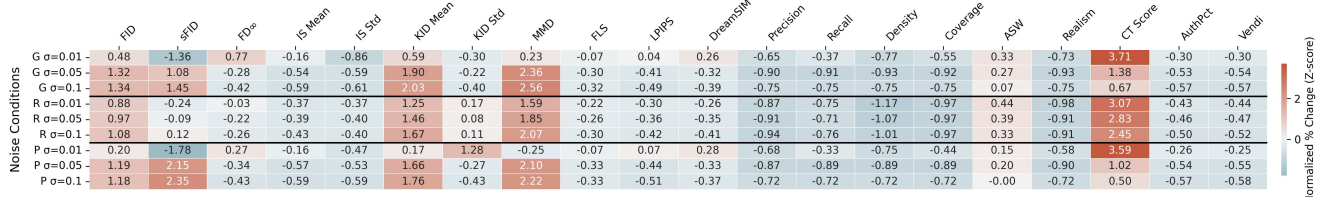
Experiments introducing localised morphological manipulations within clinical regions (specifically tumour boundary blurring and internal intensity gradient alterations in BRaTS images) yielded slightly contrasting results between upstream metrics and downstream task performance. As visualised in Fig. 5, a profound insensitivity was observed across nearly all tested upstream NRIQMs — including distance-based (FID, sFID, KID, MMD), perceptual (LPIPS, DreamSIM), PRDC metrics, and others such as AuthPct and Realism Score. These metrics showed negligible changes from baseline values even with increasing levels of boundary blurring or internal intensity variation, with only the CT Score exhibiting a minimal reaction.

In contrast, downstream segmentation performance, assessed via the Evanyseg score (Table 3), did show sensitivity to these morphological changes. Both tumour boundary blurring and the introduction of radial intensity gradients caused a measurable decrease in the segmentation quality score, with the degradation generally increasing as the manipulation strength ( $\sigma$ ) increased. While the absolute changes in the Evanyseg score might not be dramatic, the consistent downward trend indicates that the downstream segmentation task was negatively impacted by these subtle alterations to tumour morphology.

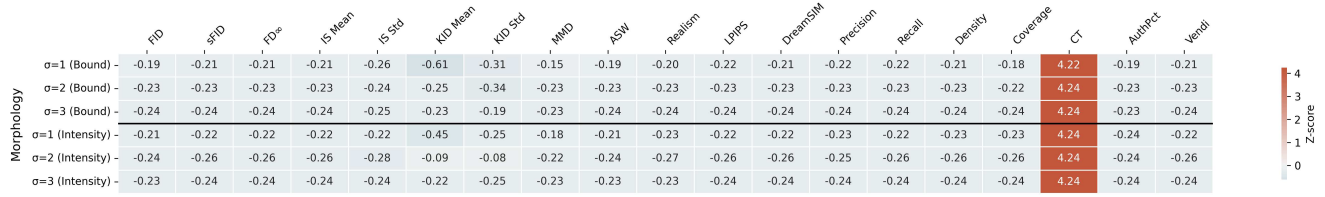
This comparison highlights a critical finding: while the downstream task registered a response to these clinically relevant morphological manipulations, the vast majority of upstream NRIQMs completely failed to detect these changes. This underscores the potential unreliability of relying solely on upstream metrics for evaluating the anatomical fidelity and fine-grained structural accuracy of generated medical images.

**Table 3. Downstream Evanyseg scores for Morphological Manipulation experiments (Baseline: 0.948).**

Manipulation	Evanyseg Score
<i>Intensity Gradient</i>	
$\sigma = 1$	0.934
$\sigma = 2$	0.902
$\sigma = 3$	0.881
<i>Boundary Blur</i>	
$\sigma = 1$	0.932
$\sigma = 2$	0.893
$\sigma = 3$	0.845



**Figure 4. Heatmap illustrating the sensitivity of each evaluation metric to various forms of noise perturbation.**



**Figure 5. Heatmap illustrating the sensitivity of upstream evaluation metrics to morphological perturbations. Note the widespread lack of response across most metrics, contrasting with the downstream sensitivity shown in Table 3.**

#### 4.1.3 Data Memorisation Results

Analysis of data memorisation effects, visualised via upstream metric sensitivities in Fig. 6, highlights more potential assessment pitfalls. As observed, metrics sensitive to distribution overlap (FID, sFID, KID, MMD) showed apparent improvement (lower distance scores) as the level of external duplication (training data leakage) increased. While reflecting increased feature similarity, this trend could be highly misleading if interpreted as better generative model quality in a real scenario. Concurrent with this, AuthPct dropped significantly, correctly identifying the decreased authenticity due to leakage, suggesting that combining distance metrics with AuthPct offers a potential signature for diagnosing external duplication. Conversely, most upstream metrics, including distance-based ones and even diversity metrics such as Vendi Score, showed minimal sensitivity to internal duplication (within-set repetition).

Downstream segmentation performance (Table 4) remained remarkably stable and very close to the baseline score (0.907) across all levels of both external and internal data duplication. This suggests that, for this specific downstream task and evaluation method, performance was largely unaffected by the presence of duplicated images in perturbed image sets, whether leaked from training or repeated internally. This contrasts sharply with the significant shifts observed in

upstream metrics such as FID and AuthPct under external duplication, emphasising that upstream distribution similarity does not always predict downstream task robustness or behaviour.

**Table 4. Downstream Evanyseg scores for Data Memorisation experiments (Baseline: 0.907).**

Condition	Evanyseg Score
<i>External Duplication</i>	
5%	0.908
15%	0.908
30%	0.907
45%	0.909
<i>Internal Duplication</i>	
5%	0.907
15%	0.908
30%	0.912
45%	0.912

#### 4.1.4 Density Shift, Mode Collapse and Mode Invention Results

Experiments altering data distribution (simulating density shifts between scanners and mode collapse by





**Figure 6. Heatmap illustrating the sensitivity of upstream evaluation metrics to increasing levels of data memorisation (External: Training set leakage; Internal: Within-set duplication).**

manipulating class proportions) and simulating mode invention (via artificial contrast manipulation) revealed varied sensitivities among upstream metrics (Fig. 7). Distance metrics (FID, sFID, KID, MMD, ASW) generally reacted to these shifts and manipulations, detecting deviations from the baseline distribution. However, their behaviour under mode collapse was potentially misleading, showing score ‘improvements’ when rare classes were removed. Recall and potentially Realism Score appeared more indicative of such mode coverage failures. AuthPct again showed contrasting diagnostic behaviour, decreasing with external data duplication, but increasing with mode invention, suggesting divergence from the real data manifold. LPIPS remained relatively insensitive to density shifts and mode collapse but reacted moderately to mode invention, while DreamSIM reacted counter-intuitively to mode invention.  $FD\infty$  and Vendi Score continued to show limited sensitivity across these diverse challenges.

Remarkably, downstream segmentation performance again remained almost entirely unaffected by density shifts, as well as by mode collapse and invention (Table 5). The Evanyseg scores stayed very close to the respective baselines across all tested conditions, including varying scanner proportions (density shift), removal or imbalance of anatomical classes (mode collapse), and artificial contrast invention (mode invention). This striking insensitivity of the downstream task performance contrasts significantly with the responses observed in many upstream metrics, particularly the distance-based ones and AuthPct, which clearly detected these distributional manipulations. This divergence strongly suggests that detecting a statistically significant shift in upstream feature distributions does not necessarily imply a meaningful impact on the utility of the images for this specific downstream segmentation task.

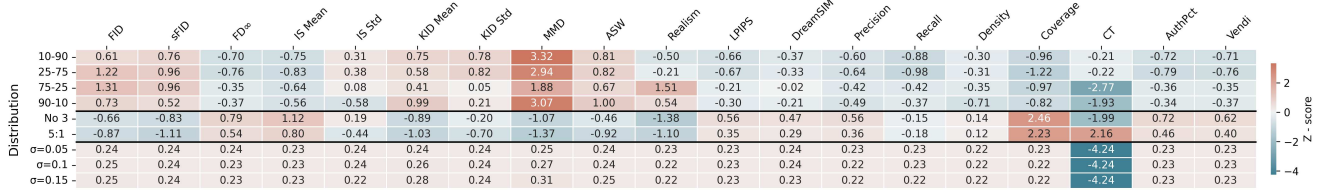
**Table 5. Downstream Evanyseg scores for Distribution Shift and Mode Invention experiments.**

Condition	Evanyseg Score
<i>Density Shift (Scanner Proportions)</i>	
Baseline (50-50)	0.907
10-90	0.904
25-75	0.909
75-25	0.902
90-10	0.901
<i>Mode Collapse (PCoA Classes)</i>	
Baseline (All Classes)	0.907
No Class 3 (50-50)	0.904
No Class 3 (5:1 Imbalance)	0.901
<i>Mode Invention (Contrast Manip.)</i>	
Baseline	0.907
$\sigma = 0.05$	0.909
$\sigma = 0.07$	0.909
$\sigma = 0.1$	0.908

## 4.2 Phase 2: Generative Model Performance Evaluation Results

Moving beyond controlled perturbations, we evaluated synthetic images produced by representative pre-trained generative models — a VAE, a GAN, and a DDPM — on the IXI and BRaTS datasets. We applied both the upstream NRIQMs (detailed in Sec. 2.1) and the downstream segmentation task evaluation (using the Evanyseg method detailed in Sec. 2.2.1), with results presented in Fig. 8 and Table 6 respectively to enable direct comparison of these different evaluation modalities on generative model outputs.

The upstream metric results reveal a striking pattern: nearly all evaluated NRIQMs indicated substantial deviations between the generated images (regardless of model type) and the real data baseline dis-



**Figure 7. Heatmap illustrating the sensitivity of upstream evaluation metrics to density shift, mode collapse and mode invention.**

tributions. As shown in Fig. 8, distance-based metrics (FID, sFID, KID, MMD, ASW) yielded markedly high dissimilarity scores for VAE, GAN, and DDPM compared to real data. While minor differences existed between models (e.g., DDPM often slightly better than GAN/VAE on some distance metrics for IXI), the overall magnitude of the deviation from the real baseline was typically large for all architectures, making nuanced quality ranking based solely on these metrics challenging. Furthermore, PRDC metrics generally collapsed to near-zero values, and AuthPct saturated at 100%, confirming that although these metrics effectively distinguished all generated distributions from the real one, they provided little discrimination between the generative models themselves. LPIPS showed moderate degradation across models, DreamSIM exhibited potentially counter-intuitive increases, and IS varied (notably high for DDPM on BRaTS). Overall, the upstream metrics strongly signalled a distributional mismatch for all models but struggled to differentiate them in a way that reflected expected visual quality differences.

In contrast, evaluating the models via the downstream segmentation task using Evanyseg scores, as shown in Table 6, provided a distinctly different and more discriminative assessment, aligning better with anticipated visual fidelity. On the IXI vessel segmentation task, the DDPM achieved the highest score (0.91), significantly outperforming the GAN (0.67) and also surpassing the VAE (0.86). Similarly, for BRaTS tumour segmentation, the DDPM (0.78) demonstrated clearly superior performance compared to the GAN (0.56). This downstream ranking reflects the typical expectation that DDPMs often produce images with higher structural coherence suitable for such tasks compared to VAEs or standard GANs. Intriguingly, results on IXI dataset highlight a direct contradiction: while the GAN produced ‘better’ upstream scores than the VAE on several distance metrics (visible in Fig. 8 data), its output was significantly less suitable for the downstream segmentation task compared to the VAE’s

output (0.67 vs 0.86). This discrepancy starkly underscores the potential unreliability of using only upstream metrics for model selection or quality assessment.

This observed divergence between the generally poor and undiscriminating upstream metric scores and the more differentiated downstream task performance reinforces a critical limitation of relying solely on many common distribution-focused NRIQMs. While these upstream metrics can effectively detect the *presence* of a distributional shift between generated and real images, they appear inadequate in characterising the *nature* of this shift in a way that consistently aligns with downstream task suitability across different generative architectures. These results strongly advocate for incorporating downstream task evaluations as a more meaningful and reliable method for assessing the practical utility of generative models in medical imaging applications.

**Table 6. Calculated Evanyseg Dice scores for downstream evaluation of generative models.**

Model	IXI Vessel Segmentation	Brats Tumour Segmentation
VAE	0.86	-
GAN	0.67	0.56
DDPM	0.91	0.78

## 5 Discussion

Our systematic evaluation investigated the reliability of commonly used NRIQMs for assessing generative models in medical imaging, comparing quantitative upstream metric scores against downstream task performance. Multidisciplinary perspectives on simulation, data, and model evaluation help to contextualise our findings, which reveal significant limitations in relying solely on upstream metrics and emphasise the need for

**Table 7. This table summarises some of the key findings and Recommendations for Evaluating Generative Medical Images with NRIQMs**

#	Recommendation	Context	Implication / Why it Matters for Reliable Evaluation
1	<b>Prioritise Task-Based (Downstream) Evaluation:</b> It proved more reliable and discriminative for assessing practical model utility than upstream metrics alone.	Upstream vs. Downstream Comparison (All Models); Evanyseg results vs. Upstream metrics, e.g. FID/KID etc.	Task performance better reflects clinical relevance, effectively differentiates between models (e.g., DDPM vs. GAN), and aligns more closely with desired visual/functional quality.
2	<b>Use Distance Metrics With Extreme Caution:</b> These metrics are highly prone to misleading scores despite sensitivity to global data distribution shifts.	FID, KID, MMD, ASW, sFID	Scores can falsely suggest improvement with data leakag/memorization or mode collapse. Correlated poorly with task performance across models. Unreliable for quality ranking alone.
3	<b>Recognise Metric Blindness to Local Detail:</b> Most tested upstream metrics fail to detect fine-grained anatomical or pathological errors.	All Upstream Metrics remained largely unchanged in Local Boundary&Intensity Manipulation experiments.	Critical Gap: Cannot ensure local anatomical fidelity, potentially missing subtle but clinically significant errors. Poses a safety concern if used as primary validation.
4	<b>Actively Investigate Mode Collapse &amp; invention:</b> Do not rely solely on distance metrics, which can be deceptively optimistic.	FID/KID (misleading improvement), Recall potentially signals degradation, Downstream testing insensitive.	Models might fail to generate rare but critical conditions, necessitating targeted evaluation beyond global scores, e.g. monitoring recall and performing visual & task-specific inspection.
5	<b>Diagnose Data Memorisation Using Combined Signals:</b> Look beyond seemingly 'improved' distance scores.	FID/KID (misleading improvement), AuthPct (authenticity score drop).	Combining these signals helps differentiate genuine model performance from artificial score inflation caused by training data repetition in the test set.
6	<b>Avoid Metrics Demonstrated to be Unreliable or Insensitive:</b> Certain metrics consistently provided little useful information or behaved erratically.	Vendi Score, $FD_{\infty}$ (Insensitive); IS, DreamSIM (Inconsistent/Counter-intuitive).	These specific metrics were unsuitable or unreliable for the challenges tested in these controlled perturbation experiments
8	<b>Adopt a Multifaceted Evaluation Strategy:</b> No single metric provides a complete or trustworthy picture of generative model quality.	Synthesis of all experimental findings.	Robust assessment requires combining cautiously interpreted upstream metrics, essential downstream task validation, and potentially expert visual inspection.



**Figure 8. Heatmap illustrating the sensitivity of upstream evaluation metrics across different generative model outputs (VAE, GAN, DDPM on IXI and BRaTS datasets). Scores represent standardised differences relative to a baseline or between models, highlighting how metrics differentiate model outputs.**

purpose-driven validation frameworks [9, 64]. Generated images, unlike direct clinical measurements, are outputs of complex computational processes – akin to simulations – whose epistemic status is inherently indirect [62, 63]. They represent the model’s learned approximation of reality, potentially lacking fidelity due to uncaptured factors or ‘unknown unknowns’ inherent in the real-world data generating process, placing them epistemically ‘one step behind’ direct experimentation or measurement [69].

A primary empirical observation was the sensitivity of distance-based metrics (e.g., FID, KID) to global distributional characteristics, yet their potential to mislead. Their ‘improvement’ with data duplication or mode collapse highlights a failure to capture crucial aspects of generative quality beyond statistical moments. Furthermore, their poor correlation with downstream task performance across different model architectures (VAE, GAN, DDPM) suggests they often fail the test of ‘adequacy-for-purpose’ [9, 64]. The hypothesis that the homogeneity of medical images compared to diverse natural scenes might challenge these metrics, especially those using features from natural image models, warrants further investigation. If metrics cannot reliably distinguish subtle yet critical pathological variations, their utility diminishes significantly.

Other metrics showed varied utility. The saturation of PRDC metrics limited their discriminative power. AuthPct and Recall, however, showed promise as diagnostic tools for specific failure modes (leakage, mode invention, mode collapse), aligning with the idea that data evaluation involves targeted assessment for specific purposes [9] rather than seeking a single quality score. The consistent insensitivity of Vendi Score and  $FD\infty$  underscores the difficulty of finding universally effective NRIQMs.

Perhaps the most critical finding, viewed through the lens of distinguishing data from models [47] and experiment from simulation [69], is the profound in-

sensitivity of almost all upstream metrics to localised morphological manipulations. These metrics, treating generated images as data points in a feature space, failed to detect alterations in boundary sharpness or internal gradients – features crucial for clinical interpretation and downstream tasks such as segmentation. While our downstream evaluation did register performance degradation (although limited) under these conditions, the upstream metrics largely failed. This highlights a significant risk: generative models might produce anatomically flawed images that pass common quantitative metrics checks but would be unsuitable – or even hazardous – for applications such as augmenting training datasets for developing clinical decision support tools, potentially leading these tools to learn incorrect anatomical representations. These findings highlight that upstream metrics often lack the representational granularity needed to validate anatomical fidelity. This aligns with the philosophical view that evaluating representations requires considering the specific purpose; metrics optimised for global distribution matching may be inadequate for purposes requiring local structural accuracy [9].

The divergence between upstream and downstream results, particularly the VAE/GAN contradiction on IXI, strongly supports the ‘adequacy-for-purpose’ approach to evaluation [64, 9]. It suggests that synthetic data which misrepresent the real distribution according to upstream metrics can still be adequate for specific downstream tasks. However, our finding that the downstream task was insensitive to various distributional shifts (data memorisation, density shifts, mode invention) that were detected by upstream metrics serves as a crucial counterpoint. It implies that downstream evaluation, while necessary for assessing utility, is not sufficient alone. It evaluates fitness for one purpose and may overlook other significant flaws, aligning with Parker’s discussion of model evaluation challenges [64].

These findings carry significant implications for deploying generative models in safety-critical medical applications. Relying on easily computed upstream metrics - without validating downstream utility - risks deploying flawed models or clinical tools trained on inadequate synthetic data. Safety is highly context-dependent, and even slight variation in the clinical settings can undermine the validity of the existing evidence [30]. Further, viewing generated images as akin to simulation outputs, offering potentially valuable insights [63], demands a rigorous validation process focused on their fitness for specific clinical purposes [64]. Therefore, we advocate strongly for a multifaceted evaluation framework, integrating downstream task validation with context-aware interpretation of appropriate upstream metrics to ensure both utility and safety.

**Limitations of this Study:** Our findings are primarily based on two specific MRI datasets and one downstream task (segmentation). Metric performance and the upstream-downstream relationship may differ across other modalities, anatomical regions, or clinical tasks. The specific generative models and perturbation methods used may not cover all possibilities. Furthermore, the downstream evaluation relied on a specific pre-trained segmentor and the Evanyseg method, introducing dependencies on these components.

**Suggestions for Future Research:** Our findings underscore several critical areas for future research, aligning with recognised needs in the evaluation of generative medical imaging. Firstly, addressing the **limitations of existing metrics**, particularly the noted issues with widely-used distance metrics such as FID in medical contexts [2] and the general insensitivity of upstream metrics to localised anatomical details observed in our work, necessitates the development of novel metrics. Future metrics should aim for enhanced sensitivity to fine-grained, clinically relevant features and validation against specific medical image failure modes. Secondly, the significant divergence we observed between upstream metric scores and downstream task performance reinforces the growing consensus on the importance of downstream task evaluation [36]. Further work is needed to validate findings across a broader spectrum of clinical tasks and imaging modalities, and to develop robust, potentially reference-free, methods for assessing downstream utility reliably, possibly incorporating insights from approaches using human feedback[61]. Integrating insights from both improved upstream metrics and rigorous downstream task validation will be key to building truly trustworthy generative models for healthcare. An integrated approach, grounded in a purpose-driven evaluation philosophy [9, 64], is vital for the responsible

progress of generative models in healthcare.

## 6 Conclusion

Our comprehensive assessment revealed significant limitations in the utility of many commonly used NRIQMs for evaluating generative models in medical imaging. While sensitive to certain global distributional characteristics, these metrics frequently correlate poorly with practical utility, often fail to detect clinically relevant localised inaccuracies, and can provide misleading assessments regarding issues such as mode collapse or data memorisation. We observed a stark divergence between upstream metric scores and downstream task performance, with the latter often proving more discriminative but also exhibiting its own limitations, such as insensitivity to certain distributional shifts. Therefore, we conclude that neither upstream nor downstream evaluation alone is sufficient for reliable validation. Robust and trustworthy assessment of generative models necessitates a multifaceted framework, integrating insights from relevant downstream task evaluations with the cautious interpretation of appropriately selected upstream metrics. Addressing the identified gaps requires future research focused on developing novel metrics with enhanced sensitivity to localised anatomical features and clinical relevance. This should be accompanied by broader validation across diverse medical tasks and modalities, as well as continued refinement of robust downstream task evaluation methodologies. Moving forward with such integrated and improved evaluation strategies is essential for the safe and effective translation of generative models into clinical practice.

## 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 Research 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).

## Competing interest

The authors declare no competing interests.

## References

- [1] Ehsan Abadi, William P Segars, Benjamin MW Tsui, Paul E Kinahan, Nick Bottenus, Alejandro F Frangi, Andrew Maidment, Joseph Lo, and Ehsan Samei. Virtual clinical trials in medical imaging: a review. *J Med Imag*, 7(4):042805–042805, 2020.
- [2] A. Bobomirzaevich Abdusalomov, Rakhmatillo Rakhmatullaevich Nasimov, Nodirabegim Rakhmatullaevna Nasimova, Bobir Mirzayevich Muminov, and Taeg Keun Whangbo. Evaluating synthetic medical images using artificial intelligence with the gan algorithm. *Sensors*, 23(7):3440, 2023.
- [3] Waqar Ahmad, Hazrat Ali, Zubair Shah, and Shoaib Azmat. A new generative adversarial network for medical images super resolution. *Sci Rep*, 12(1):9533, 2022.
- [4] 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*, 6(1):015022, 2025.
- [5] 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*, pages 290–306. PMLR, 2022.
- [6] Spyridon Bakas, Hamed Akbari, Aristeidis Sotiras, Michel Bilello, Martin Rozycki, Justin S Kirby, John B Freymann, Keyvan Farahani, and Christos Davatzikos. Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. *Scientific data*, 4(1):1–13, 2017.
- [7] 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.
- [8] Mikołaj Bińkowski, Danica J. Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs, 2021.
- [9] Alisa Bokulich and Wendy S Parker. Data models, representation, and adequacy-for-purpose. *European Journal for Philosophy of Science*, 11(1):31, 2021.
- [10] Sihong Chen, Kai Ma, and Yefeng Zheng. Med3d: Transfer learning for 3d medical image analysis, 2019.
- [11] Nina Cheng, Zhengji Liu, Yash Deo, Haoran Dou, Ning Bi, Kun Wu, Fengming Lin, Zeike A Taylor, Nishant Ravikumar, and Alejandro F Frangi. Synthesising 3d cardiac cine-mr images and corresponding segmentation masks using a latent diffusion model. In *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE, 2024.
- [12] Min Jin Chong and David Forsyth. Effectively unbiased FID and Inception score and where to find them. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6069–6078. IEEE, 2020.
- [13] Li Sze Chow and Raveendran Paramesran. Review of medical image quality assessment. *Biomed Signal Process Control*, 27:145–154, 2016.
- [14] Li Sze Chow, Heshalini Rajagopal, Raveendran Paramesran, Alzheimer’s Disease Neuroimaging Initiative, et al. Correlation between subjective and objective assessment of magnetic resonance (MR) images. *Magn Reson Imaging*, 34(6):820–831, 2016.
- [15] Hyungiin Chung and Jong Chul Ye. Score-based diffusion models for accelerated MRI. *Med Image Anal*, 80:102479, 2022.
- [16] Matthew C Clark, Lawrence O Hall, Dmitry B Goldgof, Richard F Murtagh, and Michael L Silbiger. Automatic tumor segmentation using knowledge-based techniques. *IEEE Trans Med Imag*, 17(2):187–201, 2006.
- [17] Gian Marco Conte, Alexander D Weston, David C Vogelsang, Kenneth A Philbrick, Jason C Cai, Maurizio Barbera, Francesco Sanvito, Daniel H Lachance, Robert B Jenkins, W Oliver Tobin, et al. Generative adversarial networks to synthesize missing t1 and flair mri sequences for use in a multisequence brain tumor segmentation model. *Radiology*, 299(2):313–323, 2021.
- [18] Yash Deo, Rodrigo Bonazzola, Haoran Dou, Yan Xia, Tianyou Wei, Nishant Ravikumar, Alejandro F Frangi, and Toni Lassila. Learned local

- attention maps for synthesising vessel segmentations from T2 MRI. In *International Workshop on Simulation and Synthesis in Medical Imaging*, pages 32–41. Springer, 2023.
- [19] Yash Deo, Haoran Dou, Nishant Ravikumar, Alejandro F Frangi, and Toni Lassila. Shape-guided conditional latent diffusion models for synthesising brain vasculature. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 164–173. Springer, 2023.
- [20] Yash Deo, Fengming Lin, Haoran Dou, Nina Cheng, Nishant Ravikumar, Alejandro F Frangi, and Toni Lassila. Few-shot learning in diffusion models for generating cerebral aneurysm geometries. In *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE, 2024.
- [21] M. Dohmen, M.A. Klemens, I.M. Baltruschat, T. Truong, and M. Lenga. Similarity and quality metrics for MR image-to-image translation. *Sci Rep*, 15(1):3853, 2025.
- [22] Zolnamar Dorjsembe, Sodtavilan Odonchimed, and Furen Xiao. Three-dimensional medical image synthesis with denoising diffusion probabilistic models. In *Medical Imaging with Deep Learning*, 2022.
- [23] Ricard Durall, Avraam Chatzimichailidis, Peter Labus, and Janis Keuper. Combating mode collapse in GAN training: An empirical analysis using hessian eigenvalues, 2020.
- [24] Azin Shokraei Fard, David C Reutens, and Viktor Vegh. From CNNs to GANs for cross-modality medical image estimation. *Comput Biol Med*, 146:105556, 2022.
- [25] Virginia Fernandez, Walter Hugo Lopez Pinaya, Pedro Borges, Mark S. Graham, Petru-Daniel Tudosi, Tom Vercauteren, and M. Jorge Cardoso. Generating multi-pathological and multi-modal images and labels for brain mri. *Medical Image Analysis*, 97:103278, 2024.
- [26] Virginia Fernandez, Walter Hugo Lopez Pinaya, Pedro Borges, Petru-Daniel Tudosi, Mark S. Graham, Tom Vercauteren, and M. Jorge Cardoso. *Can Segmentation Models Be Trained with Fully Synthetically Generated Data?*, page 79–90. Springer International Publishing, 2022.
- [27] Dan Friedman and Adji Bousso Dieng. The Vendi score: A diversity evaluation metric for machine learning, 2023.
- [28] H. Gudbjartsson and S. Patz. The Rician distribution of noisy MRI data. *Magn Reson Med*, 34(6):910–4, December 1995. Erratum in: *Magn Reson Med* 1996 Aug;36(2):332.
- [29] Alper Güngör, Salman UH Dar, Şaban Öztürk, Yilmaz Korkmaz, Hasan A Bedel, Gokberk Elmas, Muzaffer Ozbey, and Tolga Çukur. Adaptive diffusion priors for accelerated mri reconstruction. *Med Image Anal*, 88:102872, 2023.
- [30] Ibrahim Habli, Richard Hawkins, Colin Paterson, Philippa Ryan, Yan Jia, Mark Sujan, and John McDermid. The big argument for ai safety cases. *arXiv preprint arXiv:2503.11705*, 2025.
- [31] Changhee Han, Hideaki Hayashi, Leonardo Rundo, Ryosuke Araki, Wataru Shimoda, Shinichi Muramatsu, Yujiro Furukawa, Giancarlo Mauri, and Hideki Nakayama. GAN-based synthetic brain MR image generation. In *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*, pages 734–738. IEEE, 2018.
- [32] Changhee Han, Hideaki Hayashi, Leonardo Rundo, Ryosuke Araki, Wataru Shimoda, Shinichi Muramatsu, Yujiro Furukawa, Giancarlo Mauri, and Hideki Nakayama. GAN-based synthetic brain MR image generation. In *2018 IEEE 15th international Symposium on Biomedical Imaging (ISBI 2018)*, pages 734–738. IEEE, 2018.
- [33] 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*, 30, 2017.
- [34] Pham Ngoc Huy and Tran Minh Quan. Denoising diffusion medical models. In *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE, 2023.
- [35] Information eXtraction from Images Consortium. IXI dataset – brain development. <https://brain-development.org/ixi-dataset/>. Accessed: 2023-02-14.
- [36] Abhinav K Jha, Kyle J Myers, Nancy A Obuchowski, Ziping Liu, Md Ashequr Rahman, Babak Saboury, Arman Rahmim, Barry A Siegel, Eric C Frey, and Ehsan Samei. Objective task-based



- evaluation of artificial intelligence-based medical imaging methods: Framework, strategies, and role of the physician. *Journal of Medical Imaging*, 9(1):011004, 2022.
- [37] Lan Jiang, Ye Mao, Xiangfeng Wang, Xi Chen, and Chao Li. Cola-diff: Conditional latent diffusion model for multi-modal MRI synthesis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 398–408. Springer, 2023.
- [38] Yifan Jiang, Han Chen, Murray Loew, and Hanseok Ko. COVID-19 CT image synthesis with a conditional generative adversarial network. *IEEE J Biomed Health Inform*, 25(2):441–452, 2020.
- [39] Marco Jiralerspong, Avishek Joey Bose, Ian Gemp, Chongli Qin, Yoram Bachrach, and Gauthier Gidel. Feature likelihood divergence: Evaluating the generalization of generative models using samples, 2024.
- [40] Sergey Kastruyulin, Jamil Zakirov, Nicola Pezzotti, and Dmitry V Dylov. Image quality assessment for magnetic resonance imaging. *IEEE Access*, 11:14154–14168, 2023.
- [41] Amirhossein Kazerouni, Ehsan Khodapanah Aghdam, Moein Heidari, Reza Azad, Mohsen Fayyaz, Ilker Hacihaliloglu, and Dorit Merhof. Diffusion models in medical imaging: A comprehensive survey. *Med Image Anal*, 88:102846, 2023.
- [42] Aghiles Kebaili, Jérôme Lapuyade-Lahorgue, and Su Ruan. Deep learning approaches for data augmentation in medical imaging: a review. *J Imaging*, 9(4):81, 2023.
- [43] Firas Khader, Gustav Müller-Franzes, Soroosh Tayebi Arasteh, Tianyu Han, Christoph Haarbuerger, Maximilian Schulze-Hagen, Philipp Schadt, Sandy Engelhardt, Bettina Baeßler, Sebastian Försch, et al. Denoising diffusion probabilistic models for 3D medical image generation. *Sci Rep*, 13(1):7303, 2023.
- [44] D.P. Kingma and M. Welling. Auto-encoding variational Bayes. *arXiv.org*, abs/1312.6114, 2013.
- [45] Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. *Adv Neural Inf Process Syst*, 32, 2019.
- [46] Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models, 2019.
- [47] Sabina Leonelli. What distinguishes data from models? *European Journal for Philosophy of Science*, 9(2):22, 2019.
- [48] Lucie Lévêque, Meriem Outtas, Hantao Liu, and Lu Zhang. Comparative study of the methodologies used for subjective medical image quality assessment. *Phys Med Biol*, 66(15):15TR02, 2021.
- [49] Chaofeng Li and Alan C. Bovik. Content-partitioned structural similarity index for image quality assessment. *Signal Processing: Image Communication*, 25(7):517–526, 2010. Special Issue on Image and Video Quality Assessment.
- [50] Yunxiang Li, Hua-Chieh Shao, Xiaoxue Qian, and You Zhang. Fddm: Unsupervised medical image translation with a frequency-decoupled diffusion model, 2024.
- [51] Fengming Lin, Yan Xia, Shuang Song, Nishant Ravikumar, and Alejandro F Frangi. High-throughput 3DRA segmentation of brain vasculature and aneurysms using deep learning. *Comput Methods Programs Biomed*, 230:107355, 2023.
- [52] Yintong Liu, U Rajendra Acharya, and Jen Hong Tan. Preserving privacy in healthcare: A systematic review of deep learning approaches for synthetic data generation. *Comput Methods Programs Biomed*, page 108571, 2024.
- [53] Allister Mason, James Rioux, Sharon E Clarke, Andreu Costa, Matthias Schmidt, Valerie Keough, Thien Huynh, and Steven Beyea. Comparison of objective image quality metrics to expert radiologists’ scoring of diagnostic quality of MR images. *IEEE Trans Med Imag*, 39(4):1064–1072, 2019.
- [54] Casey Meehan, Kamalika Chaudhuri, and Sanjoy Dasgupta. A non-parametric test to detect data-copying in generative models, 2020.
- [55] 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 transactions on medical imaging*, 34(10):1993–2024, 2014.



- [56] Kimia Nadjahi, Alain Durmus, Pierre E Jacob, Roland Badeau, and Umut Simsekli. Fast approximation of the sliced-Wasserstein distance using concentration of random projections. *Adv Neural Inf Process Syst*, 34:12411–12424, 2021.
- [57] Muhammad Ferjad Naeem, Seong Joon Oh, Youngjung Uh, Yunjey Choi, and Jaejun Yoo. Reliable fidelity and diversity metrics for generative models. In *International Conference on Machine Learning*, pages 7176–7185. PMLR, 2020.
- [58] Charlie Nash, Jacob Menick, Sander Dieleman, and Peter Battaglia. Generating images with sparse representations. In *International Conference on Machine Learning*, pages 7958–7968. PMLR, 2021.
- [59] Richard Osuala, Grzegorz Skorupko, Noussair Lazrak, Lidia Garrucho, Eloy García, Smriti Joshi, Socayna Jouide, Michael Rutherford, Fred Prior, Kaisar Kushibar, et al. medigan: a python library of pretrained generative models for medical image synthesis. *Journal of Medical Imaging*, 10(6):061403, 2023.
- [60] Muzaffer Özbey, Onat Dalmaz, Salman UH Dar, Hasan A Bedel, Şaban Öztürk, Alper Güngör, and Tolga Çukur. Unsupervised medical image translation with adversarial diffusion models. *IEEE Transactions on Medical Imaging*, 42(12):3524–3539, 2023.
- [61] Neha S. Pak, Yuyin Yong, Hakjoo Lee, Letian Sun, Semir Beraki, Linda Wang, Roger Chen, Mihela van der Schaar, and Ahmed M Alaa. Aligning synthetic medical images with clinical knowledge using human feedback. In *Advances in Neural Information Processing Systems*, volume 36.
- [62] Wendy S Parker. Virtually a measurement. *Nature Physics*, 17(2):146, 2021.
- [63] Wendy S Parker. Evidence and knowledge from computer simulation. *Erkenntnis*, 87(4):1521–1538, 2022.
- [64] Wendy S Parker. Model evaluation. In *The Routledge Handbook of Philosophy of Scientific Modeling*, pages 208–219. Routledge, 2024.
- [65] Cheng Peng, Pengfei Guo, S Kevin Zhou, Vishal M Patel, and Rama Chellappa. Towards performant and reliable undersampled mr reconstruction via diffusion model sampling. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 623–633. Springer, 2022.
- [66] Walter H. L. Pinaya, Petru-Daniel Tudosiu, Jessica Dafflon, Pedro F da Costa, Virginia Fernandez, Parashkev Nachev, Sebastien Ourselin, and M. Jorge Cardoso. Brain imaging generation with latent diffusion models, 2022.
- [67] Walter HL Pinaya, Petru-Daniel Tudosiu, Jessica Dafflon, Pedro F Da Costa, Virginia Fernandez, Parashkev Nachev, Sebastien Ourselin, and M Jorge Cardoso. Brain imaging generation with latent diffusion models. In *MICCAI Workshop on Deep Generative Models*, pages 117–126. Springer, 2022.
- [68] Zhihang Ren, X Yu Stella, and David Whitney. Controllable medical image generation via generative adversarial networks. In *IS&T International Symposium on Electronic Imaging*, volume 33. NIH Public Access, 2021.
- [69] Sherrilyn Roush. The epistemic superiority of experiment to simulation. *Synthese*, 195(11):4883–4906, 2018.
- [70] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training GANs. *Adv Neural Inf Process Syst*, 29, 2016.
- [71] Ahjol Senbi, Tianyu Huang, Fei Lyu, Qing Li, Yuhui Tao, Wei Shao, Qiang Chen, Chengyan Wang, Shuo Wang, Tao Zhou, and Yizhe Zhang. Towards ground-truth-free evaluation of any segmentation in medical images, 2024.
- [72] Nripendra Kumar Singh and Khalid Raza. Medical image generation using generative adversarial networks: A review. *Health informatics: A computational perspective in healthcare*, pages 77–96, 2021.
- [73] Yang Song, Liyue Shen, Lei Xing, and Stefano Ermon. Solving inverse problems in medical imaging with score-based generative models. *arXiv preprint arXiv:2111.08005*, 2021.
- [74] George Stein, Jesse Cresswell, Rasa Hosseinzadeh, Yi Sui, Brendan Ross, Valentin Vilecroze, Zhaoyan Liu, Anthony L Caterini, Eric Taylor, and Gabriel Loaiza-Ganem. Exposing flaws of generative model evaluation metrics and their unfair treatment of diffusion models. *Adv Neural Inf Process Syst*, 36, 2024.

- [75] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Adv Neural Inf Process Syst*, 30, 2017.
- [76] Z. Wang, E.P. Simoncelli, and A.C. Bovik. Multiscale structural similarity for image quality assessment. In *The Thirty-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, volume 2, pages 1398–1402 Vol.2, 2003.
- [77] Krishna Pandu Wicaksono, Koji Fujimoto, Yasutaka Fushimi, Akihiko Sakata, Sachi Okuchi, Takuya Hinoda, Satoshi Nakajima, Yukihiro Yamao, Kazumichi Yoshida, Kanae Kawai Miyake, et al. Super-resolution application of generative adversarial network on brain time-of-flight mr angiography: image quality and diagnostic utility evaluation. *European Radiology*, 33(2):936–946, 2023.
- [78] McKell Woodland, Austin Castelo, Mais Al Taie, Jessica Albuquerque Marques Silva, Mohamed Eltaher, Frank Mohn, Alexander Shieh, Supruteek Kundu, Joshua P. Yung, Ankit B. Patel, and Kristy K. Brock. Feature extraction for generative medical imaging evaluation: New evidence against an evolving trend. *arXiv.org*, abs/2311.13717, 2024.
- [79] Yueyue Xiao, Chunxiao Chen, Liang Wang, Jie Yu, Xue Fu, Yuan Zou, Zhe Lin, and Kunpeng Wang. A novel hybrid generative adversarial network for ct and mri super-resolution reconstruction. *Physics in Medicine & Biology*, 68(13):135007, 2023.
- [80] Kaifeng Zhang. On mode collapse in generative adversarial networks. In *Artificial Neural Networks and Machine Learning–ICANN 2021: 30th International Conference on Artificial Neural Networks, Bratislava, Slovakia, September 14–17, 2021, Proceedings, Part II 30*, pages 563–574. Springer, 2021.
- [81] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018.