Counterfactual Stress-Testing for Fixing Lung Cancer Segmentation Models

Ainkaran Santhirasekaram[2,*], Charlie Johnson[1,*], and Mitchell Chen[1]

[1] Department of Surgery and Cancer, Imperial College London, UK
[2] MRC Laboratory of Medical Sciences, Imperial College London, UK
a.santhirasekaram19@imperial.ac.uk

Abstract. Segmentation models for lung cancer on CT can achieve strong performance yet remain brittle to clinically relevant morphological and textural variations, with failures more prevalent in certain phenotypes. We introduce a counterfactual stress-testing framework that makes this brittleness measurable and actionable. We learn a conditional 3D tumour generator that produces identity-preserving counterfactual CT crops and corresponding pseudo-label masks by intervening on a compact set of radiomic "parent" features. Sweeping each feature across its empirical quantiles produces Counterfactual Robustness Curves (CRCs) for a baseline segmentation model, highlighting the feature regimes in which performance collapses. We define hard regimes as CRC regions where performance falls below a user-defined threshold and use them to select counterfactuals for targeted fine-tuning. On a publicly available CT cohort, CRCs expose consistent failure regions for small volumes, low sphericity, and high fractal dimension. Counterfactual fidelity analysis confirms that interventions achieve the intended feature changes with limited drift in non-target features. Targeted counterfactual fine-tuning guided by CRCs improves test segmentation performance compared with classical data augmentation and unconditional generative sampling. Code is available at https://github.com/AinkaranSanthi/counterfactual-stress-testing-lung-segmentation

Keywords: Counterfactual Images · Radiomics · Lung cancer.

1 Introduction

Lung cancer remains the world's leading cause of cancer-related mortality. CT underpins its staging, radiotherapy planning, and disease surveillance [4]. In radiotherapy, accurate definition of the gross tumour volume is clinically critical because substantial inter-observer contouring variability can lead to inadequate target coverage or unnecessary irradiation of healthy tissue, increasing toxicity [26,27]. Automatic segmentation can reduce workload and improve reproducibility, but safe deployment depends on robustness to the wide range of tumour phenotypes encountered in routine practice.

* Joint first authors.

Deep learning dominates lung cancer segmentation on CT, with 3D U-Net-based models delivering the strongest benchmark performances [17,10]. However, mean Dice score can conceal brittle behaviour, as performance collapses for certain imaging phenotypes including very small lesions and those with poorly defined tumour boundaries, which can arise from peritumoural ground glass opacification, or adjacent anatomical structures with similar attenuation characteristics, such as atelectatic lung, pleura or mediastinal structures. [13]. Generic robustness evaluations are important but do not necessarily isolate which tumour attributes drive failure [11].

Radiomics provides a compact language for clinically interpretable attributes such as size, boundary complexity, and texture heterogeneity [14]. If we could intervene on such descriptors within an individual patient context; changing only the tumour attribute while preserving surrounding anatomy; we could construct controlled distribution shifts to stress-test segmentation models mechanistically.

Contributions. 1. We introduce a counterfactual stress-testing framework for lung tumour segmentation that intervenes on clinically interpretable radiomic parents while preserving the individual patient context. This yields identity-preserving counterfactual crops with pseudo-label masks. We summarize robustness using Counterfactual Robustness Curves (CRCs) evaluated over deciles. 2. We propose CRC-guided repair via hard mining of failure-inducing counterfactuals and benchmark it against four other augmentation based methods.

2 Background and Related Work

Robustness in medical segmentation is often pursued via domain generalization and synthetic transformations [7,12]. Deep stacked transformations (BigAug) combine intensity and geometric augmentations to mimic unseen domains [31]. However, these do not target clinically meaningful tumour attributes (size, compactness, texture heterogeneity). Generative modelling offers an alternative by mixing synthetic images sampled from noise with real data. In CT, Generative adverserial Networks (GANs) have been used to augment contrast and appearance variation to improve robustness across scanners and acquisition protocols [23]. While this increases diversity, it often fails to preserve patient anatomy, limiting its utility for probing individual-level failure modes [9].

Stress testing reframes evaluation as mapping a model's performance under controlled shifts rather than relying on a single test set [22,28]. Current robustness frameworks quantify sensitivity to noise, motion and other corruptions, but the resulting failure modes are rarely clinically interpretable [3,24,25].

Counterfactual reasoning answers "what if" questions by modelling interventions within a structural causal model (SCM) [19]. Deep SCMs extend this to images by learning a generative mechanism that supports the full abduction–intervention–prediction cycle, enabling counterfactual inference at scale while preserving the semantics of interventions [18]. Recent theory further links counterfactual identifiability to dynamic optimal transport, showing that when the learned generator is equivalent to the time-1 map of a dynamic OT flow, the in-

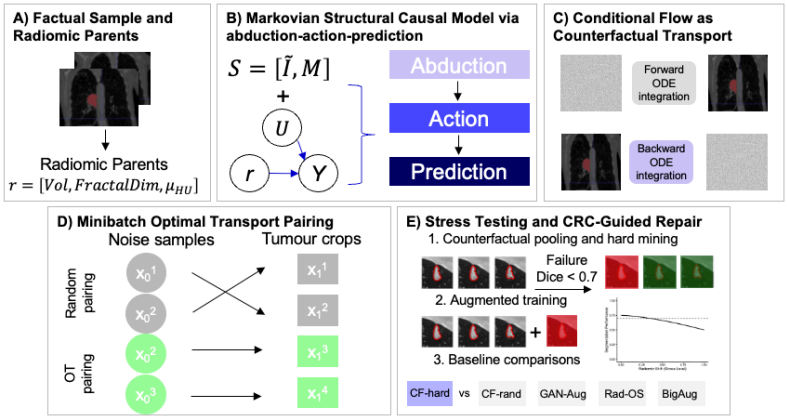

Fig. 1. Overview of our Counterfactual Stress-Testing Framework. (A) Tumour-centered CT crops and masks are summarized by a set of radiomic parents. **(B)** An SCM generates image–mask pairs from radiomic features (parents), enabling identity-preserving counterfactuals via abduction–intervention–prediction. **(C-D)** A conditional flow model with mini-batch optimal transport learns an invertible mapping between noise and data. **(E)** Decile-wise interventions on each parent produce CRCs; low-performing counterfactuals are used for targeted fine-tuning to improve robustness under controlled distribution shifts.

duced transport is vector-monotone and almost-everywhere bijective [21]. These properties make abduction unique, so reusing the same sampled exogenous noise across interventions yields identifiable, identity-preserving counterfactuals [21].

3 Methods

We study the robustness of a lung-cancer segmenter under controlled, clinically interpretable shifts. Given a tumour-centred CT crop I and mask M, the model f_θ predicts $M = f_\theta(I)$. We extract radiomic parents r with PyRadiomics [30] using a compact morphology/texture set: volume, sphericity, fractal dimension, mean Hounsfield unit (HU), HU standard deviation, and a first order wavelet entropy feature that captures fine-scale textural heterogeneity. We then generate identity-preserving counterfactual crops by intervening on one parent; keeping patient-specific factors fixed then evaluating how f_θ responds along these shifts. A summary of our method is shown in Figure 1.

3.1 Counterfactual image generation

We model paired outputs $S = [I, M]$ (normalized HU crop I and binary mask M) with a Markovian SCM of the form; $U \sim P_U$, $U \perp\!\!\!\perp r$, and $S = G_\phi(U, r)$, where U captures patient-specific exogenous factors (background anatomy, acquisition idiosyncrasies) and $r \in \mathbb{R}^K$ is the vector of radiomic parents.

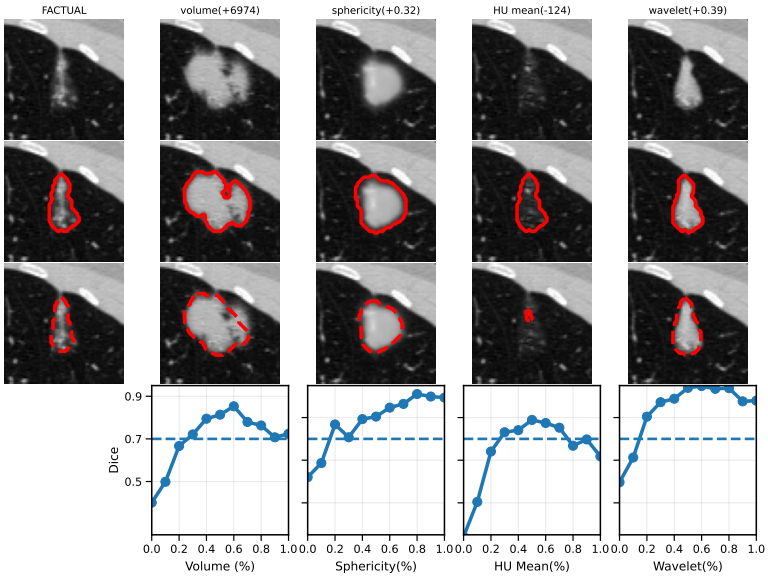

Fig. 2. Example counterfactual stress test for one case. The first column is the factual crop. Subsequent columns are counterfactuals generated by changing one radiomic parent at a time (header shows the parent and Δ from factual). The CT crop (top row), generator-derived counterfactual mask (second row; used as a pseudo-label), and the nnU-Net prediction (dashed red boundary, third row) are shown. Bottom panels plot the corresponding individual CRCs with a Dice = 0.7 reference line.

Interventions and counterfactuals. To intervene on a single parent we write $do(r_j = v)$ and form r^{cf} by copying r and setting its jth entry to v (all other entries unchanged). The counterfactual output is $S^{\mathrm{cf}} = G_\phi(U, r^{\mathrm{cf}})$, i.e., we keep the same exogenous noise U and change only the targeted parent value.

Abduction–intervention–prediction. Given a factual sample (S, r), we (i) *abduce* $\hat{U}$ by inverting the generator, $\hat{U} = G_\phi^{-1}(S, r)$, (ii) *intervene* by forming r^{cf}, and (iii) *predict* $S^{\mathrm{cf}} = G_\phi(\hat{U}, r^{\mathrm{cf}})$.

Conditional flow parameterization. We parameterize G_ϕ as the time-1 map of a conditional ODE:

$$\frac{\mathrm{d}S_t}{\mathrm{d}t} = v_\phi(S_t, t, r), \qquad t \in [0,1], \qquad S_0 = U, \qquad S_1 = S. \tag{1}$$

For fixed r, integrating Eq. (1) forward maps $S_0 \mapsto S_1$, while integrating the same ODE backward from S_1 recovers S_0 or equivalently U for abduction.

Flow-matching training. We train v_ϕ using flow matching [15,16]. Given paired endpoints (S_0, S_1), sample $t \sim \mathcal{U}[0,1]$ and set $S_t = (1-t)S_0 + tS_1$ with target displacement $W_t = S_1 - S_0$, which we regress the conditional velocity to:

$$\mathcal{L}(\phi) = \mathbb{E}\left[\|v_\phi(S_t, t, r) - W_t\|_2^2\right]. \tag{2}$$

Table 1. Counterfactual fidelity on the validation dataset (mean over cases and deciles). We report fidelity for the generated image (I) and mask (M). ρ is Spearman rank correlation (higher is better). MSE and drift are mean±SD (lower is better). Mask metrics are reported only for shape parents.

Feature	Image I			Mask M		
	$\rho\uparrow$	MSE↓	Drift↓	$\rho\uparrow$	MSE↓	Drift↓
Volume	0.91	0.15±0.07	0.17±0.06	0.93	0.10±0.05	0.17±0.05
Sphericity	0.92	0.17±0.08	0.19±0.07	0.94	0.13±0.06	0.18±0.06
Fractal dim.	0.85	0.29±0.12	0.26±0.09	0.88	0.22±0.10	0.20±0.07
Mean HU	0.90	0.11±0.05	0.12±0.04	–	–	–
HU std	0.82	0.23±0.10	0.19±0.07	–	–	–
Wavelet entropy	0.91	0.26±0.11	0.11±0.04	–	–	–

Minibatch OT couplings. Naïve flow matching pairs noise and data endpoints independently, which can encourage curved trajectories and make reverse-time inversion not monotonic. Instead, within each minibatch we compute an optimal-transport (OT) coupling between S_0 and data endpoints under a quadratic cost [29,20]. Let μ_0 and μ_1 denote the empirical minibatch distributions over (S_0, S_1). We solve

$$\pi^\star = \arg\min_{\pi \in \Pi(\mu_0,\mu_1)} \mathbb{E}_{(S_0,S_1)\sim\pi}\left[\|S_0 - S_1\|_2^2\right], \tag{3}$$

and sample paired endpoints $(S_0, S_1) \sim \pi^\star$ when forming the flow-matching targets in Eq. (2). This OT pairing encourages the learned ODE to follow a dynamic-OT style displacement path, yielding a counterfactual transport that is monotone and rank-preserving (and therefore invertible) under the conditions studied in [21]. Consequently, reverse-time integration corresponds to a unique abduction map leading to identifiable counterfactuals.

3.2 Counterfactual stress testing and CRC-guided repair

For each parent r_j, we choose intervention values from the deciles of its training distribution, $q \in \{0.1, \ldots, 1.0\}$. We evaluate the nnUNet on each counterfactual image and compute Dice against the corresponding generator mask pseudo-label.

Counterfactual robustness curves (CRCs) summarise robustness under these controlled interventional shifts. Due to the counterfactual identifiability of our model we can compute an individual level CRC plot from a single case, revealing patient-specific sensitivity. At the population level, we aggregate individual CRCs across cases to identify cohort-wide failure regimes and their severity.

For CRC-guided repair (CF-hard), we pool counterfactuals across parents and deciles and mark a sample as hard if the baseline model's Dice falls below a fixed threshold τ (we use $\tau = 0.7$). In our validation stress test (100 cases, 6 radiomic features, 10 deciles), we yielded 319 hard counterfactuals. We fine-tune nnU-Net on the original training cases augmented with these 319 samples. To isolate the effect of hard mining, we fine-tune on the same number of counterfac-

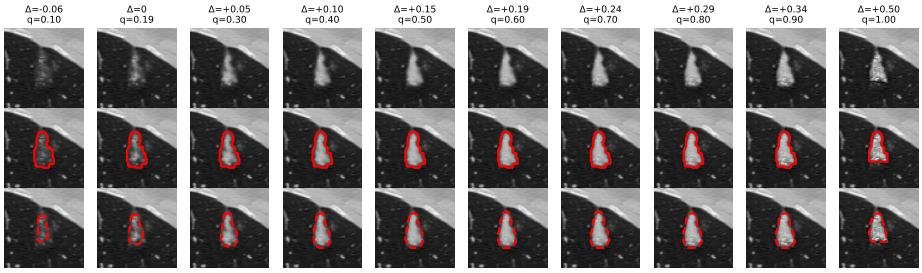

Fig. 3. Wavelet-entropy decile sweep around one case showing second column (factual) with a wavelet entropy value on the 0.19 quantile. Columns correspond to interventions at quantiles $q \in \{0.1, \ldots, 1.0\}$ of the wavelet-entropy distribution. The counterfactual images (top row), generator-derived masks (middle), and nnU-Net predictions (bottom; dashed red) illustrate sensitivity to texture variation.

Table 2. Stress-test summary by feature quintiles (0–20% to 80–100%). Cells report baseline Dice (mean ± 1 SD). Values with mean Dice < 0.7 (failure regions) highlighted.

Feature	Q1	Q2	Q3	Q4	Q5
Volume	**0.62**±0.12	0.74±0.09	0.81±0.06	0.84±0.05	0.86±0.04
Sphericity	**0.58**±0.13	0.71±0.10	0.80±0.07	0.83±0.06	0.84±0.05
Fractal dim.	0.84±0.05	0.82±0.06	0.78±0.08	**0.65**±0.11	**0.61**±0.12
Mean HU	0.85±0.06	0.77±0.09	0.80±0.08	**0.69**±0.12	**0.49**±0.16
HU std	0.78±0.08	0.81±0.07	0.80±0.07	0.73±0.10	**0.64**±0.12
Wavelet entropy	0.78±0.08	0.82±0.07	0.83±0.06	0.74±0.10	**0.67**±0.11

tuals (319) sampled uniformly from the full counterfactual pool. We term this process random counterfactual sampling (CF-rand) as an ablation experiment.

We compare CF-hard to three baselines. **BigAug** [31] applies a strong composition of intensity and geometric transformations. **GAN-Aug** [23] generates synthetic crops by sampling $U \sim P_U$ without abduction and conditioning on randomly sampled parent vectors. **Rad-OS** [6,8] performs radiomics-stratified oversampling by binning each parent into quintiles and sampling inversely to bin frequency. For fairness, each method uses the same additional sample budget as CF-hard (322 extra samples for training)

4 Experiments and Results

4.1 Implementation details

All models were implemented in PyTorch and trained on two NVIDIA A100 GPUs (40,GB). We train the segmentation model on 322 cases from the NSCLC-Radiogenomics dataset, with 100 cases retained for validation [2]. CRC stress testing and checkpoint selection are performed on this validation set, and final segmentation results are reported on 144 cases from the independent NSCLC-Radiomics dataset [1]. The validation set spans the tumour radiomic distribution

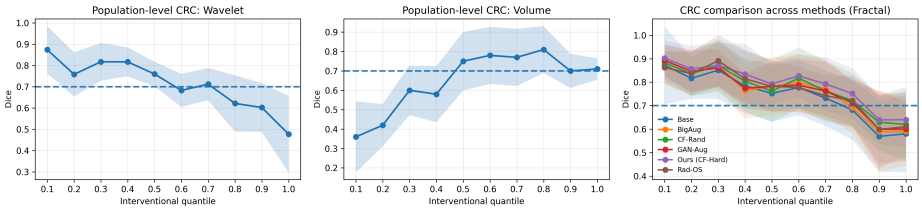

Fig. 4. Population-level CRCs on the validation dataset. Left/middle: cohort-aggregated CRCs (mean dice ± 1 SD) for wavelet entropy and volume interventions. Right: method comparison under a fractal-dimension intervention, showing how training strategies affect robustness across interventional quantiles.

used for CRC stress testing: volume 46–129,cm^3, sphericity 0.33–0.80, fractal dimension 2.38–2.78, mean HU −20 to 118, HU SD 35–199, and wavelet entropy 2.1–6.9. The counterfactual generator is trained on 1,472 in-house tumour-centred CT crops with expert masks. Segmentation uses 3D nnU-Net trained using stochastic gradient descent with polynomial LR decay (LR 10^{-2}, weight decay (WD) $3{\times}10^{-5}$, momentum 0.99, batch size 2, 1,000 epochs). We tuned LR, WD, and Dice threshold ($\tau \in \{0.65, 0.70, 0.75\}$), plus method-specific settings for BigAug and Rad-OS, selected by mean validation Dice. The generator is a BigGAN-style conditional 3D U-Net with self-attention [5], conditioned on parents via FiLM (Feature-wise Linear Modulation) and cross-attention. We train with AdamW (LR $2{\times}10^{-4}$, WD 10^{-4}), batch size 100, for 300,000 steps, and use a fixed-step Euler solver (50 steps) for sampling and reverse-time abduction.

4.2 Evaluation protocol and metrics.

For counterfactual fidelity, we re-extract radiomic parents on each counterfactual and report (i) MSE between intended and achieved feature values (z-scored within the training distribution), (ii) Spearman correlation across deciles, and (iii) non-target drift (mean absolute z-score change across non-intervened parents). For stress testing, we evaluate segmentation with Dice similarity coefficient and 95th percentile Hausdorff distance (HD95) on the independent test and summarise results as mean ± 1 SD across cases.

4.3 Counterfactual fidelity and mask plausibility

Table 1 reports counterfactual fidelity for both the generated image $\tilde{I}$ and mask M. For image-based interventions, the generator shows strong rank agreement between the target level and the feature re-measured on counterfactuals ($\rho = 0.82$–0.92), with low MSE and modest non-target drift. Fidelity is highest for volume, sphericity, mean HU and wavelet entropy, while fractal dimension and HU standard deviation are less controlled, consistent with boundary complexity and intra-tumour heterogeneity being harder to change without perturbing

Table 3. Independent test-set segmentation after training across different augmentation-based methods compared to the base nnU-Net. Dice is higher-is-better; HD95 is lower-is-better.

	Baseline(nnU-Net)	BigAug	GAN-Aug	Rad-OS	CF-Rand	CF-hard(Ours)
Dice (mean±SD)	0.77±0.11	0.81±0.09	0.80±0.09	0.78±0.10	0.81±0.11	**0.84±0.08**
HD95 (mm, mean±SD)	8.7±5.5	7.2±4.6	7.5±4.8	8.1±5.1	6.9±4.7	**5.6±3.8**

correlated radiomic features. The mask channel tracks volume, sphericity and fractal-dimension interventions with $\rho = 0.88$–0.94 and low MSE (0.10–0.22) at modest drift (0.17–0.20), including lower drift than the image channel for fractal dimension. The observed non-target drift should be interpreted as finite-sample and finite-capacity error, not perfect parent isolation. In particular, minibatch OT with batch size 100 cannot exactly recover the population dynamic-OT coupling, so leakage between correlated radiomic features is expected. To assess identity preservation outside the tumour, we computed background preservation on the overlapping non-tumour region after removing both factual and counterfactual masks. Background MSE was 0.03 ± 0.01 and background SSIM was 0.89, supporting that patient-specific context is largely preserved. Visually, counterfactual images and masks remain spatially coherent across intervention levels preserving background lung anatomy (Figures 2 and 3), supporting their use as pseudo-labels for CRC stress testing and CRC-guided repair.

4.4 Counterfactual Robustness Curves

Our CRCs probe nnU-Net robustness under controlled radiomic shifts, revealing failure modes at both the individual and population levels (Figures 2–4). For morphology, Dice improves with larger volume and degrades with reduced sphericity or more irregular boundaries (higher fractal dimension), matching the quintile trends in Table 2. Texture shifts shows Dice degradation for low-density, high-heterogeneity tumours even when shape is fixed, as clearly seen under mean HU where reduced HU yields faint lesions (Figure 2) and low Dice in the lowest quintiles (Table 2). Cohort CRCs highlight that method differences concentrate in the extreme quantiles where the baseline falls below Dice = 0.7, and CRC-guided repair consistently raises performance in these tail regimes (Figure 4).

Table 3 compares test-set Dice before and after retraining for each strategy. Augmentation and oversampling baselines yield modest gains, consistent with broadly distributed extra data. In contrast, CRC-guided repair (CF-hard) produces the largest improvement because added samples are concentrated in counterfactual regimes where the baseline fails, rather than spread across easy and hard phenotypes. This targeted sampling increases exposure to extreme intervention quantiles where CRCs drop sharply, improving test Dice. CF-hard improves boundary accuracy, reducing HD95 from 8.7 ± 5.5,mm to 5.6 ± 3.8,mm.

5 Discussion

We introduce an individual-level counterfactual stress-testing framework for lung tumour segmentation and showed how it can be used to repair brittle models. CRCs complement standard domain-shift evaluation by identifying which clinically interpretable tumour attributes drive degradation and the severities at which these failures emerge. Our counterfactuals preserve background anatomy, isolating the effect of the intervened radiomic parent. A key limitation is that counterfactual masks are pseudo-labels, so residual generator bias can propagate into repair. More broadly, we intervene on a fixed set of radiomic parents and generate counterfactuals from tumour-centred crops, which may miss failure sources that depend on wider context. The repair procedure also depends on the hardness threshold and the number of mined counterfactuals, which can affect the balance between improving difficult regimes and maintaining performance on easier cases. Future work will explore principled strategies for selecting and weighting counterfactuals during repair (e.g., curriculum sampling or threshold-free objectives) to reduce sensitivity to pseudo-label quality.

Acknowledgment

A. S. is supported by an NIHR clinical Lectureship. M.C. is funded by Medical Research Council (MRC) Clinician Scientist Fellowship (PB1626) and acknowledges support from the Image Research Team (ImRes) at Imperial College Healthcare Trust.

Disclosure of interests

No competing interests.

References

1. Aerts, H., Wee, L., Rios Velazquez, E., Leijenaar, R., Parmar, C., Grossmann, P., Carvalho, S., Bussink, J., Monshouwer, R., Haibe-Kains, B., et al.: Data from nsclc-radiomics (version 4)[data set]. The Cancer Imaging Archive **10**, K9 (2014)
2. Bakr, S., Gevaert, O., Echegaray, S., Ayers, K., Zhou, M., Shafiq, M., Zheng, H., Zhang, W., Leung, A., Kadoch, M., et al.: Data for nsclc radiogenomics (version 4)[data set]. The Cancer Imaging Archive (2017)
3. Boone, L., Biparva, M., Forooshani, P.M., Ramirez, J., Masellis, M., Bartha, R., Symons, S., Strother, S., Black, S.E., Heyn, C., Martel, A.L., Swartz, R.H., Goubran, M.: Rood-mri: Benchmarking the robustness of deep learning segmentation models to out-of-distribution and corrupted data in mri. NeuroImage **278** (9 2023)
4. Bray, F., Laversanne, M., Sung, H., Ferlay, J., Siegel, R.L., Soerjomataram, I., Jemal, A.: Global cancer statistics 2022: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. Cancer J Clin (2024)

5. Brock, A., Donahue, J., Simonyan, K.: Large scale gan training for high fidelity natural image synthesis. arXiv preprint arXiv:1809.11096 (2018)
6. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: Smote: synthetic minority over-sampling technique. Journal of artificial intelligence research **16**, 321–357 (2002)
7. Cheng, Z., Liu, M., Yan, C., Wang, S.: Dynamic domain generalization for medical image segmentation. Neural Networks (4 2025)
8. Demircioğlu, A.: The effect of data resampling methods in radiomics. Scientific Reports **14**(1), 2858 (2024)
9. Ibrahim, M., al Khalil, Y., Amirrajab, S., sun, C., Breeuwer, M., Pluim, J., Elen, B., Ertaylan, G., Dumontier, M.: Generative ai for synthetic data across multiple medeical modalities: A systematic review of recent developments and challenges. Computers in Biology and Medicine **189** (5 2025)
10. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods **18**(2), 203–211 (2021)
11. Isler, I.S., Mohaisen, D., Lisle, C., Turgut, D., Bagci, U.: Uncertainty-guided coarse-to-fine tumor segmentation with anatomy-aware post-processing (2025)
12. Khalil, Y.A., Amirrajab, S., Lorenz, C., Weese, J., Pluim, J., Breeuwer, M.: On the usability of synthetic data for improving robustness of deep learning-based segmentation of cardiac magnetic resonance images. Medical Image Analysis **84** (2023)
13. Kocak, B., Klontzas, M.E., Stanzione, A., Meddeb, A., Demircioğlu, A., Bluethgen, C., Bressem, K.K., Ugga, L., Díaz, N.M.O., Cuocolo, R.: Evaluation metrics in medical imaging ai: fundamentals, pitfalls, misapplications, and recommendations. European Journal of Radiology Artificial Intelligence (2025)
14. Lambin, P., Rios-Velazquez, E., Leijenaar, R., Carvalho, S., van Stiphout, R.G.P.M., Granton, P., Zegers, C.M.L., Gillies, R., Boellard, R., Aerts., A.D.H.J.W.L.: Radiomics: extracting more information from medical images using advanced feature analysis. European Journal of Cancer pp. 441–446 (2012)
15. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. In: The Eleventh International Conference on Learning Representations (2023)
16. Liu, X., et al.: Flow straight and fast: Learning to generate and transfer data with rectified flow. In: The Eleventh International Conference on Learning Representations (2023)
17. Olaf Ronneberger, Philipp Fischer, T.B.: U-net: Convolutional networks for biomedical image segmentation. Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015 **9351** (2015)
18. Pawlowski, N., Castro, D.C., Glocker, B.: Deep structural causal models for tractable counterfactual inference. Advances in Neural Information Processing Systems **33**, 857–869 (2020)
19. Pearl, J.: Causality. Cambridge university press (2009)
20. Pooladian, A.A., Ben-Hamu, H., Domingo-Enrich, C., Amos, B., Lipman, Y., Chen, R.T.: Multisample flow matching: Straightening flows with minibatch couplings. In: International Conference on Machine Learning. pp. 28100–28127. PMLR (2023)
21. Ribeiro, F.D.S., Santhirasekaram, A., Glocker, B.: Counterfactual identifiability via dynamic optimal transport. arXiv preprint arXiv:2510.08294 (2025)
22. Rostamzadeh, N., Hutchinson, B., Greer, C., Prabhakaran, V.: Thinking beyond distributions in testing machine learned models (2021)

23. Sandfort, V., Yan, K., Pickhardt, P.J., Summers, R.M.: Data augmentation using generative adversarial networks (cyclegan) to improve generalizability in ct segmentation tasks. Scientific reports **9**(1), 16884 (2019)
24. Santhirasekaram, A., Kori, A., Winkler, M., Rockall, A., Glocker, B.: Vector quantisation for robust segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 663–672. Springer (2022)
25. Santhirasekaram, A., Winkler, M., Rockall, A., Glocker, B.: A geometric approach to robust medical image segmentation. Medical Image Analysis **97**, 103260 (2024)
26. de Steenea, J.V., Linthouta, N., de Meyb, J., Vinh-Hunga, V., Claassensa, C., Noppenc, M., Bela, A., Stormea, G.: Definition of gross tumor volume in lung cancer: inter-observer variability. Radiotherapy and Oncology (2002)
27. Stroom, J.C., Heijmen, B.J.: Geometrical uncertainties, radiotherapy planning margins, and the icru-62 report. Radiotherapy and Oncology (2002)
28. Subbaswamy, A., Adams, R., Saria, S.: Evaluating model robustness and stability to dataset shift (2021)
29. Tong, A., Fatras, K., Malkin, N., Huguet, G., Zhang, Y., Rector-Brooks, J., Wolf, G., Bengio, Y.: Improving and generalizing flow-based generative models with minibatch optimal transport. Transactions on Machine Learning Research (2024)
30. Van Griethuysen, J.J., Fedorov, A., Parmar, C., Hosny, A., Aucoin, N., Narayan, V., Beets-Tan, R.G., Fillion-Robin, J.C., Pieper, S., Aerts, H.J.: Computational radiomics system to decode the radiographic phenotype. Cancer research **77**(21), e104–e107 (2017)
31. Zhang, L., Wang, X., Yang, D., Sanford, T., Harmon, S., Turkbey, B., Wood, B.J., Roth, H., Myronenko, A., Xu, D., et al.: Generalizing deep learning for medical image segmentation to unseen domains via deep stacked transformation. IEEE transactions on medical imaging **39**(7), 2531–2540 (2020)