False Confidence: Automated Labels Confound Fairness Audits in Cervical Spine Segmentation

Linus Juni*, Aasa Feragen, and Aditya Parikh

Section for Visual Computing, DTU Compute,
Technical University of Denmark, Kongens Lyngby, Denmark
{s225224, afhar, adipa}@dtu.dk

Abstract. Automated segmentation of cervical-spine MRI is increasingly used in clinical workflows, yet no fairness audit exists for this anatomy. We show that auditing these segmentation tasks is complicated by a common property of modern segmentation datasets: expert-annotated gold labels are expensive, so abundant machine-generated (silver) labels are added to limit annotation cost. This matters because the reference used to judge a model can itself be biased. In this study, we present the first fairness audit of cervical-spine MRI segmentation across sex, age, and race using the CSpineSeg dataset. We observe no significant demographic disparities for the deployed model, but the choice of reference label is not neutral. Because a dataset's silver labels are generated by a model trained on its gold labels, any new model trained on those same gold labels agrees more with the silver labels than with expert truth: scoring identical predictions against silver rather than gold overestimates performance by ∼8 Dice points and turns the significance verdict for age from non-significant to significant – not by the gap inflation Parikh et al. report (which we term *false magnitude*) but by collapsing within-group variance (which we term *false confidence*). Reference-label provenance is thus a first-order confounder in segmentation evaluation: performance and fairness should be reported against expert labels, and any fairness claim stated together with the provenance of its reference.

Keywords: Fairness · Segmentation · Spine MRI · Bias · Label Noise.

1 Introduction

Automated segmentation of the cervical spine – labelling the vertebral bodies and intervertebral discs on MRI helps clinicians grade stenosis and track degeneration over time [11]. A model that works well for some patient groups and produces incorrect segmentation for others would quietly widen existing gaps in patient care and surgical planning. Bias audits are increasingly required before clinical use, and are not optional [6]. Yet no fairness audit exists for this anatomy.

Auditing a segmentation model is harder than auditing a classification model. Its output is millions of voxels rather than a single decision, so fairness is condensed to a single scalar (Dice, HD95) [17] and the metric is itself biased by

* Corresponding author: s225224@dtu.dk

structure size when anatomy differs between groups [20]. Expert annotation is also time- and cost-intensive, so modern segmentation datasets increasingly mix a few expert (gold) labels with many machine-generated (silver) labels [21]. This carries its own problems, and a biased ruler can mislead an audit in two opposite ways. Parikh et al. [16] showed the first in breast MRI (a *biased ruler effect*): a silver reference biased against a group *exaggerates* the true gap, so a disparity looks larger than it is (*false magnitude*). Carrying their framework to a new anatomy, we find the opposite mode: a silver reference that is a near-copy of the model under audit agrees with it so tightly that score variation collapses, making a tiny, negligible gap look statistically certain (*false confidence*).

Fairness in segmentation is much less studied [18, 3]. The only prior spine fairness study, FairMedFM [10], covers the lumbar spine, sex only, and uses no volume correction; FairSeg [22] is in ophthalmology. None audit cervical-spine MRI across several protected attributes, and none ask whether the labels come from training or evaluation, and can, by themselves, carry bias.

Our Contributions. Building on the biased-ruler framework of Parikh et al. [16, 15], we contribute:

i **The first demographic fairness audit of cervical-spine MRI segmentation**, across sex, race, and age – on which we detect no significant demographic disparities by any standard metric.
ii **Evidence that machine-generated labels leak from the expert labels they derive from.** Scoring identical predictions against silver rather than expert labels overestimates performance by $\sim$8 Dice points, affecting *any* user who validates on such a dataset, not only fairness auditors.
iii **A complementary mode of the biased-ruler effect.** From that same leakage, our correlated ruler flips age from non-significant to significant by collapsing within-group variance (*false confidence*) – the opposite mechanism to Parikh et al.'s *false magnitude* [16].
iv **Practitioner guidance.** Report performance and fairness against expert labels, and treat a reference that scores both much higher and much more tightly than experts ($\sim$4$\times$ variance collapse) as too close to the model to be an independent benchmark.

2 Dataset and Exploratory Analysis

2.1 The CSpineSeg Cohort

We work with CSpineSeg [24], a public collection of 1,255 sagittal T2-weighted cervical-spine MRI exams from 1,232 patients at Duke University. We drop a single localizer series with anomalous dimensions, leaving a working set $\mathcal{D}_0$ of $N_0 = 1{,}254$ exams from 1,231 patients. The task is semantic segmentation: labelling every voxel of a scan as background, vertebral body, or intervertebral disc ($\mathcal{L} = \{0, 1, 2\}$). About 40% of the exams carry expert (gold) reference segmentations, drafted by one post-doctoral researcher without medical training and reviewed and revised by one of six board-certified radiologists; the rest were

Table 1. Composition of the working set $\mathcal{D}_0$ (N_0 = 1,254 exams, counted at the exam level). Percentages are of N_0; age is mean ± SD and the 13 missing-age exams (confirmed >89) are grouped into ≥60. The cohort is overwhelmingly non-Hispanic (92.7%); the 7.7% Other/unknown race is mostly Asian and unreported.

Characteristic		n (%)
Sex	Female	683 (54.5)
	Male	571 (45.5)
Race	White	809 (64.5)
	Black	349 (27.8)
	Other/unknown	96 (7.7)
Field strength	1.5 T	746 (59.5)
	3.0 T	508 (40.5)
Age, years (mean ± SD)		54.6 ± 16.3
Age group	<40	241 (19.2)
	40–60	486 (38.8)
	≥60	527 (42.0)
Manufacturer	Siemens	788 (62.8)
	GE	466 (37.2)

segmented automatically by the authors' own model, and these silver masks were never human-reviewed – though the mid-sagittal slice of each unannotated scan was checked for image quality.

We write the cohort as $\mathcal{D} = \{(X_i, Y_i, A_i, q_i)\}_{i=1}^{N}$: exam i pairs a 3D image X_i with a reference segmentation Y_i, an attribute value A_i over the protected attributes $\mathcal{A} = \{\text{sex}, \text{race}, \text{age}\}$, and a provenance flag $q_i \in \{\mathsf{g}, \mathsf{s}\}$ marking a gold (expert) or silver (machine-generated) label. We write $\mathcal{D}_{A=a}$ for the subgroup with $A = a$. The analysis cohort $\mathcal{D}$ (N = 1,142) is the sex-balanced subset of $\mathcal{D}_0$ defined in "Data, Splits, and Label Regimes".

Table 1 summarises $\mathcal{D}_0$: it leans slightly female, is predominantly White with a sizeable Black minority, overwhelmingly non-Hispanic, spans the adult age range, and splits across two manufacturers and two field strengths. The volumes are strongly anisotropic – high in-plane resolution (∼0.53 mm) but thick, few sagittal slices (∼4 mm, 12–25 per scan). These group sizes decide which axes we test formally – sex (F/M), race (White/Black), and age have enough exams per group for reliable inference – while the smaller race categories (8% combined) and ethnicity are too sparse and reported descriptively only.

2.2 Anatomical Confounders

A performance gap is biased only if anatomy does not explain it, and structure size is the main confound: men's vertebral bodies and discs are ∼23–24% larger, and the vertebral body also grows with age as cervical degeneration accrues [11] and differs by race [7], while the disc stays essentially constant – the least size-confounded cross-check. We therefore report vertebral body and disc separately alongside their macro average, and score with a volume-aware nDSC ("Metrics and statistical inference") beside raw Dice so a size effect is not read as bias. The size gap reflects anatomy, not scanner (GE vs. Siemens differ ∼2.15× in voxel count yet ∼6% in mm^3), and no attribute is entangled with the scanner; only

age is mildly entangled with race (White patients slightly older), so we control for age when comparing race.

3 Methodology

Our design carries the label-bias and biased-ruler framework of Parikh et al. [16, 15] from breast DCE-MRI over to cervical-spine segmentation, with one structural difference that shapes everything below: each CSpineSeg exam carries a gold or a silver label, never both.

3.1 Data, Splits, and Label Regimes

The provenance flag q_i partitions the cohort into disjoint gold and silver sets, $\mathcal{D}^{\mathrm{g}}$ and $\mathcal{D}^{\mathrm{s}}$ ($\mathcal{D}^{\mathrm{g}} \cap \mathcal{D}^{\mathrm{s}} = \varnothing$). We split the cohort at the patient level into training, validation, and test sets in a 70/10/20 ratio, stratified by the protected attributes $\mathcal{A}$; splitting by patient keeps all of a multi-exam patient's exams on one side, preventing leakage. We then balance the sexes – the working set leans female ("The CSpineSeg Cohort"), and an unequal base rate would conflate a true performance gap with mere representation – by randomly downsampling female exams within each split until the two are equal, dropping 112 exams ($N_0 = 1{,}254 \rightarrow N = 1{,}142$); this removes representation as an explanation rather than attempting to close the gap [16]. Because nnU-Net is cross-validation-native, we pool $\mathcal{D}_{\mathrm{train}}$ and $\mathcal{D}_{\mathrm{val}}$ into a single development set and hold out $\mathcal{D}_{\mathrm{test}}$.

We train two models that differ only in the labels they see, both sharing the architecture, recipe, patient-level split, and stratification: the deployment-realistic M_{mix} on the full split (798 exams: 318 gold, 480 silver) – combining scarce expert labels with abundant automated ones is the standard response to annotation cost [21] – and a gold-only M_{gold} (288 exams), which doubles as the generated-silver ruler for E2 (defined below, but in short, it never sees the test cases). Each is sex-balanced independently within its split.

3.2 Models, Metrics, and Experimental Design

Model and training. We segment every regime with nnU-Net v2 (ResEnc-L preset) [8, 9]: a strong equal-resource baseline in one fixed configuration shared across regimes, so every comparison is like-for-like and any gap reflects the data and labels, not a weak model. Per regime we train the standard 2D and 3D full-resolution configurations as five-fold cross-validations over the pooled development set and ensemble the ten folds (twenty total); the one audit-driven change is demographically stratified folds, with post-processing limited to a global largest-foreground cleanup. Training cost ∼1,105 GPU-hours (∼53–74 kg CO_2e) [13, 4].

Metrics and statistical inference. For class c on exam i, $\mathrm{Dice} = 2\,\mathrm{TP}/(2\,\mathrm{TP} + \mathrm{FP} + \mathrm{FN})$. Because its false-positive penalty is not scaled by structure size – the volume bias of "Anatomical Confounders" – we also report the normalised Dice

score (nDSC) [19], which rescales that term to the dataset-mean volume (recovering Dice when a class is already average-sized), and HD95 (mm; lower better) [20]; all three are reported per class and macro-averaged. For group disparities we use the binarized, rate-based metrics canonical in this literature [16, 2, 5]: an exam *succeeds* when its score clears τ (0.8 for Dice/nDSC, 5 mm for HD95), giving per-group success rates whose spread (DPD) and ratio (DIR, <0.8 flags adverse impact) we read off a τ-sweep. For significance we test the continuous scores – Mann–Whitney [14] (sex, race) and Kruskal–Wallis [12] (age) with rank-based effect sizes, FDR-corrected across the attribute×metric family [1] – backed by bootstrap DIR intervals, a joint OLS on $\mathcal{A}$, and permutation checks. These continuous tests are our *primary* instrument: at macro Dice ≈0.89 success rates sit near the ceiling and DIR is near-degenerate, so the rank tests carry more power.

Experimental design (E1–E2). Two experiments isolate where label provenance acts; each applies the full battery to every attribute in $\mathcal{A}$. **E1** asks whether the deployed model shows significant demographic disparities: it audits M_{mix} on the full 228-case test set against the dataset's own (mixed) labels. **E2** asks whether the *reference* distorts that verdict, on the 76 gold test cases scored against expert labels: it needs two references per image, but CSpineSeg gives only one, so we generate the second as M_{gold}'s predictions (it never saw these 76), mirroring how Zhou et al. produced the dataset's silver labels [24], and score M_{mix}'s *same* predictions against both the gold and this generated-silver ruler to isolate the pure ruler effect.

4 Experiments and Results

For every regime nnU-Net selected the softmax-averaged ensemble of the 2D and 3D full-resolution configurations, giving ten model-folds per regime (twenty total). We report E1 and E2 ("Experimental design (E1–E2)").

4.1 Global Fairness Audit (E1)

On the full 228-case test set, M_{mix} reaches a macro Dice of 0.947 (vertebral body 0.959, disc 0.935), a median macro HD95 of 0.34 mm, and a macro nDSC of 0.949. On the 76 gold-labelled cases – the only ones scored against expert references – macro Dice is 0.897, comparable to the 0.916 Zhou et al. [24] report on a different split with a smaller training pool.

We detect no significant demographic disparities by any measure: across sex, race, and age on all three metrics every DIR exceeds 0.96 and every BCa lower confidence bound clears the four-fifths threshold of 0.80. None of the 63 tests reach significance after FDR correction (smallest $p_{\text{fdr}} = 0.72$), effect sizes are negligible throughout ($|r_{rb}| \leq 0.14$, $\varepsilon^2 < 0.01$), and a joint OLS of macro Dice on sex, race, and age explains 1.3% of variance ($R^2 = 0.013$, $p = 0.43$).

4.2 The Biased Ruler: Performance Inflation and False Confidence (E2)

We now ask whether the choice of reference label can change a verdict. We take the same model, the same 76 gold-test images, and score them twice: against the expert (gold) labels, and against M_{gold}'s predictions on those images – the generated-silver ruler of "Experimental design (E1–E2)". The predictions are identical, so any change in the verdict is a pure ruler effect.

Performance overestimation from label leakage. Against gold, M_{mix} scores 0.897 macro Dice; against silver, the same predictions score 0.973 – an inflation of nearly 8 points. The silver ruler is M_{gold}'s output, and M_{gold} and M_{mix} were both trained, in part, on the *same* expert labels; two models fitted to overlapping data come to resemble each other more than either resembles the ground truth, so they share systematic errors and the silver ruler flatters them. This leakage is the dataset's, not an artefact of our setup: CSpineSeg's published silver labels were likewise generated by a model trained on its gold labels, so *any* user who validates against them overestimates performance, not just fairness auditors.

The verdict flip. Under the gold ruler, 0/63 tests are FDR-significant (smallest $p_{\text{fdr}} = 0.13$); under the silver ruler, 11/63 are, and every one is an age comparison (Table 2). The significance verdict reverses purely from the choice of reference label.

Mechanism: variance collapse, not magnitude. The age trend is *smaller* against silver: the worst-vs-best median spread falls from 2.7 Dice points (gold) to 0.6 (silver). What changes is noise – the within-group standard deviation drops from 0.058 to 0.014 ($\sim 4\times$) – so the Kruskal–Wallis H *rises* from 7.0 ($p_{\text{fdr}} = 0.27$) to 11.4 ($p_{\text{fdr}} = 0.026$). The rate-based DIR saturates at 1.0 on silver (zero failures above τ) even as the continuous test turns significant; a threshold sweep confirms the saturation. Figure 1 shows the same downward staircase on both rulers – gold's wide boxes bury the signal, silver's razor-thin boxes expose it. We interpret the two modes in "Discussion".

5 Discussion

Two modes of ruler bias. Our silver ruler is M_{gold}'s predictions, so scoring against it measures inter-model *agreement*, not accuracy. That agreement ($\sim$0.97 Dice, with the within-group SD collapsing $\sim 4\times$ from 0.058 to 0.014) far exceeds the $\sim$0.89 the same predictions reach against expert labels, and the cause is label leakage: M_{mix} and the M_{gold} ruler are both trained, in part, on the same expert labels, so two models fitted to overlapping data resemble each other more than either resembles the truth. Because the Kruskal–Wallis test's power scales with signal-to-noise, the same clinically negligible age gradient that gold's noise buries clears FDR correction once the leakage-driven variance collapse sharpens it: the ruler changed the test's power, not the patient. This complements Parikh et

Table 2. E2 biased ruler on the 76 gold-test images (macro Dice). The same M_{mix} predictions are scored against expert labels (gold) and M_{gold}'s predictions (silver). DIR/DPD are rate-based ($\tau = 0.8$); p_{fdr} is Kruskal–Wallis after FDR correction. The silver ruler saturates (DIR $\equiv 1.0$) yet manufactures significance on age.

	Gold ruler			Silver ruler		
Grouping	DIR	DPD	p_{fdr}	DIR	DPD	p_{fdr}
Sex	1.000	0.000	0.997	1.000	0.000	0.718
Race (W/B)	0.956	0.044	0.997	1.000	0.000	0.328
Age (3 bins)	0.952	0.048	0.270	1.000	0.000	0.026*

* FDR-significant at $\alpha = 0.05$. All 11 significant tests on the silver ruler are age groupings (three-bin and median split, Dice/nDSC).

al. [16], whose *independently biased* silver ruler inflates the between-group gap (DIR $0.871 \rightarrow 0.815$) – *false magnitude*; our *correlated* ruler instead shrinks the gap yet manufactures significance by removing noise (*false confidence*). Same practical hazard – **the verdict depends on the ruler**, opposite mechanism.

The age effect is intrinsic, not unfairness. The silver-ruler significance is an artefact, but the age *trend* is real: it runs the same direction on both rulers (60+ worst) and matches the degenerative anatomy of "Anatomical Confounders" – intrinsic difficulty, not a label artefact, and not unfairness either, the magnitudes being clinically negligible ($\leq$2.7 Dice points, every group > 0.89, all gold-ruler DIRs > 0.93).

Limitations. (i) a high-performance ceiling (all groups > 0.89 Dice) leaves little room for disparity to emerge; (ii) a single institution, anatomy, sequence, and architecture, with no external validation, and a small gold test set (76, $\sim$20 Black) that limits power for race; (iii) a correlated silver ruler – a near-clone of the model under test – so we can demonstrate only false confidence, not false magnitude; and (iv) our silver ruler is a *reconstruction* of the dataset's label-generation process (M_{gold}'s predictions, the original generator being unavailable), so E2 measures agreement between two related models rather than the influence of CSpineSeg's original silver labels directly; and the gold and silver pools – split by enrollment order, not at random – differ modestly in race and age, so any gold-vs-silver *training* comparison confounds provenance with composition (the E2 ruler result, on identical images, is immune); and (v) the gold labels are themselves unvalidated for inter-rater agreement (one annotator, one radiologist reviewer per case), so part of the silver ruler's variance collapse may reflect the absence of this human annotation noise, not leakage alone.

Recommendations. (i) **Report against expert labels.** On a mixed-provenance dataset, machine labels leak information from the expert labels they were derived

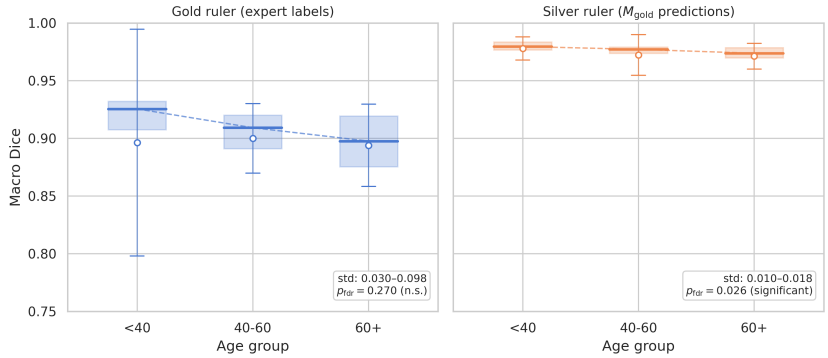

Fig. 1. The variance collapse behind the flip. Macro Dice by age bin for the same 76 predictions, scored against the gold ruler (left) and silver ruler (right). The gradient runs the same direction on both (60+ worst), but gold's per-group SDs (0.03–0.10) bury it below FDR correction while silver's collapsed variance (0.01–0.02) exposes it. The ruler changed the test's power, not the patient.

from, so validating on them overestimates performance (∼8 Dice points here); the only safe reference is the expert subset, which presupposes users are told which labels are which. (ii) **State where reference labels come from**: swapping expert for machine labels moved our verdict from 0/63 to 11/63 significant on identical predictions. (iii) **Check ruler–model similarity**: scores much higher *and* much tighter against the ruler than against experts (∼0.97 vs. ∼0.89, 4× variance collapse) signal a ruler too close to be an independent benchmark. (iv) **Report magnitude, not just significance**: the flagged age gap spans 0.6 Dice points where every group exceeds 0.97.

Future work. The clearest next step is to replace the correlated silver ruler with an independent one – either an externally trained segmenter (e.g. TotalSpineSeg [23]) or a deliberately biased generator – to test whether false magnitude emerges as ruler correlation decreases. A second direction is to formalise the variance-collapse diagnostic of recommendation (iii) into a general pre-audit check for over-correlated references. Finally, validating the effect across additional datasets, anatomies, and architectures – and, within CSpineSeg itself, on smaller and harder structures such as the spinal cord – would clarify how common false confidence is beyond this setting.

6 Conclusion

We presented the first demographic fairness audit of cervical-spine MRI segmentation: across sex, race, and age we detect no significant demographic disparities by any standard metric. The decisive finding, however, is methodological and reaches beyond this anatomy. Because a mixed-provenance dataset's automated

silver labels come from a model trained on its expert-gold labels, any new model trained on the same gold labels resembles the silver labels more than expert truth – so scoring against silver overestimates performance (∼8 Dice points here) and can flip a significance verdict by collapsing within-group variance (*false confidence*, the complement to Parikh et al.'s [16] *false magnitude*).

We therefore **urge the community** to treat reference-label provenance as a first-order, reportable variable: datasets should mark which labels are expert versus machine-generated, and every fairness claim should name its reference's provenance, be measured against expert labels where possible, and report a disparity's *magnitude*, not only its significance. As a **limitation**, our evidence is from a single institution, anatomy, and architecture in a high-performance regime; whether the same leakage distorts harder tasks remains open.

Code Availability

Code and models are available at https://github.com/linusjuni/false-confidence, along with setup and usage instructions.

Acknowledgments

The imaging and associated clinical data downloaded from MIDRC (The Medical Imaging and Data Resource Center) and used for research in this publication was made possible by the National Institute of Biomedical Imaging and Bioengineering (NIBIB) of the National Institutes of Health under contract 75N92020D00021 and through The Advanced Research Projects Agency for Health (ARPA-H). The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government.

Disclosure of Interests

The authors have no competing interests to declare that are relevant to the content of this article.

Bibliography

[1] Benjamini, Y., Hochberg, Y.: Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society: Series B (Methodological) **57**(1), 289–300 (1995). https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

[2] Bird, S., Dudík, M., Edgar, R., Horn, B., Lutz, R., Milan, V., Sameki, M., Wallach, H., Walker, K.: Fairlearn: A toolkit for assessing and improving fairness in AI. Tech. Rep. MSR-TR-2020-32, Microsoft (2020)

[3] Danaee, G., Niethammer, M., Rushmore, J., Bouix, S.: Investigating demographic bias in brain MRI segmentation: A comparative study of deep-learning and non-deep-learning methods. Machine Learning for Biomedical Imaging **3**, 792–808 (2025). https://doi.org/10.59275/j.melba.2025-d1g3

[4] Danish Energy Agency (Energistyrelsen): Key figures for Denmark 2023: CO2 emissions per kwh electricity. https://ens.dk/en/analyses-and-statistics/key-figures (2023), observed (physical) 99 g/kWh; adjusted (trade-corrected) 138 g/kWh. Accessed 2026-06-11

[5] Equal Employment Opportunity Commission, et al.: Uniform guidelines on employee selection procedures (1978). 29 C.F.R. § 1607.4(D) (the "four-fifths rule"); adopted jointly by the EEOC, Civil Service Commission, Department of Labor, and Department of Justice (1978), accessed 2026-06-12

[6] European Parliament and Council of the European Union: Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (AI Act). Official Journal of the European Union, L series, 12 July 2024 (2024)

[7] Ezra, D., Masharawi, Y., Salame, K., Slon, V., Alperovitch-Najenson, D., Hershkovitz, I.: Demographic aspects in cervical vertebral bodies' size and shape (C3-C7): a skeletal study. The Spine Journal **17**(1), 135–142 (2017). https://doi.org/10.1016/j.spinee.2016.08.022

[8] Isensee, F., Jaeger, P.F., Kohl, S.A.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). https://doi.org/10.1038/s41592-020-01008-z

[9] Isensee, F., Wald, T., Ulrich, C., Baumgartner, M., Roy, S., Maier-Hein, K.H., Jäger, P.F.: nnU-Net revisited: A call for rigorous validation in 3D medical image segmentation. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2024. Lecture Notes in Computer Science, vol. 15009, pp. 488–498. Springer Nature Switzerland, Cham (2024). https://doi.org/10.1007/978-3-031-72114-4_47

[10] Jin, R., Xu, Z., Zhong, Y., Yao, Q., Dou, Q., Zhou, S.K., Li, X.: FairMedFM: Fairness benchmarking for medical imaging foundation models. In: Advances in Neural Information Processing Systems (NeurIPS) – Datasets and Benchmarks Track (2024)

[11] Kato, F., Yukawa, Y., Suda, K., Yamagata, M., Ueta, T.: Normal morphology, age-related changes and abnormal findings of the cervical spine. Part II: magnetic resonance imaging of over 1,200 asymptomatic subjects. European Spine Journal **21**(8), 1499–1507 (2012). https://doi.org/10.1007/s00586-012-2176-4

[12] Kruskal, W.H., Wallis, W.A.: Use of ranks in one-criterion variance analysis. Journal of the American Statistical Association **47**(260), 583–621 (1952). https://doi.org/10.1080/01621459.1952.10483441

[13] Lacoste, A., Luccioni, A., Schmidt, V., Dandres, T.: Quantifying the carbon emissions of machine learning. arXiv preprint arXiv:1910.09700 (2019)

[14] Mann, H.B., Whitney, D.R.: On a test of whether one of two random variables is stochastically larger than the other. The Annals of Mathematical Statistics **18**(1), 50–60 (1947). https://doi.org/10.1214/aoms/1177730491

[15] Parikh, A., Das, S., Feragen, A.: Who does your algorithm fail? Investigating age and ethnic bias in the MAMA-MIA dataset. In: EurIPS 2025 Workshop on Medical Imaging (MedEurIPS) (2025)

[16] Parikh, A., Das, S., Feragen, A.: Investigating label bias and representational sources of age-related disparities in medical segmentation. In: 2026 IEEE 23rd International Symposium on Biomedical Imaging (ISBI). pp. 1–5 (2026). https://doi.org/10.1109/ISBI61048.2026.11515928

[17] Parikh, A., Frank, S., Das, S., Feragen, A.: Towards fairness under label bias in image segmentation: Impact, measurement and mitigation (2026), https://arxiv.org/abs/2605.06891

[18] Puyol-Antón, E., Ruijsink, B., Piechnik, S.K., Neubauer, S., Petersen, S.E., Razavi, R., King, A.P.: Fairness in cardiac MR image analysis: An investigation of bias due to data imbalance in deep learning based segmentation. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2021. Lecture Notes in Computer Science, Springer (2021). https://doi.org/10.1007/978-3-030-87199-4_39

[19] Raina, V., Molchanova, N., Graziani, M., Malinin, A., Müller, H., Bach Cuadra, M., Gales, M.: Tackling bias in the Dice similarity coefficient: Introducing nDSC for white matter lesion segmentation. In: 2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI). IEEE (2023). https://doi.org/10.1109/ISBI53787.2023.10230755

[20] Taha, A.A., Hanbury, A.: Metrics for evaluating 3d medical image segmentation: analysis, selection, and tool. BMC medical imaging **15**(1), 29 (2015). https://doi.org/10.1186/s12880-015-0068-x

[21] Tajbakhsh, N., Jeyaseelan, L., Li, Q., Chiang, J.N., Wu, Z., Ding, X.: Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation. Medical Image Analysis **63**, 101693 (2020). https://doi.org/10.1016/j.media.2020.101693

[22] Tian, Y., Shi, M., Luo, Y., Kouhana, A., Elze, T., Wang, M.: FairSeg: A large-scale medical image segmentation dataset for fairness learning using segment anything model with fair error-bound scaling. In: International Conference on Learning Representations (ICLR) (2024)

[23] Warszawer, Y., Molinier, N., Valošek, J., Benveniste, P.L., Bédard, S., Shirbint, E., Mohamed, F., Tsagkas, C., Kolind, S., Lynd, L., Oh, J., Prat, A., Tam, R., Traboulsee, A., Patten, S., Lee, L.E., Achiron, A., Cohen-Adad, J.: TotalSpineSeg: Robust spine segmentation with landmark-based labeling in MRI. ResearchGate preprint (2025), https://www.researchgate.net/publication/389881289_TotalSpineSeg_Robust_Spine_Segmentation_with_Landmark_Based_Labeling_in_MRI

[24] Zhou, L., Wiggins, W., Zhang, J., Colglazier, R., Willhite, J., Dixon, A., Malinzak, M., Gu, H., Mazurowski, M.A., Calabrese, E.: The Duke University Cervical Spine MRI Segmentation Dataset (CSpineSeg). Scientific Data **12**(1), 1695 (2025). https://doi.org/10.1038/s41597-025-05975-w