# Sex-based Bias Inherent in the Dice Similarity Coefficient: A Model Independent Analysis for Multiple Anatomical Structures

Hartmut Häntze<sup>1,2,3</sup>, Myrthe Buser<sup>2</sup>, Alessa Hering<sup>2</sup>, Lisa C. Adams<sup>3</sup>, and Keno K. Bressem<sup>3</sup>

<sup>1</sup> Charité - Universitätsmedizin Berlin, 12203 Berlin, Germany hartmut.haentze@charite.de

<sup>2</sup> Radboudumc, 6525 GA Nijmegen, Netherlands

Abstract. Overlap-based metrics such as the Dice Similarity Coefficient (DSC) penalize segmentation errors more heavily in smaller structures. As organ size differs by sex, this implies that a segmentation error of equal magnitude may result in lower DSCs in women due to their smaller average organ volumes compared to men. While previous work has examined sex-based differences in models or datasets, no study has yet investigated the potential bias introduced by the DSC itself. This study quantifies sex-based differences of the DSC and the normalized DSC in an idealized setting independent of specific models. We applied equally-sized synthetic errors to manual MRI annotations from 50 participants to ensure sex-based comparability. Even minimal errors (e.g., a 1 mm boundary shift) produced systematic DSC differences between sexes. For small structures, average DSC differences were around 0.03; for medium-sized structures around 0.01. Only large structures (i.e., lungs and liver) were mostly unaffected, with sex-based DSC differences close to zero. These findings underline that fairness studies using the DSC as an evaluation metric should not expect identical scores between men and women, as the metric itself introduces bias. A segmentation model may perform equally well across sexes in terms of error magnitude, even if observed DSC values suggest otherwise. Importantly, our work raises awareness of a previously underexplored source of sex-based differences in segmentation performance. One that arises not from model behavior, but from the metric itself. Recognizing this factor is essential for more accurate and fair evaluations in medical image analysis.

**Keywords:** Fairness · Segmentation · Dice Similarity Coefficient.

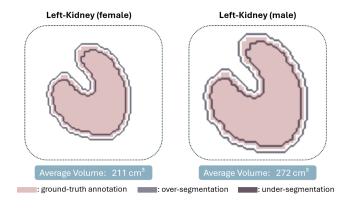
## 1 Introduction

The Dice Similarity Coefficient (DSC) is the most commonly used segmentation metric in medical imaging. Together with Intersection over Union it is the default recommendation of frameworks such as Metrics Reloaded [13]. Its pitfalls,

<sup>&</sup>lt;sup>3</sup> Klinikum rechts der Isar, Technical University of Munich, 81675 Munich, Germany

such as indifference to multiple classes [17] or size-dependent performance [6,17] are well reported in the literature and, dependent on the task, it can be advisable to pair it with other non-overlap metrics, such as Hausdorff Distance or Normalized Surface Distance [13]. Alternatively, volume-corrected metrics such as the normalized DSC (nDSC) have been proposed and demonstrate reduced bias in contexts with substantial target volume variation, such as white matter lesion segmentation [16], although they have yet to see widespread adoption.

The DSC has been widely used to evaluate the fairness of segmentation models. Typically, models are trained on mixed-sex datasets, and DSCs are calculated separately for male and female subgroups. A common assumption is that a fair model should yield comparable DSCs across sexes, which can be controlled by statistical tests. Some studies report higher DSCs for men [9,10], women [1], or equal outcomes for both sexes [3,12,14,15]. When a performance gap is observed, it can be attributed to unbalanced training data or anatomical differences that make one sex more difficult to segment. However, a third possibility is frequently overlooked: the DSC itself may introduce bias related to sex-specific organ size.



**Fig. 1.** Example annotation-mask of a left kidney with simulated over- and undersegmentation using a uniform 3 mm margin (grey and black outlines). In our cohort, female kidneys are on average smaller than male ones. Applying the same fixed error systematically across all kidneys in the cohort may result in different DSC values despite an identical error margin, as the increased size of male kidneys positively influences the outcome of the DSC calculation.

It is well documented that the DSC disproportionately penalizes errors in smaller structures [17], yet few discuss the implications this has on comparisons of subgroups with different organ volumes. Ferrante and Echeveste [7] caution against directly comparing DSC values between pediatric and adult patients for this reason and explicitly raise concerns about sex-based comparisons, given known volume differences between male and female organs [8]. Despite these known limitations of the DSC, their specific impact on sex-based analyses has not been systematically quantified. For example, Häntze et al. [9] reported average

DSC scores of 0.89 for men and 0.87 for women, but it remains unclear whether this reflects true model performance differences or biases inherent to the metric itself.

In this study, we explicitly quantify sex-related volume bias in the DSC. Unlike previous work, we do not evaluate the output of a specific segmentation model. Instead, to ensure comparability, we simulate uniform over- and under-segmentation by adding or removing a fixed voxel margin along structure boundaries (Fig. 1). This approach allows us to precisely control the error magnitude and apply identical modifications to subjects of both sexes. Consequently, any observed differences in evaluation metrics cannot be attributed to model behavior or training data composition but must stem from subject-specific characteristics such as sex and organ volume.

### 2 Methods and Materials

#### 2.1 Study Design

This study is a retrospective simulation. Explicit ethical consent was not required, as the analysis was performed on data from the German National Cohort [2,5] accessed through application number 836.

## 2.2 Dataset Description

We used a dataset of whole-body in-phase gradient echo sequences of 50 participants (25 male, 25 female) from the German National Cohort with similar age range and BMI (Table 1). The dataset includes annotations for 40 anatomical structures that were quality-controlled by two board-certified radiologists as described in [9]. We excluded left and right femur annotations, as these structures were not fully captured in all sequences. We further removed one annotation of the left kidney that we identified as a pathologic outlier. Additionally, six participants were missing a gallbladder, either due to cholecystectomy or annotation mistakes. A complete list of structures, along with their average volumes, is provided in Table 3. Based on these volumes we categorized the structures into three groups: small (volume < 100 cm³), medium (100 to 1,000 cm³) and large (volume > 1,000 cm³). We chose these round number thresholds as they are easy to interpret and to apply consistently.

**Table 1.** Summary statistics (median with interquartile range in parentheses).

Measure	Male (n=25)	Female (n=25)
		52.0 (13.0)
Weight (kg)	81.0 (17.0)	64.0 (18.0)
Height (m)	1.78(0.07)	1.64 (0.09)
BMI	25.69(5.05)	24.14 (5.42)

#### 2.3 Evaluation Metrics

We evaluated segmentation quality using the DSC, as it is the most frequently used metric in medical segmentation. Additionally, we used the nDSC due to its proposed advantages; that is robustness against large volumetric variance. The nDSC achieves this by adjusting the DSC by the mean fraction of the positive class for the target structure and group. In this study, this adjustment was applied separately for male and female subjects using the mean volume of each anatomical structure by sex. We did not test distance-based metrics such as normalized surface distance or Hausdorff distance. They are, by definition, not affected by volume, hence, errors that we induce with binary dilation will yield the same metrics.

## 2.4 Experiments

To assess how much the DSC and nDSC differentiate between men and women, we introduced controlled segmentation errors into our ground-truth masks. These synthetic errors are designed to resemble common failures of automated segmentation models, namely over-segmentation and under-segmentation. Because the DSC depends only on the overlap of two shapes, not on the spatial distribution of errors, we can simulate both under- and over-segmentation by applying binary dilation or erosion with a uniform margin around each reference structure, using the scikit-image [18] and Monai [4] python libraries. We resampled the sequences to a isotopic spacing of 1 mm and conducted two experiments: First, to simulate a very good model with almost perfect segmentations we added an error margin for dilation and erosion of 1 mm. Second, to simulate a good but not perfect model, we added an error margin of 3 mm. Note, that due to the isotopic spacing 1 mm equals exactly the width of one voxel. During simulation, we applied both dilation and erosion independently to the reference masks of each participant and structure. We then computed the evaluation metrics for both error types and averaged the results to obtain a single score per structure and participant. To prevent adjacent structures from merging during dilation, all organs were processed independently during both the error simulation and evaluation steps.

#### 2.5 Statistical Analysis

Since organ volumes differ between men and women, and DSC values are inherently linked to organ size, we expect to find corresponding DSC differences, as this relationship can be derived analytically. Consequently, this article's aim is not to only test whether differences exist, but rather to quantify how large these differences are. For this, we focused on the magnitude of the effects, reporting mean DSC and nDSC differences between male and female participants along with 95% confidence intervals. We performed this analysis for each anatomical structure, across the three volume-based groups, and for both error margins (1 and 3 mm). Within each volume group, we assessed statistical significance

using Welch's t-test and controlled the false discovery rate using the Benjamini-Hochberg correction.

## 3 Results

Applying the 1 mm error margin to the annotations resulted in average DSC values from 0.97  $\pm$  0.00 (both lungs, and liver), up to 0.72  $\pm$  0.10 (right adrenal gland). Table 3 shows the specific DSC values for each structure for this error margin. The 3 mm error margin further reduced the DSC values, which range from 0.92  $\pm$  0.01 up to 0.35  $\pm$  0.10, respectively. The volume corrected nDSC values were generally larger than their DSC counterparts; ranging from 0.98  $\pm$  0.00 (lungs and liver) to 0.83  $\pm$  0.07 (right adrenal gland) for 1 mm and from 0.95  $\pm$  0.00 (right lung) to 0.58  $\pm$  0.04 (left and right iliac arteries) for 3 mm.

The sex-stratified differences for the three volume categories are listed in Table 2. For the 1 mm error margin both sexes had average DSC differences of 0.00 for large, 0.01 for medium and 0.03 for small structures. The nDSC differences are 0.00, 0.00 and 0.04 respectively. The 3 mm margin further increased the differences for both metrics; all differences are significant with adjusted p < 0.001. The DSC value distributions by sex for the 1 mm margin are visualized in Fig. 2.

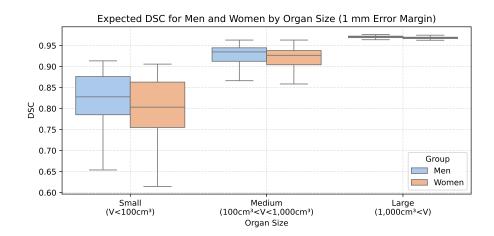
**Table 2.** Expected metric differences plus 95% confidence intervals for segmentations with an equal error margin between men and women. All differences are significant with adjusted p < 0.001.

Category	Error Margin	DSC	nDSC
Large			0.00 (0.00, 0.00)
			0.00 (0.00, 0.01)
Medium	1 mm	$0.01 \ (0.00, \ 0.01)$	0.01 (0.00, 0.01)
	3 mm	$0.02 \ (0.01, \ 0.03)$	0.02 (0.01, 0.02)
Small			0.02 (0.01, 0.02)
	3 mm	$0.06 \ (0.04, \ 0.08)$	0.04 (0.03, 0.05)

#### 4 Discussion

While multiple papers [17] and frameworks, such as metrics reloaded [13], mention the Dice Similarity Coefficient's (DSC) dependence on volume, none discuss the direct implication this has on sex: A subgroup with a larger average organ volume can be expected to have a higher DSC due to inherent bias in the metric itself, independent of model performance.

To quantify this bias we added synthetic segmentation errors to ground-truth annotations and calculated the resulting DSC. Our results demonstrate that even models with high overall accuracy and a small error margin of just 1 mm



**Fig. 2.** Box-plot that shows the Dice similarity Coefficient (DSC) for an error margin of 1 mm grouped by sex and organ size. Smaller structures are proportionally more affected by volume differences, which results in a higher average DSC difference between men and women. Average differences are 0.03 for small structures, 0.01 for medium structures and 0.00 for large structures.

can produce systematically different DSC values between men and women. For small anatomical structures, we observed an average DSC difference of 0.03; for medium-sized structures, the difference was around 0.01. Only large structures (i.e., lungs and liver) were mostly unaffected by this bias, with DSC differences close to zero (while this difference was statistically significant, we consider it to be of limited practical relevance). When simulating a larger error margin of 3 mm, these discrepancies became more pronounced, with differences reaching up to 0.06 for small structures. In contrast, the normalized DSC (nDSC) was less sensitive towards sex-specific volume differences: for the 1 mm error margin nDSC differences were close to zero for both large and medium structures and reduced to 0.02 for small structures, compared to the standard DSC.

The choice of an appropriate error margin is closely tied to image resolution. For example, with a slice thickness of 3 mm, a 1 mm segmentation error is not feasible due to voxel size constraints. In such cases, the 3 mm error simulations more accurately reflect the impact of a one-voxel segmentation error and are therefore more applicable. Notably, the average DSC values from our 1 mm error simulations closely align with those reported for real-world models in the literature. For instance, Kart et al. [11] reported DSC values of 0.97 (liver), 0.95 (spleen), 0.95 (left kidney), 0.95 (right kidney), and 0.87 (pancreas) on the same cohort. Our simulations yielded comparable values: 0.97, 0.94, 0.94, 0.94, and 0.89, respectively.

It is important to understand that our results do not indicate a flaw in the DSC's calculation itself. The metric behaves as intended: the relative impact

Table 3. Average volumes and DSC values for all structures Modifying the ground truth annotations with a uniform 1 mm error margin results in different Dice Similarity Coefficients (DSC) across structures. Column three reports the average DSC values ( $\pm$  standard deviation) across all participants. Column four shows the average DSC differences between men and women, along with the 95% confidence intervals. Volume categories (large, medium, small) are separated by dashed lines.

Class	Avg. Volume	DSC (1mm)	$\Delta \mathrm{DSC} \; (1\mathrm{mm})$
right lung	$1948 \text{ cm}^3$	$0.97 \pm 0.00$	0.00 [0.00, 0.00]
left lung	$1670   \mathrm{cm^3}$	$0.97 \pm 0.00$	0.00 [0.00, 0.00]
liver	$1575   \mathrm{cm^3}$	$0.97 \pm 0.00$	0.00 [0.00, 0.00]
small bowel	$723 \text{ cm}^3$	$-0.92 \pm 0.01$	$0.01 \ [0.00, 0.01]$
colon	$650 \; {\rm cm^3}$	$0.92 \pm 0.01$	0.00 [-0.01, 0.01]
heart	$622 \text{ cm}^3$	$0.96 \pm 0.00$	0.00 [0.00, 0.01]
right gluteus maximus	$568 \text{ cm}^3$	$0.95 \pm 0.01$	0.01 [0.01, 0.01]
left gluteus maximus	$527   {\rm cm^3}$	$0.95 \pm 0.01$	0.01 [0.00, 0.01]
left autochthonous muscle	$418 \text{ cm}^3$	$0.94 \pm 0.01$	0.01 [0.01, 0.01]
right autochthonous muscle		$0.94 \pm 0.01$	0.01 [0.00, 0.01]
stomach	$350   \mathrm{cm^3}$	$0.94 \pm 0.01$	0.01 [0.00, 0.01]
spleen	$321   {\rm cm^3}$	$0.94 \pm 0.01$	0.00 [0.00, 0.01]
right kidney	$275 \text{ cm}^3$	$0.94 \pm 0.01$	0.00 [0.00, 0.01]
right hip	$262 \text{ cm}^3$	$0.90 \pm 0.01$	0.01 [0.01, 0.01]
left hip	$261 \text{ cm}^3$	$0.90 \pm 0.01$	0.01 [0.00, 0.01]
left iliopsoas muscle	$257   \mathrm{cm^3}$	$0.92 \pm 0.01$	0.02 [0.01, 0.02]
left kidney	$241 \text{ cm}^3$	$0.94 \pm 0.01$	0.01 [0.00, 0.01]
right iliopsoas muscle	$241   \text{cm}^3$	$0.92 \pm 0.01$	0.02 [0.01, 0.02]
right gluteus medius	$227 \text{ cm}^3$	$0.94 \pm 0.01$	0.01 [0.01, 0.01]
spine	$221 \text{ cm}^3$	$0.89 \pm 0.01$	0.01 [0.01, 0.01]
aorta	$221 \text{ cm}^3$	$0.91 \pm 0.01$	0.01 [0.01, 0.01]
left gluteus medius	$211   \text{cm}^3$	$0.94 \pm 0.01$	0.01 [0.00, 0.01]
urinary bladder	$191   \mathrm{cm^3}$	$0.94 \pm 0.01$	0.00 [-0.01, 0.01]
sacrum	$170   \mathrm{cm^3}$	$0.91 \pm 0.01$	0.01 [0.00, 0.01]
pancreas	$130 \text{ cm}^3$	$0.89 \pm 0.01$	0.01 [0.01, 0.02]
gallbladder	$123 \text{ cm}^3$	$0.85 \pm 0.05$	0.01 [-0.02, 0.04]
inferior vena cava	$91 \text{ cm}^3$	$0.88 \pm 0.01$	$0.01 \ [0.01, 0.02]$
right adrenal gland	$86 \text{ cm}^3$	$0.72 \pm 0.10$	0.04 [-0.02, 0.10]
left adrenal gland	$84 \text{ cm}^3$	$0.75 \pm 0.06$	0.05 [0.02, 0.08]
left iliac vena	$63 \text{ cm}^3$	$0.83 \pm 0.02$	0.02 [0.02, 0.03]
right iliac vena	$63 \text{ cm}^3$	$0.82 \pm 0.02$	0.03 [0.02, 0.04]
duodenum	$63 \text{ cm}^3$	$0.86 \pm 0.03$	0.03 [0.01, 0.04]
right gluteus minimus	$57 \text{ cm}^3$	$0.89 \pm 0.01$	0.01 [0.00, 0.01]
esophagus	$55 \text{ cm}^3$	$0.81 \pm 0.02$	0.02 [0.01, 0.03]
left gluteus minimus	$54 \text{ cm}^3$	$0.89 \pm 0.01$	0.01 [0.00, 0.01]
left iliac artery	$54 \text{ cm}^3$	$0.76 \pm 0.03$	0.04 [0.03, 0.06]
portal vein and splenic vein	$50 \text{ cm}^3$	$0.78\pm0.02$	0.01 [0.00, 0.03]
right iliac artery	$44 \text{ cm}^3$	$0.75\pm0.04$	0.05 [0.03, 0.06]

of a constant-sized error increases as the size of the target structure decreases; and the DSC correctly captures this. However, interpreting such differences as evidence of model unfairness can be misleading, as the absolute error may be identical for both groups. Even a segmentation model that operates in a sexneutral manner can yield different DSC values between men and women due to inherent anatomical volume differences.

The findings of this study are expected to generalize to other 3D imaging modalities such as CT, as our analysis was based solely on annotations and did not incorporate MRI-specific features. For 2D modalities like X-ray, results are likely to differ as the dimensionality shift from three to two alters the voxel count per structure, and therefore also the calculation of the DSC.

This paper has limitations. Real-world segmentation errors are highly heterogeneous and can include not only over- and under-segmentation but also false positives and complete omissions of structures. By simulating errors through binary dilation and erosion, we simplify this complexity and inevitably lose some similarity to real model behavior. However, the purpose of this study is not to reproduce all error types but to quantify potential DSC differences between men and women under idealized conditions. Further, it should be noted that our analysis was conducted on a cohort of German participants, and results may not generalize to populations with different anatomical or sex-based characteristics. Hence, the outcomes presented here should be interpreted as a rough estimate of the potential magnitude of sex-related differences of the DSC. To support comparability, we reported the average organ volumes for our cohort alongside the corresponding metrics (Table 3).

# 5 Conclusion

In this study, we quantified the expected DSC differences between men and women under the assumption of an ideal non-discriminatory segmentation model. Observed DSC differences vary in magnitude from 0.00 to 0.06. This shows that differences in DSC values between sexes do not necessarily indicate model bias. Rather, these discrepancies may be due to a volume-dependent and thus sexdependent bias intrinsic to the DSC itself, particularly when evaluating small structures below 100 cm<sup>3</sup>. Importantly, our results do not aim to replace the DSC, as multiple studies have clearly demonstrated its value in highlighting both negligible [3,12,14,15] and substantial [1,10] sex-based differences. However, we show that systematic differences can arise even under optimal conditions. Consequently, when a segmentation model appears to underperform for one sex, especially for women, it is worth considering whether the metric itself contributes to the observed disparity. Simulations like those presented in this article can help isolate metric-induced effects, and comparing results with alternative metrics such as distance-based metrics or the nDSC [16] may offer a more complete assessment of fairness.

### References

- Afzal, M.M., Khan, M.O., Mirza, S.: Towards equitable kidney tumor segmentation: bias evaluation and mitigation. In: Machine Learning for Health (ML4H). pp. 13–26. PMLR (2023), https://proceedings.mlr.press/v225/afzal23a.html
- 2. Bamberg, F., Kauczor, H.U., Weckbach, S., Schlett, C.L., Forsting, M., Ladd, S.C., Greiser, K.H., Weber, M.A., Schulz-Menger, J., Niendorf, T., et al.: Wholebody MR imaging in the german national cohort: rationale, design, and technical background. Radiology 277(1), 206–220 (2015). https://doi.org/10.1148/radiol. 2015142272
- de Boer, S., Häntze, H., Venkadesh, K.V., Buser, M.A., Mamani, G.E.H., Xu, L., Adams, L.C., Nawabi, J., Bressem, K.K., van Ginneken, B., et al.: Robust kidney abnormality segmentation: A validation study of an AI-based framework. arXiv preprint arXiv:2505.07573 (2025). https://doi.org/10.48550/arXiv.2505.07573
- Cardoso, M.J., Li, W., Brown, R., Ma, N., Kerfoot, E., Wang, Y., Murrey, B., Myronenko, A., Zhao, C., Yang, D., et al.: Monai: An open-source framework for deep learning in healthcare. arXiv preprint arXiv:2211.02701 (2022). https://doi.org/10.48550/arXiv.2211.02701
- Consortium, G.N.C.G.: The german national cohort: aims, study design and organization. Eur. J. Epidemiol. 29(5), 371–382 (2014). https://doi.org/10.1007/s10654-014-9890-7
- Dice, L.R.: Measures of the amount of ecologic association between species. Ecology 26(3), 297–302 (1945). https://doi.org/10.2307/1932409
- Ferrante, E., Echeveste, R.: Open challenges on fairness of artificial intelligence in medical imaging applications. In: Trustworthy AI in Medical Imaging, pp. 265–276. Elsevier (2025). https://doi.org/10.1016/B978-0-44-323761-4.00023-7
- 8. Geraghty, E.M., Boone, J.M., McGahan, J.P., Jain, K.: Normal organ volume assessment from abdominal CT. Abdom. Imaging  $\bf 29$ , 482–490 (2004). https://doi.org/10.1007/s00261-003-0139-2
- Häntze, H., Xu, L., Mertens, C.J., Dorfner, F.J., Donle, L., Busch, F., Kader, A., Ziegelmayer, S., Bayerl, N., Navab, N., et al.: MRSegmentator: Multi-modality segmentation of 40 classes in MRI and CT. arXiv preprint arXiv:2405.06463 (2024). https://doi.org/10.48550/arXiv.2405.06463
- Ioannou, S., Chockler, H., Hammers, A., King, A.P., Initiative, A.D.N.: A study of demographic bias in CNN-based brain MR segmentation. In: International Workshop on machine learning in clinical neuroimaging. pp. 13–22. Springer (2022). https://doi.org/10.1007/978-3-031-17899-3 2
- Kart, T., Fischer, M., Küstner, T., Hepp, T., Bamberg, F., Winzeck, S., Glocker, B., Rueckert, D., Gatidis, S.: Deep learning-based automated abdominal organ segmentation in the UK biobank and german national cohort magnetic resonance imaging studies. Invest. Radiol. 56(6), 401–408 (2021). https://doi.org/10.1097/ RLI.00000000000000755
- 12. Lee, T., Puyol-Antón, E., Ruijsink, B., Shi, M., King, A.P.: A systematic study of race and sex bias in CNN-based cardiac MR segmentation. In: International Workshop on Statistical Atlases and Computational Models of the Heart. pp. 233–244. Springer (2022). https://doi.org/10.1007/978-3-031-23443-9 22
- Maier-Hein, L., Reinke, A., Godau, P., Tizabi, M.D., Buettner, F., Christodoulou, E., Glocker, B., Isensee, F., Kleesiek, J., Kozubek, M., et al.: Metrics reloaded: recommendations for image analysis validation. Nat. Methods 21(2), 195–212 (2024). https://doi.org/10.1038/s41592-023-02151-z

- Pettit, R.W., Marlatt, B.B., Corr, S.J., Havelka, J., Rana, A.: nnU-Net deep learning method for segmenting parenchyma and determining liver volume from computed tomography images. Ann. Surg. Open 3(2), e155 (2022). https://doi.org/10.1097/AS9.000000000000155
- Puyol-Antón, E., Ruijsink, B., Mariscal Harana, J., Piechnik, S.K., Neubauer, S., Petersen, S.E., Razavi, R., Chowienczyk, P., King, A.P.: Fairness in cardiac magnetic resonance imaging: assessing sex and racial bias in deep learning-based segmentation. Front. Cardiovasc. Med. 9, 859310 (2022). https://doi.org/10.3389/fcvm.2022.859310
- Raina, V., Molchanova, N., Graziani, M., Malinin, A., Muller, H., Cuadra, M.B., 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). pp. 1–5. IEEE (2023). https://doi.org/10.1109/ ISBI53787.2023.10230755
- Reinke, A., Tizabi, M.D., Sudre, C.H., Eisenmann, M., Rädsch, T., Baumgartner, M., Acion, L., Antonelli, M., Arbel, T., Bakas, S., et al.: Common limitations of image processing metrics: A picture story. arXiv preprint arXiv:2104.05642 (2021). https://doi.org/10.48550/arXiv.2104.05642
- Van der Walt, S., Schönberger, J.L., Nunez-Iglesias, J., Boulogne, F., Warner, J.D., Yager, N., Gouillart, E., Yu, T.: scikit-image: image processing in python. PeerJ 2, e453 (2014). https://doi.org/10.7717/peerj.453

Acknowledgments. This project was conducted with data from the German National Cohort (NAKO) (www.nako.de). The NAKO is funded by the Federal Ministry of Education and Research (BMBF) [project funding reference numbers: 01ER1301A/B/C, 01ER1511D and 01ER1801A/B/C/D], federal states of Germany and the Helmholtz Association, the participating universities and the institutes of the Leibniz Association. We thank all participants who took part in the NAKO study and the staff of this research initiative. Much of the computation resources required for this research was performed on computational hardware generously provided by the Charité HPC cluster. Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Health and Digital Executive Agency (HADEA). Neither the European Union nor the granting authority can be held responsible for them.



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