A Multi-Site Label Trap in Shape-Based Endometriosis Detection: What Female Pelvic-Organ Shape Does (and Does Not) Encode

Gijs Luijten[1,2,*][0009−0000−3404−1917], Merlin Engelke[1][0000−0002−8463−8786], Richard Ruppel[3][0009−0002−7081−0300], and Jan Egger[1,2,4,5][0000−0002−5225−1982]

[1] Institute for AI in Medicine (IKIM), University Medicine Essen, Hufelandstr. 55, 45147 Essen, Germany
[2] Faculty of Computer Science, University of Duisburg-Essen, Schützenbahn 70, 45127 Essen, Germany
[3] Department of Radiology, Charité – Universitätsmedizin Berlin, Charitéplatz 1, 10117 Berlin, Germany
[4] Cancer Research Center Cologne Essen (CCCE), West German Cancer Center Essen (WTZ), 45122 Essen, Germany
[5] Medical Faculty, University of Duisburg-Essen, Hufelandstraße 55, 45122 Essen, Germany
*Corresponding author: gijs.luijten@uk-essen.de

Abstract. We analyze the 3D shape of female pelvic organs in two public MRI datasets, using only the segmentation masks, so the remaining signal is purely morphological. Our central finding is a multi-site label trap. On UT-EndoMRI (124 patients from two sites), ovarian and uterine shape predicts whether an endometrioma annotation is present within one site (D2, 15% positive, ROC-AUC 0.826 ± 0.073). However, the prevalence of this annotation differs sharply between sites (78% vs. 15%), so pooled cross-validation is misleading (AUC 0.638), and a model trained at one site fails at the other (AUC 0.274, n=73). Two further results support this caution. First, on UMD (300 women), uterine-corpus shape carries a weak age signal (R^2=0.14, MAE 8.9 years) that stems from organ size rather than scale-free shape, while fibroid shape does not predict age at all ($R^2 = -0.31$). Second, scale-free shape descriptors vary less across raters (annotators) than size descriptors (mean coefficient of variation, uterus 0.117 vs. 0.142; ovary 0.147 vs. 0.228). We recommend that shape-based studies in women's health imaging report label prevalence per site, evaluate cross-site transfer alongside pooled metrics, and quantify the rater robustness of their descriptors. Code is publicly available.

Keywords: Women's health · Pelvic MRI · Statistical shape analysis · Uterus · Endometriosis · Multi-site confounding · Inter-rater reliability.

1 Introduction

Pelvic MRI is the main imaging modality for characterizing uterine fibroids, mapping endometriosis, and planning gynecologic surgery [2]. The organs it depicts are usually assessed visually or through simple size measurements; their 3D shape is rarely analyzed as a quantitative phenotype of its own. We ask three questions using two public datasets that ship 3D segmentations. First, does uterine morphology encode age, and does the signal come from the uterus itself or from its fibroids? Second, can ovarian and uterine shape predict which patients carry an endometrioma annotation, and does such a model survive a change of site? Third, because every shape phenotype rests on a segmentation, how reproducible are shape descriptors across human raters (annotators)?

Along the way we surface a pitfall that is easy to miss and damaging to reproducibility: a clinically plausible label that is confounded with site, producing pooled performance that does not transfer. Multi-site confounding is well documented in deep-learning radiology [6,14,18], but shape-based studies often skip the check because shapes feel "objective" once the vendor-specific intensity texture is discarded. We show that it bites just as hard when the labeling practice, rather than the image appearance, differs across sites.

This is a women's-health application of an intensity-free shape pipeline (segmentation $\rightarrow$ surface point cloud $\rightarrow$ shape descriptors): image intensity is discarded after segmentation, so the remaining signal is purely morphological. The methods are deliberately standard; the contribution is empirical.

2 Related Work

Statistical shape models (SSMs) underpin medical morphometry [7,3,10]; point-cloud encoders analyze anatomical shape [13,16]. Shape encodes demographics in pelvis, brain, and mandible [4,5,8]. Medical AI models can encode site information and exhibit performance differences across groups and sites [18,6,14], motivating careful multi-site evaluation and reliability reporting, both of which are central here.

Most shape-based demographic studies focus on a single organ (e.g., brain [5], mandible [8], cardiac chambers [1,11]), and they do not always make explicit whether the reported signal stems from organ size or from scale-free shape.

Female reproductive organs have received little attention from the correspondence-free shape-analysis community, despite the clinical relevance of uterine and ovarian morphology. We address this gap.

3 Materials and Methods

3.1 Datasets

UMD [12] provides T2-weighted uterine MRI with multi-label segmentations (uterine wall, cavity, fibroid, cysts) for 300 women; age (21–86 years, mean 49,

standard deviation 13.0) is read from the DICOM `PatientAge` tag. The median voxel spacing is $0.5 \times 0.5 \times 6.1$ mm. We define the uterine corpus by merging the uterine wall and cavity labels (segmentation labels $\{1, 2\}$) into a single mask; fibroids (label $\{3\}$) are kept separate. Not every subject was segmented: a usable uterine-corpus mask exists for 266 of the 300 subjects and a fibroid mask for 264.

UT-EndoMRI [9] provides multi-sequence pelvic MRI for 124 endometriosis patients from two institutions (D1, n=51; D2, n=73), with segmentations of the uterus, ovaries, and, where annotated, endometriomas; the median voxel spacing is $0.7 \times 0.7 \times 6.0$ mm. Three independent raters segmented the D1 cases. Where several raters segmented the same organ, the prediction experiments use one fixed annotation per subject, from the highest rater index (r3 before r2 before r1), so the choice is deterministic; the rater comparison in Sec. 4.3 uses all of them. We define the endometrioma label as the presence of an endometrioma segmentation file for a subject. This label reflects annotation availability, not surgically confirmed pathology, and we phrase all results accordingly. Its prevalence differs sharply between sites: 78% at D1 (40 of 51, leaving 11 negatives) versus 15% at D2 (11 of 73). Again, not every organ was segmented: a usable uterus mask exists for 111 of the 124 subjects and an ovary mask for 101.

3.2 Shape Extraction Pipeline

For each structure, we load the NIfTI segmentation mask, apply marching cubes in physical (mm) coordinates to obtain a watertight triangular mesh, and uniformly sample a fixed-size surface point cloud of N=2,048 points. Working in physical coordinates absorbs voxel-spacing differences between subjects and sites. Masks with fewer than 50 (UMD) or 30 (UT-EndoMRI) voxels are also treated as absent, and missing organs enter the models as missing values; see per-organ counts in Sec. 3.1. No subject was excluded for touching the field-of-view boundary, as in pelvic MRI the organ of interest is the scan target.

3.3 Shape Descriptors: Size vs. Shape

From each point cloud we compute correspondence-free descriptors (no point-to-point matching between subjects is required) separated into two families:

- **Size:** centroid size (the square root of the summed squared distances to the centroid), volume (mm^3), surface area (mm^2), axis-aligned bounding-box extents sorted from largest to smallest ($x \geq y \geq z$), and bounding-box diagonal.
- **Shape (scale-free):** covariance-eigenvalue elongation (λ_2/λ_1, λ_3/λ_1), flatness, anisotropy, sphericity, aspect ratios (y/x, z/x), and normalized surface-area-to-volume ratio.

Put simply, size descriptors grow with the organ, while shape descriptors capture its form regardless of how large it is: points are centroid-subtracted and normalized to unit centroid size before computing shape descriptors, so these

are invariant to translation and isotropic scaling. The bounding box is axis-aligned in patient coordinates, not fitted to the organ, so its extents mix organ size with organ orientation; for an organ as variably oriented as the uterus, this adds noise, which the inter-rater analysis in Sec. 4.3 partly captures.

3.4 Prediction Models

Gradient-boosted trees (XGBoost; 300 trees for classification and 400 for regression, maximum depth 4, learning rate 0.05, row and column subsampling 0.8) predict age and endometrioma annotation from the descriptors of Sec. 3.3. Age regression reports the mean absolute error (MAE) and the coefficient of determination (R^2); classification reports the area under the receiver operating characteristic curve (ROC-AUC), the area under the precision-recall curve (PR-AUC, whose random-ranker baseline equals the prevalence), and balanced accuracy (the mean of sensitivity and specificity). All experiments use subject-level 5-fold cross-validation, stratified for classification, with per-fold class rebalancing via inverse-prevalence weighting; we report mean $\pm$ standard deviation across folds and, for the key classification results, bootstrap 95% confidence intervals over pooled out-of-fold predictions (2,000 resamples). For the endometrioma task only uterus and ovary descriptors enter the model; the endometrioma segmentations define the label and are never used as features, so the label cannot leak into the input. The size-vs-shape decomposition runs three models: size features only, shape features only, and both combined. What follows says what these descriptors and this model can detect, not what pelvic shape contains; a negative result means no signal was found, not that none exists.

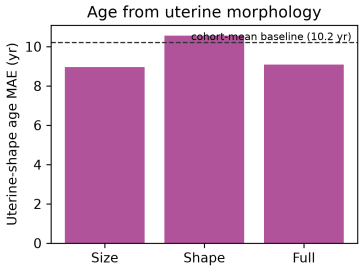
Multi-site evaluation. For UT-EndoMRI, we run (i) pooled 5-fold CV, (ii) within-site CV at each site separately, and (iii) cross-site transfer (train on D1, test on D2). Both within-site evaluations rest on roughly a dozen minority-class subjects (11 annotation-negative cases at D1, 11 annotation-positive cases at D2), so we treat both with caution and attach confidence intervals rather than relying on fold-level spread alone.

3.5 Inter-Rater Robustness

Shape descriptors inherit variability from the segmentations they are computed on. For D1 subjects segmented by ≥ 2 raters we therefore compute each descriptor once per rater and summarize the disagreement as the coefficient of variation (CoV) across raters: the standard deviation divided by the absolute mean. A lower CoV means the descriptor changes less when a different rater segments the same organ; this measures the stability of the descriptors, not agreement between readers on a diagnosis. We report the CoV averaged within the size and shape families and, because family averages can hide descriptor-level differences, also per descriptor. The uterus has 25 multi-rater subjects and the ovary 18.

Table 1. Age from uterine morphology (UMD, 300 women, 5-fold CV). Size dominates; fibroids carry no age signal. Right: age MAE by feature regime; the dashed line marks the cohort-mean baseline (10.2 yr).

Features	Age MAE (yr)	Age R^2
Size only	9.0 ± 0.5	0.08 ± 0.09
Shape only	10.6 ± 0.7	-0.10 ± 0.11
Size + shape	9.1 ± 0.8	0.09 ± 0.13
Uterine corpus shape	8.9 ± 0.5	0.14 ± 0.09
Fibroid shape	11.5 ± 0.7	-0.31 ± 0.10



4 Results

4.1 Uterine Corpus Shape Encodes Age Weakly, and It Is Not the Fibroids

Uterine-corpus shape predicts age with MAE 8.9 years and $R^2 = 0.14$ (Table 1). To put these numbers in context: always predicting the cohort mean age gives MAE 10.2 years, so the model removes about 13% of the baseline error and explains 14% of the age variance, with substantial fold-to-fold spread (Table 1). The association is real but weak: a covariate to control for, not a usable age predictor.

Two controls matter. First, the signal is size-dominated: size features alone reach MAE 9.0 years ($R^2 = 0.08$), while scale-free shape alone performs worse than the cohort-mean baseline ($R^2 = -0.10$). Adding shape to size does not improve prediction ($R^2 = 0.09$ versus 0.08; the difference lies within fold-level standard deviations), so the age–uterus relationship is primarily volumetric. Second, fibroid shape also performs worse than the baseline ($R^2 = -0.31$), so the weak age signal resides in the uterine corpus, not in the myomas. This is consistent with uterine involution: the corpus shrinks after menopause, a process distinct from fibroid growth.

4.2 Shape Predicts the Endometrioma Annotation Within One Site, but the Label Is a Site Trap

Within site D2 (15% annotation-positive), ovarian and uterine shape predicts the endometrioma annotation well (ROC-AUC 0.826 ± 0.073 across folds; Table 2). The uncertainty analysis supports this: the bootstrap 95% CI spans 0.67–0.94, a permutation test makes a lucky ranking unlikely ($p = 0.004$), and the PR-AUC of 0.569 sits well above the random-ranker baseline of 0.15 (balanced accuracy 0.74 ± 0.12). The trap is that this within-site result does not survive pooling or transfer. The annotation prevalence is 78% at D1 versus 15% at D2, so a pooled classifier conflates site with pathology (AUC 0.638) and a model trained on D1 is

Table 2. Endometrioma-annotation prediction from pelvic-organ shape (UT-EndoMRI). Prevalence: D1 78%, D2 15%.

Setting	n	ROC-AUC	PR-AUC	Bal. acc.
Pooled (5-fold CV)	124	0.638 ± 0.036	0.556	0.55 ± 0.02
Within site D2 (15% pos.)	73	0.826 ± 0.073	0.569	0.74 ± 0.12
Within site D1 (78% pos.)	51	0.458 ± 0.211	0.805	0.49 ± 0.07
Cross-site D1→D2	73	0.274	0.112	0.38

uninformative on D2 (AUC 0.274, n=73); its transfer CI (0.14–0.43) lies entirely below 0.5: the D1 model ranks D2 patients in the wrong direction, consistent with having learned the annotation protocol rather than the pathology.

The within-site evaluation at D1 gives AUC 0.458 ± 0.211 (bootstrap 95% CI 0.29–0.67). Its PR-AUC of 0.805 in Table 2 looks high only in isolation: it barely exceeds D1's own prevalence baseline of 0.78, whereas D2's 0.569 stands well above its baseline of 0.15. This does not refute the shape signal; it fails to test it. With 78% positive cases, D1 contributes only 11 annotation-negative subjects, the mirror image of D2's 11 positives, so both within-site estimates rest on roughly a dozen minority-class cases and their intervals are wide. The difference is that D2's interval lies entirely above 0.5, the AUC of random ranking, while D1's straddles it. The lesson: pooled cross-validation hides a site-confounded label; per-site prevalence and cross-site transfer expose it.

4.3 Inter-Rater Shape Robustness

Because shape phenotypes inherit segmentation variability, we quantify this variability using the three D1 raters (Table 3); the CoV expresses the across-rater spread as a fraction of the descriptor's value. Averaged within families, scale-free shape descriptors vary less across raters than size descriptors for both organs (mean CoV, uterus 0.117 versus 0.142; ovary 0.147 versus 0.228). The medians are lower than the means, so a few subjects with strong rater disagreement inflate the family averages; for the uterus the ordering even reverses at the median (shape 0.088 versus size 0.074). The advantage of the shape family therefore comes from avoiding the occasional extreme disagreement that hits volume and surface area, not from being uniformly more stable. At the descriptor level the families overlap: centroid size is the single most rater-stable descriptor for the uterus (median CoV 0.055), while surface area and the λ_3/λ_1 elongation vary most (0.147 and 0.141). The uterus is more robust than the ovary, consistent with ovaries being small and often deformed in endometriosis. No clinically meaningful CoV threshold has been established, so these values rank descriptors rather than certify them and motivate ICC and Bland–Altman analyses in future work. The practical implication is a trade-off: pipelines deployed across institutions should prefer descriptors with demonstrated rater stability, but the age signal in Sec. 4.1 is size-dominated, while volume and surface area are among the least rater-stable descriptors.

Table 3. Inter-rater coefficient of variation of descriptors (D1, ≥ 2 raters). On family averages, shape varies less across raters than size.

Organ	Feature family	subjects	Mean CoV	Median CoV
Ovary	shape	18	0.147	0.085
Ovary	size	18	0.228	0.112
Uterus	shape	25	0.117	0.088
Uterus	size	25	0.142	0.074

5 Discussion

Size-dominated, uterine age signal. Uterine morphology carries a weak age signal that is genuinely uterine (fibroid shape does not predict age), but the signal is mostly size-driven, so "shape-based" pelvic biomarkers should always be reported with a size control. The size dominance is unsurprising: uterine volume changes substantially across the reproductive lifespan, while the shape of the corpus is comparatively stable. Studies that report shape-based age biomarkers without separating size from shape may overstate the signal's geometric content.

The multi-site label trap. The endometrioma result is the paper's main caution. Shape predicts whether an endometrioma was annotated within one site (AUC 0.826 at D2), but the label follows each site's annotation habits, so pooled numbers flatter the model and do not transfer. Models are known to inherit such human labeling bias [14,18]. Multi-site women's health shape studies should report label prevalence per site and cross-site transfer, not only pooled cross-validation.

This trap generalizes beyond endometriosis and beyond shape: any multi-site dataset where annotation completeness correlates with the label of interest could produce misleading pooled metrics. D1 produced an endometrioma segmentation for 78% of its patients and D2 for 15%; the model cannot distinguish "this patient has an endometrioma" from "this patient comes from the site that segments them more often."

Within a consistent annotation protocol, the same within-site model could assist labeling or triage. Our descriptors do not yet support reliable cross-site pseudolabeling, but as cohorts grow, within-site models may become practical tools for generating annotations where clinical labels are absent, the application that originally motivated this study.

Privacy. By analogy with prior work on the brain [15], pelvic organ shape may carry patient-specific information. Re-identification in the strict sense requires an attack scenario: an adversary holding a reference shape of a known patient to link records against. We test no such attack, and our descriptors carry limited discriminative signal; the more plausible risk is unintended attribute extraction (e.g., an age range) from shared segmentations. Data curators releasing dense point clouds or organ meshes should weigh both risks.

Limitations. Cohorts are modest (300 for UMD, 124 for UT-EndoMRI). The within-site D2 result rests on 11 annotation-positive cases, so its bootstrap

95% CI is wide (0.67–0.94) and 0.826 should be read as indicative rather than precise. The endometrioma label reflects annotation availability, not confirmed surgical pathology. The age analysis comes from a uterine-myoma cohort: fibroid burden, menopausal status, parity, hormonal status, and the clinical indication for imaging may all confound the age signal, none are released with the dataset, and the cohort need not represent the general population. The inter-rater analysis covers only 25/18 D1 subjects (uterus/ovary), and rater variability at D2 is unknown, so we cannot confirm that the robustness pattern holds across sites; future pelvic MRI datasets should include multi-rater segmentations. The analyses are exploratory and motivate confirmatory work. A site-prediction experiment (classifying site from shape alone) would quantify how much site information the descriptors encode; we leave it to future work.

The study is MRI-based by necessity: female reproductive organs are not covered by widely used public CT segmentation tools such as TotalSegmentator [17], and the available gynecologic CT cohorts are small and heterogeneous. Extending shape phenotyping of these structures to CT is future work.

Correspondence-based SSMs and deep point-cloud encoders (PointNet [13], DGCNN [16]) remain future work; the correspondence-free descriptors are interpretable and robust but may miss subtle non-linear shape variation. A richer representation would not dissolve the label trap: more capacity makes site easier to encode.

6 Conclusion

Applying an intensity-free shape pipeline to female pelvic organs, we find a weak, size-dominated age signal that resides in the uterine corpus rather than the fibroids; we show that shape predicts the endometrioma annotation within one site while pooled and cross-site evaluations mislead; and we show that scale-free descriptors resist rater disagreement better than size ones. The label trap is the broadest contribution: it threatens any study in which annotation completeness varies across sites, whether or not shape is involved.

Code and Data Availability. The full pipeline is publicly available at `https://github.com/TIO-IKIM/shape-demographics-capi` and reproduces every number in this paper from a single entry point. Both datasets are publicly released and de-identified by their original providers; no new patient data were collected.

Acknowledgments. This research was supported by the REACT-EU funded project KITE (Grant No. EFRE-2920801977, Plattform für KI-Translation Essen, `https://kite.ikim.nrw/`) and by the German Federal Ministry of Research, Technology and Space (BMFTR) within the Network of University Medicine, third funding phase (NUM 3.0, Grant No. 01KX2524), through the RACOON project. Large language models assisted with the literature search, code scaffolding, and manuscript drafting. The study design was developed by the authors, and the first author reviewed all material.

Disclosure of Interests. The authors have no competing interests.

References

1. Bai, W., Suzuki, H., Huang, J., Francis, C., Wang, S., Tarroni, G., Guitton, F., Aung, N., Fung, K., Petersen, S.E., Piechnik, S.K., Neubauer, S., Evangelou, E., Dehghan, A., O'Regan, D.P., Wilkins, M.R., Guo, Y., Matthews, P.M., Rueckert, D.: A population-based phenome-wide association study of cardiac and aortic structure and function. Nature Medicine **26**(10), 1654–1662 (2020). https://doi.org/10.1038/s41591-020-1009-y
2. Bazot, M., Bharwani, N., Huchon, C., Kinkel, K., Cunha, T.M., Guerra, A., Manganaro, L., Buñesch, L., Kido, A., Togashi, K., Thomassin-Naggara, I., Rockall, A.G.: European society of urogenital radiology (ESUR) guidelines: MR imaging of pelvic endometriosis. European Radiology **27**(7), 2765–2775 (2017). https://doi.org/10.1007/s00330-016-4673-z
3. Cates, J., Elhabian, S.Y., Whitaker, R.T.: ShapeWorks: Particle-based shape correspondence and visualization software. In: Statistical Shape and Deformation Analysis, pp. 257–298. Elsevier (2017). https://doi.org/10.1016/B978-0-12-810493-4.00012-2
4. Coelho, J.d., Curate, F.: CADOES: An interactive machine-learning approach for sex estimation with the pelvis. Forensic Science International **302**, 109873 (2019). https://doi.org/10.1016/j.forsciint.2019.109873
5. Cole, J.H., Franke, K.: Predicting Age Using Neuroimaging: Innovative Brain Ageing Biomarkers. Trends in Neurosciences **40**(12), 681–690 (2017). https://doi.org/10.1016/j.tins.2017.10.001
6. Gichoya, J.W., Banerjee, I., Bhimireddy, A.R., et al.: AI recognition of patient race in medical imaging: a modelling study. The Lancet Digital Health **4**(6), e406–e414 (2022). https://doi.org/10.1016/S2589-7500(22)00063-2
7. Heimann, T., Meinzer, H.P.: Statistical shape models for 3D medical image segmentation: A review. Medical Image Analysis **13**(4), 543–563 (2009). https://doi.org/10.1016/j.media.2009.05.004
8. Klop, C., Schreurs, R., De Jong, G.A., Klinkenberg, E.T.M., Vespasiano, V., Rood, N.L., Niehe, V.G., Soerdjbalie-Maikoe, V., Van Goethem, A., De Bakker, B.S., Maal, T.J.J., Nolte, J.W., Becking, A.G.: An open-source, three-dimensional growth model of the mandible. Computers in Biology and Medicine **175**, 108455 (2024). https://doi.org/10.1016/j.compbiomed.2024.108455
9. Liang, X., Alpuing Radilla, L.A., Khalaj, K., Dawoodally, H., Mokashi, C., Guan, X., Roberts, K.E., Sheth, S.A., Tammisetti, V.S., Giancardo, L.: A multimodal pelvic MRI dataset for deep learning-based pelvic organ segmentation in endometriosis. Scientific Data **12**, 1292 (2025). https://doi.org/10.1038/s41597-025-05623-3
10. Mitteröcker, P., Gunz, P.: Advances in geometric morphometrics. Evolutionary Biology **36**(2), 235–247 (2009). https://doi.org/10.1007/s11692-009-9055-x
11. Moscoloni, B., Beeche, C., Chirinos, J.A., Segers, P., Peirlinck, M.: Unveiling sex dimorphism in the healthy cardiac anatomy. The Journal of Physiology (2025). https://doi.org/10.1113/JP288667
12. Pan, H., Chen, M., Bai, W., Li, B., Zhao, X., Zhang, M., Zhang, D., Li, Y., Wang, H., Geng, H., Kong, W., Yin, C., Han, L., Lan, J., Zhao, T.: Large-scale uterine myoma MRI dataset covering all FIGO types with pixel-level annotations. Scientific Data **11**, 410 (2024). https://doi.org/10.1038/s41597-024-03170-x
13. Qi, C.R., Su, H., Mo, K., Guibas, L.J.: PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation. In: IEEE Conference on Computer Vision

and Pattern Recognition (CVPR). pp. 77–85 (2017). https://doi.org/10.1109/CVPR.2017.16
14. Seyyed-Kalantari, L., Zhang, G., McDermott, M.B.A., Chen, I.Y., Ghassemi, M.: Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. Nature Medicine **27**, 2176–2182 (2021). https://doi.org/10.1038/s41591-021-01595-0
15. Wachinger, C., Golland, P., Kremen, W., Fischl, B., Reuter, M.: BrainPrint: A discriminative characterization of brain morphology. NeuroImage **109**, 232–248 (2015). https://doi.org/10.1016/j.neuroimage.2015.01.032
16. Wang, Y., Sun, Y., Liu, Z., Sarma, S.E., Bronstein, M.M., Solomon, J.M.: Dynamic Graph CNN for Learning on Point Clouds. ACM Transactions on Graphics **38**(5), 1–12 (2019). https://doi.org/10.1145/3326362
17. Wasserthal, J., Breit, H.C., Meyer, M.T., Pradella, M., Hinck, D., Sauter, A.W., Heye, T., Boll, D.T., Cyriac, J., Yang, S., Bach, M., Segeroth, M.: TotalSegmentator: Robust Segmentation of 104 Anatomic Structures in CT Images. Radiology: Artificial Intelligence **5**(5), e230024 (2023). https://doi.org/10.1148/ryai.230024
18. Zech, J.R., Badgeley, M.A., Liu, M., Costa, A.B., Titano, J.J., Oermann, E.K.: Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. PLOS Medicine **15**(11), e1002683 (2018). https://doi.org/10.1371/journal.pmed.1002683