This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

What Does Anatomical Shape Know About You? A Multi-Organ, Shape-Only Study of Demographic Prediction, Attribution, and Privacy

Gijs Luijten[1,2,*][0009−0000−3404−1917], Behrus Hinrichs-Puladi[3,4][0000−0001−5909−6105], Merlin Engelke[1][0000−0002−8463−8786], Christian Wachinger[5,6][0000−0002−3652−1874], and Jan Egger[1,2,7,8][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 Oral and Maxillofacial Surgery, University Hospital RWTH Aachen, Pauwelsstraße 30, 52074, Aachen, Germany
[4] Institute of Medical Informatics, University Hospital RWTH Aachen, Pauwelsstraße 30, 52074, Aachen, Germany
[5] Lab for Artificial Intelligence in Medical Imaging, Institute for Diagnostic and Interventional Radiology, School of Medicine and Health, TUM Klinikum, Technical University of Munich (TUM), Ismaninger Straße 22, 81675 Munich, Germany
[6] Munich Center for Machine Learning (MCML), Walther-von-Dyck-Straße 10, 85748 Garching bei München, Germany
[7] Cancer Research Center Cologne Essen (CCCE), West German Cancer Center Essen (WTZ), 45122 Essen, Germany
[8] Medical Faculty, University of Duisburg-Essen, Hufelandstraße 55, 45122 Essen, Germany
*Corresponding author: gijs.luijten@uk-essen.de

Abstract. How much does the bare shape of internal organs reveal about a person once all image intensity is removed? Using the public TotalSegmentator CT and MRI datasets (1228 whole-body CTs), we convert the provided segmentations of 16 anatomical structures into surface point clouds and study, in one controlled setting, demographic prediction, per-organ attribution, size versus scale-free shape, shape versus intensity-based features, cross-site and cross-modality generalization, and privacy. Within this single cohort, shape alone predicts sex at AUC 0.866 (bootstrap 95% CI 0.846–0.886) and age within 9.3 years; the hip bones are the strongest sex cue (AUC 0.940). Shape outperforms a Hounsfield intensity baseline for sex (0.868 vs 0.729), so against this baseline the sex signal is primarily geometric rather than a density artifact. This signal persists under leave-one-institution-out evaluation (AUC 0.827) and replicates on MRI (AUC 0.762); a classifier trained on CT shapes transfers to MRI without retraining (AUC 0.708). Much of the sex signal is organ size, yet scale-free shape alone still reaches AUC 0.791. The

same signal has a constructive side: as a calibrated auto-labeling tool it sexes the most confident half of subjects at 94% accuracy. It also carries privacy risk. A single organ's quantized shape signature makes 93% of subjects unique within the cohort, rising with more organs; this distinctiveness also depends on how finely the descriptors are binned (18% unique at 3 bins, 99% at 20). We report these as upper bounds on linkability, not re-identification rates. Even so, "de-identified" organ meshes remain highly distinctive and disclose demographics. We release the full pipeline at `https://github.com/TIO-IKIM/shape-mi-demographics`.

Keywords: Statistical shape analysis · Point clouds · Demographic prediction · Privacy · Linkability · Computational anatomy.

1 Introduction

Anatomical shape is a rich, intensity-independent descriptor of the human body and the central object of computational anatomy. Tools such as TotalSegmentator [28] produce organ segmentations at scale, and meshes and point clouds derived from such tools are increasingly shared as "de-identified" surrogates for the raw images [15]. This practice raises a basic question that has not been studied systematically across organs: how much about a person is recoverable from organ shape alone? Shape-based biomarkers are useful in the clinic and in forensics. If shape also encodes demographics and identity, shared shapes are less anonymous than they appear.

Individual threads of this question are well studied. Forensic anthropology estimates sex and age from skeletal shape, and brain imaging predicts age [6]. Deep networks recover self-reported race from raw medical images [8], a finding that triggered debate about what such models learn. What is missing is a controlled, multi-organ, shape-only study: one pipeline and one dataset in which we measure per-organ signal, separate size from shape, test cross-site robustness, and quantify privacy risk. We provide exactly that.

Our contributions are:

1. A reproducible pipeline that turns TotalSegmentator segmentations into surface point clouds with field-of-view truncation control, and predicts sex and age from shape alone across 16 organs.
2. A per-organ attribution of demographic signal that rediscovers the well-known sexual dimorphism of the pelvis without anatomical priors, and extends the ranking to soft organs.
3. A decomposition into size features and scale-free shape features.
4. Cross-site and cross-modality generalization: leave-one-institution-out, and replication/transfer to MRI with the identical pipeline.
5. A privacy analysis showing that quantized (binned) organ shape is a near-unique within-cohort signature, plus an auto-labeling use of the same signal.

We restrict prediction targets to sex and age, which are biologically grounded. We treat race and ethnicity as an ethical concern only, not as a prediction target.

2 Related Work

Shape models in medical imaging: Statistical shape models (SSMs) have long underpinned segmentation and morphometry [12]. Geometric deep learning networks such as PointNet [20], PointNet++ [21], and DGCNN [27] provide permutation-invariant encoders (their output is unchanged by reordering of the points) that operate directly on point clouds, without point correspondence.

Shape as a demographic biomarker: Brain imaging underpins "brain-age" models that estimate the biological age of the brain and predict health outcomes [6,7], and cardiac shape shows pronounced sexual dimorphism [18]. Forensic anthropology and geometric morphometrics estimate sex and age from skeletal shape [5,17], and mandibular shape models capture age and sex in a growing cohort [14]. Abdominal organ shapes predict diabetes status in population imaging [9], and point-cloud networks predict age and disease from the shapes of brain structures [10]. Organ geometry thus carries clinically relevant signal beyond demographics. The size-versus-shape question itself has a long history in geometric morphometrics, where allometry (size-driven shape change) is treated as a distinct effect to be modeled or removed [17]; our size-versus-shape decomposition follows that tradition. These works are, however, typically single-organ studies that neither isolate shape from size nor test cross-site transfer.

Identity and demographics in medical AI: Deep models recover self-reported race from images [8] and exhibit demographic performance bias [23]. Anonymized brain MRIs can be re-identified by face recognition on rendered surfaces [22], and chest radiographs are likewise biometric [19]. BrainPrint showed that brain morphology alone identifies individuals across repeated MRI scans and additionally predicts age and sex [26]. Similarly, salient keypoint signatures from brain MRI are individual-specific and shared among close relatives [3]. The linkage risk we quantify is the re-identification concern that k-anonymity formalizes for tabular data [24]: a subject is at risk when their combination of attributes, here binned shape descriptors, is unique in the dataset. We extend this line of work from rendered faces, chest radiographs, and brain morphology to internal organ shapes obtained by automatic whole-body segmentation [28,13] on a public dataset.

3 Materials and Methods

3.1 Dataset and labels

We use the public TotalSegmentator v2 dataset [28]: 1228 CT studies with segmentations of 117 structures and metadata covering age (16–99 years, mean 63), sex (716 male, 510 female; missing for two), source institution, and acquisition parameters. We select 16 organs with adequate field-of-view coverage (Sec. 3.2), spanning soft viscera, vasculature, and skeleton. After excluding subjects with no fully captured organ or with missing sex (54 of 1228), 1174 subjects enter the main analyses. For the cross-modality test (Sec. 4.6), we additionally use the TotalSegmentator-MRI release (240 MRI scans with usable shapes, of which 196

carry sex and age labels and form the MRI evaluation sets in Table 5), which shares the same layout and metadata as the CT set. We run the identical pipeline on both modalities, placing CT and MRI descriptors in a single feature space. Fig. 1 summarizes the pipeline.

3.2 Shape extraction and truncation control

For each organ mask we apply marching cubes [16] in the image's physical right-anterior-superior (RAS) coordinate frame, the patient-based anatomical axes. This yields an anatomically oriented, millimeter-scaled surface. From this surface we sample 2048 points uniformly by area, giving one surface point cloud per organ (Fig. 1c). No rotation normalization is applied: surfaces stay in this anatomical frame. This is safe for most of the descriptors below, because they are rotation invariant by construction and give the same value no matter how the organ is oriented in the scanner. The exceptions are the bounding-box features, which occur in both families: the extents and the diagonal among the size features, and the two aspect ratios among the scale-free shape features. The bounding box is axis-aligned in the anatomical frame, so its features change when the patient lies differently in the scanner. Patient positioning in CT is similar but not identical across subjects. The small orientation differences that remain add subject-to-subject variability to the bounding-box features; we accept this as measurement noise rather than correcting for it. CT fields of view vary widely, so many organs are clipped at the volume boundary. For each organ, we count the mask voxels that lie on the outer boundary of the image volume. Organ instances with more than 20 such voxels are discarded as truncated, since their shape is anatomically incomplete (Fig. 1b). The tolerance of 20 voxels absorbs boundary-touching label noise and was fixed a priori, not tuned. Only organs fully captured within the image volume are included in the analysis. Per-organ sample sizes range from 427 (aorta) to 669 (heart) subjects. The full numbers per organ are listed in Table 1. Discarding truncated organs could bias each per-organ subset, e.g., if truncation were more common in one sex or age group. We checked this: each organ's valid subset matches the full cohort within $\sim$2 yr in mean age and within $\sim$5 percentage points in female fraction, so the exclusion acts approximately like random dropout with respect to sex and age. The shape representation is also compact: the point clouds for the whole cohort occupy less than 0.5 GB, versus tens of gigabytes of raw CT, so the full analysis can be stored and re-run cheaply.

3.3 Hand-crafted descriptors: size versus scale-free shape

From each point cloud we compute descriptors in two families (Fig. 1d). All descriptors are correspondence-free: each is computed from one subject's cloud alone, with no point-to-point matching across subjects. The size family has 7 features: centroid size, volume, surface area, the three bounding-box extents, and the box diagonal. Centroid size is the square root of the summed squared distances of the surface points from their centroid, a standard scale measure

in morphometrics. The scale-free family has 8 features. Four come from the covariance eigenvalues (two elongation ratios, flatness, and anisotropy), two are bounding-box aspect ratios (ratios of the box edge lengths), and the remaining two are sphericity and the normalized surface-to-volume ratio. All eight are computed on the cloud rescaled to unit centroid size, so overall scale is removed and only form remains. The eigenvalue descriptors summarize how the cloud spreads along its three principal axes, e.g., how rod-like or plate-like an organ is. Exact formulas are in the released code (see Code and data availability). This split yields three feature regimes (size-only, shape-only, and both combined) and asks how much demographic signal survives after scale is removed. In addition, one binary visibility indicator per organ records whether that organ was fully captured in the scan or truncated. The indicators are not features: they are excluded from all descriptor models and serve only to select the per-organ subsets and to tell the deep model which organs to skip in its pooling step. When a pooled model needs descriptors for an organ that is missing in a subject, they are filled with that descriptor's median computed only over the training fold, so the held-out fold does not influence these values. Each organ thus contributes 15 descriptors (7 size, 8 shape), giving 240 features per subject. These features are the input to the predictive models (Sec. 3.4).

3.4 Compared representations and predictive models

We compare three shape representations (Fig. 1e): (1) the hand-crafted descriptors (Sec. 3.3), (2) a correspondence-based statistical shape model (SSM), and (3) learned point-cloud encoders. For the SSM, we pick one template instance per organ and align every other instance to it using the iterative closest point algorithm, with the Umeyama estimator [25] for the similarity transform at each iteration. We then resample each aligned cloud to the template's point ordering, so every subject has the same points in the same anatomical order, and take the leading principal component analysis (PCA) modes as features [12,2,17]. For the learned encoders, a shared PointNet [20] or DGCNN [27] embeds each organ point cloud rescaled to unit centroid size as in Sec. 3.3. Our DGCNN implementation connects each point to its k nearest neighbors once, on the input coordinates, rather than rebuilding the graph in feature space at every layer as in the original method [27]. Since this simplification is not faithful to the published method, we do not report DGCNN results (Sec. 4.8). The static-graph implementation itself remains available in the released code. An attention pooling layer then fuses the organ embeddings of a scan into a single subject embedding: it learns a weight for each organ and averages them accordingly, with missing organs masked out using the visibility indicators. Task-specific heads predict sex, age, and pathology from this embedding, and subjects without a given label are masked out of that task's loss. Related encoders and image-to-shape approaches exist [21,11,29,1]. On the descriptor and SSM features we train gradient-boosted trees [4], random forests, and linear models. Deep networks use validation-based early stopping, and results are averaged over 3 seeds. All model hyperparameters are fixed in the released training code (see Code and data availability);

the gradient-boosted trees, for example, use 300 (sex) or 400 (age) estimators of maximum depth 4 with learning rate 0.05.

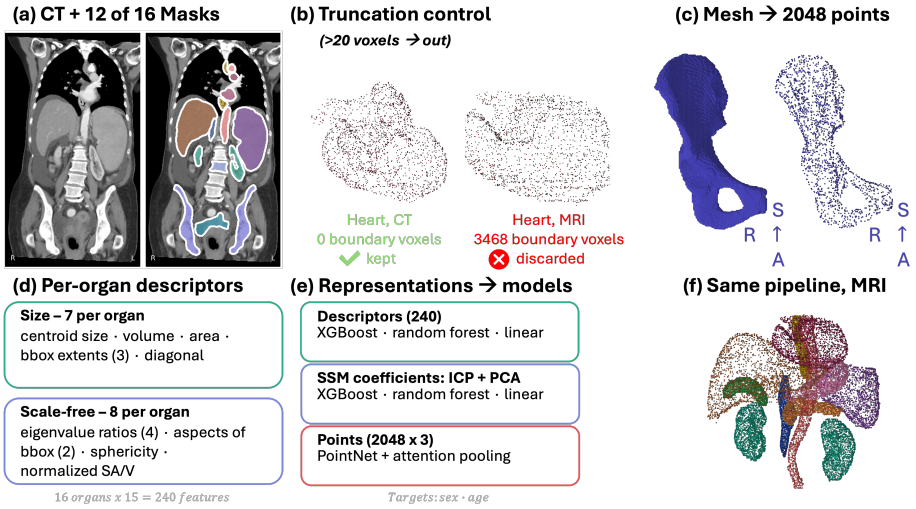

Fig. 1. Overview of the pipeline. (a) A TotalSegmentator CT subject, shown as the raw coronal slice and overlaid with the study-organ segmentations that intersect this slice. (b) Truncation control: organ instances with more than 20 mask voxels on the volume boundary, counted on the segmentation mask before point sampling, are discarded as anatomically incomplete; shown are a fully captured heart from CT (kept) and a heart clipped at the field of view in MRI (discarded). (c) Marching cubes turns each kept mask into a surface, from which 2048 points are sampled (right hip shown; R, A, S mark the anatomical axes). (d) Each organ contributes 7 size and 8 scale-free descriptors, for 240 features across 16 organs. (e) The three compared shape representations and their predictive models. (f) The identical pipeline applied to a TotalSegmentator-MRI subject.

3.5 Evaluation

Sex is evaluated by balanced accuracy and area under the ROC curve (AUC), age by mean absolute error (MAE) and R^2, under random 5-fold cross-validation with one feature row per subject. Folds are not grouped by institution, so these are within-distribution estimates. For the main results, the overall sex AUC and age MAE, we additionally report 95% confidence intervals obtained by bootstrap resampling of subjects (2000 resamples). Generalization is assessed separately in two ways. Leave-one-institution-out holds out each larger institute in turn (Sec. 4.5). Cross-modality transfer trains the gradient-boosted-tree model on descriptors from one modality and evaluates it on the other modality without change (Sec. 4.6). For privacy we measure how distinct subjects are under

quantized shape signatures (Sec. 4.10). The subject and organ exclusion rules are those of Secs. 3.1 and 3.2; descriptor definitions and model hyperparameters are fixed in the released code and configuration (see Code and data availability).

4 Results

4.1 Demographic signal in shape alone

Using all 16 organs and the full descriptor set (size and scale-free), shape alone predicts sex with AUC 0.866 (bootstrap 95% CI 0.846–0.886; the 5-fold-CV point estimate used in the tables is 0.868) and age with MAE 9.3 years (bootstrap 95% CI 8.8–9.7 years; $R^2 = 0.34$) over 1174 subjects. In contrast, a coarse scan-level pathology label is at best weakly recoverable from organ shape (descriptor AUC 0.523 under cross-validation; the deep model reaches 0.618 on held-out data). Not everything is in shape.

Table 1. Per-organ prediction from shape alone (cross-validated). Organs are ordered by sex AUC.

Organ	n	Sex AUC	Sex bal.acc	Age R^2
hip right	450	0.940 ± 0.031	0.859	0.136
hip left	453	0.923 ± 0.028	0.850	0.049
vertebrae L3	556	0.903 ± 0.026	0.800	0.248
esophagus	594	0.794 ± 0.036	0.732	0.134
aorta	427	0.790 ± 0.028	0.708	0.298
heart	669	0.780 ± 0.069	0.708	0.017
kidney left	538	0.761 ± 0.040	0.677	0.175
sacrum	513	0.759 ± 0.023	0.682	-0.005
kidney right	548	0.746 ± 0.050	0.675	0.128
liver	531	0.740 ± 0.035	0.661	-0.166
pancreas	574	0.719 ± 0.054	0.633	0.073
stomach	575	0.702 ± 0.032	0.626	-0.077
urinary bladder	516	0.688 ± 0.021	0.627	-0.320
inferior vena cava	479	0.680 ± 0.039	0.616	-0.181
spleen	602	0.660 ± 0.045	0.609	-0.123
gallbladder	504	0.549 ± 0.057	0.560	-0.225

4.2 Per-organ attribution

Table 1 and Fig. 2 rank organs. The hip bones are the strongest sex predictors (hip right, AUC 0.940), followed by the vertebrae L3 (AUC 0.903). The gallbladder is near chance (AUC 0.549). The pelvis result recovers well-known sexual dimorphism with no anatomical priors, i.e., every organ receives the same generic descriptors and models. Age is best predicted by the aorta ($R^2 = 0.30$),

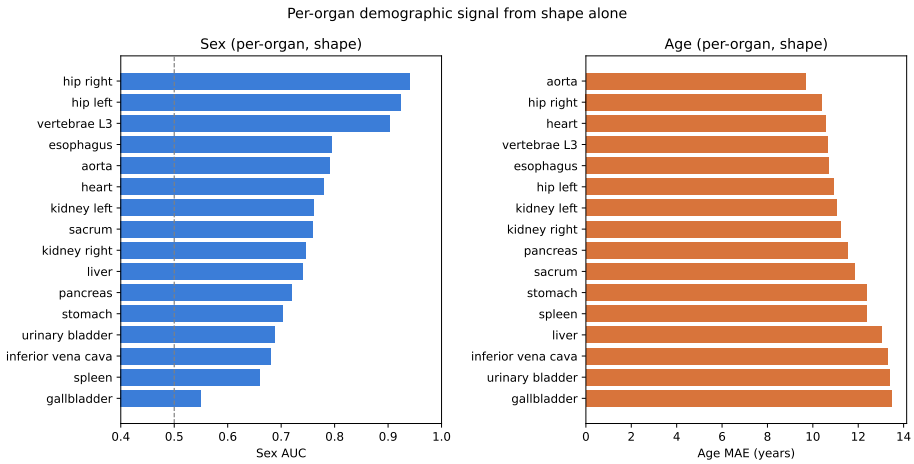

Fig. 2. Per-organ sex AUC (left) and age MAE in years (right) from shape alone. Table 1 lists sex AUC with fold-to-fold standard deviations and age R^2; per-organ MAE values are in the released results (see Code and data availability).

consistent with vascular and skeletal aging. No multiple-comparison correction (e.g., Bonferroni) is applied across the 16 organs. The ranking is exploratory, adjacent organs' intervals overlap, and organs near chance (e.g., gallbladder) should not be read as carrying signal. Several organs show negative R^2 values for age (Table 1), indicating predictions worse than always guessing the cohort mean age. Thus, these organs carry no usable age signal in isolation.

4.3 Size versus shape

Table 2 decomposes the signal. Sex is largely, but not only, a matter of size: size features alone give AUC 0.863, yet scale-free shape still reaches AUC 0.791, well above the 0.5 chance level. Age is predicted comparably by size and shape, and best by their union (MAE 9.3 years).

Table 2. Size-only vs scale-free-shape vs combined feature regimes.

Feature set	dim	Sex AUC	Sex bal.acc	Age MAE	Age R^2
Size only	112	0.863 ± 0.017	0.762	9.6 ± 0.7	0.293
Shape only (scale-free)	128	0.791 ± 0.035	0.699	9.6 ± 0.7	0.297
Size + shape	240	0.868 ± 0.016	0.755	9.3 ± 0.7	0.336

4.4 How much is in shape alone? Shape vs intensity

To test whether shape is merely a weak proxy for the image, we compare it with an intensity baseline on the same subjects (Table 3): per organ, seven Hounsfield-unit statistics (mean, standard deviation, and the 10th, 25th, 50th, 75th, and 90th percentiles), giving 112 intensity features. A coincidence of the design gives partial control: the size family alone has exactly 112 features, matching the intensity baseline's dimensionality while still exceeding it (AUC 0.863 vs 0.729). For sex, shape alone (AUC 0.868) exceeds intensity (0.729) and equals the combined model (0.865). The feature counts differ (shape: 240, intensity: 112), so the comparison reflects each modality's natural representation rather than feature-for-feature parity. A dimensionality-matched comparison (e.g., the 112 strongest shape features) is left for future work, and we report no paired statistical test. While the absence of a combined-model gain may also reflect model design or redundancy between the feature sets, the margin over this intensity-summary baseline suggests the sex signal is largely geometric rather than a density artifact. For age, intensity is complementary. Combining modalities lowers MAE from 9.3 to 8.8 years. Against the tested baseline, shape is therefore the dominant carrier of the sex signal, which sharpens the privacy implication in Sec. 5.

Table 3. Shape vs image-intensity features (same subjects, 5-fold CV).

Feature modality	dim	Sex AUC	Age MAE
Geometric descriptors (all)	240	0.868	9.3
Intensity (HU stats)	112	0.729	9.5
Shape + intensity	352	0.865	8.8

4.5 Cross-site generalization

Under leave-one-institution-out (each of the 5 larger institutes held out in turn; Table 4), the shape signal persists out-of-site rather than collapsing. The most informative case holds out the dominant, diverse site (I, n=854), giving sex AUC 0.827. The smaller held-out sites score higher, with sex AUCs up to 0.974. These estimates are noisy: the test sets are small (n as low as 52) and their confidence intervals span roughly 0.1 AUC. The sites also differ in composition, with female fractions ranging from 29% to 53%. We therefore read the inflated AUCs as artifacts of small test sets and demographic skew, not as evidence of better out-of-site generalization. Relatedly, a held-out site's AUC can exceed the within-distribution 5-fold CV estimate (Sec. 4.1), plausibly because the multi-site training pool is larger and more diverse than any single site alone.

Table 4. Leave-one-institution-out generalization. Institute codes are the anonymized site labels of the TotalSegmentator metadata. Shown are the five sites with at least 50 subjects, ordered by size. All institutes come from one curated dataset, largely share scanner vendors, and use the same segmentation pipeline, so this is within-dataset transfer, not external-domain generalization.

Held-out institute	n	% female	Sex AUC (95% CI)	Age MAE
I	854	41	0.827 (0.80–0.85)	10.5
C	96	41	0.922 (0.86–0.96)	7.9
A	81	53	0.843 (0.75–0.92)	8.4
B	52	29	0.935 (0.86–0.99)	8.9
E	52	44	0.974 (0.93–1.00)	8.2

4.6 Cross-modality: does the signal hold on MRI?

A stronger test of whether the signal is genuinely geometric (rather than a CT-specific artifact) is to change imaging modality entirely. We apply the identical pipeline to TotalSegmentator-MRI. Of its scans, 240 yield usable organ shapes, and 196 of these also carry sex and age labels. These 196 subjects form the evaluation set for all MRI analyses (Table 5). The dataset covers only abdominal organs, so the per-organ n is small. Sex prediction replicates on MRI (within-MRI AUC 0.762; Table 5) and transfers across modality above chance: a classifier trained on CT shapes scores AUC 0.708 on MRI, and MRI→CT 0.677. The above-chance transfer indicates that much of the sex signal is geometry shared across modalities. The drop from the within-modality AUC 0.866 likely reflects genuine differences between the CT and MRI segmentations, from both the modalities themselves and their separate segmentation pipelines. Age, by contrast, does not replicate on this smaller MRI set (MAE 14.8 yr, negative R^2). We attribute this to the reduced cohort and the abdominal field of view: the strongest age carriers of Table 1, the aorta and the L3 vertebra, are rarely or never fully captured on MRI, so their descriptors are median-imputed for most subjects and carry no age information.

Table 5. Cross-modality sex prediction with the same pipeline: within-MRI and CT↔MRI transfer. n is the evaluation-set size.

Setting	n	Sex AUC
Within MRI (5-fold CV)	196	0.762 ± 0.029
CT → MRI (transfer)	196	0.708
MRI → CT (transfer)	1174	0.677

4.7 Deep multi-task model

A multi-organ PointNet shares one shape encoder across organs and predicts sex, age, and pathology from a single subject embedding (Sec. 3.4). It is trained with validation-based early stopping and averaged over 3 seeds. On the held-out test split ($n = 80$) it reaches sex AUC 0.780 ± 0.017, below the gradient-boosted trees on the full descriptor set (AUC 0.868). Age is harder for this model ($R^2 = 0.02$), plausibly because the encoder sees size-normalized clouds and therefore cannot use organ scale. The pathology head stays near chance (AUC 0.618, on the subset of test subjects that carry a pathology label), consistent with the descriptor result reported above. Training the three tasks jointly gave no measurable gain over single-task variants. We therefore treat this network as a learned-shape baseline rather than a contribution, and compare it against the other representations in Sec. 4.8.

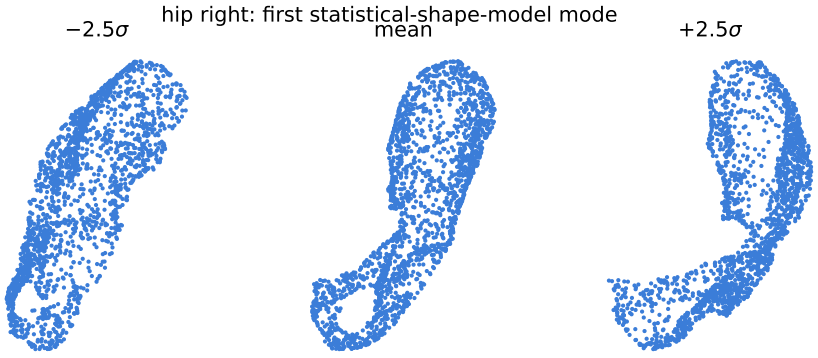

Fig. 3. First statistical-shape-model mode of the hip (mean $\pm 2.5\sigma$, coronal view), illustrating shape variation in the most sex-discriminative structure. This mode captures maximum population variance, which need not coincide with the sex-discriminative direction.

4.8 Comparison of representations and models

Table 6 compares the three shape representations of Sec. 3.4 and the models trained on them. Gradient-boosted trees on the full descriptor set are strongest (sex AUC 0.868). Gradient-boosted trees on SSM coefficients reach AUC 0.636, and the PointNet reaches 0.780 ± 0.017. Neither matches the full descriptor set. The PointNet row comes from a single held-out split of 80 subjects rather than 5-fold cross-validation over 1174, so that comparison is suggestive rather than definitive. The matched comparison carries the main evidence: organ volumes alone, one feature per organ, reach AUC 0.815 against 0.868 for the full descriptor set, with the same model and the same 5-fold protocol. Adding the remaining

size features (Table 2, AUC 0.863) and then the scale-free descriptors changes little. The demographic signal therefore appears to live in low-order geometry, that is, overall size and gross shape, rather than in fine surface detail or high-frequency point correspondence. The PointNet result points the same way, but its protocol differs, so it does not establish the claim on its own. Why the SSM underperforms is untested here. Plausible factors are the quality of the approximate ICP correspondence, the choice of template, and the number of retained modes. PCA ranks modes by how much subjects differ along them, not by how well they separate the sexes, so keeping only the leading modes can discard a direction that carries the sex signal. Fig. 3 shows the leading SSM mode of the hip, the most sex-discriminative structure. The PointNet falls further behind on age than on sex: MAE 12.3 years against 9.3 years for the full descriptor set. This is plausibly because the encoder sees size-normalized clouds, so organ scale, a strong age cue (Table 2), is not available to it, on top of the smaller evaluation split.

Table 6. Representations and models compared on sex (AUC) and age (MAE, years). Descriptor and SSM rows use 5-fold cross-validation (n=1174); deep-net rows report a single held-out test split (n=80) averaged over 3 seeds (PointNet seed-to-seed std $\pm$0.017) and are therefore noisier and not directly comparable. DGCNN is not listed: our static-graph implementation is a simplification of the original dynamic-graph method, so its numbers would not represent the published method (see text).

Method	Eval	Sex AUC	Age MAE
Naive (majority / mean age)	5-fold CV	0.500	11.7
Organ volumes only (XGB)	5-fold CV	0.815 ± 0.023	10.5
Size descriptors (XGB)	5-fold CV	0.863 ± 0.017	9.6
Shape descriptors (XGB)	5-fold CV	0.791 ± 0.035	9.6
All descriptors (linear)	5-fold CV	0.854 ± 0.024	10.0
All descriptors (RandomForest)	5-fold CV	0.849 ± 0.026	9.6
All descriptors (XGBoost)	5-fold CV	0.868 ± 0.016	9.3
SSM-PCA coeffs (XGB)	5-fold CV	0.636	11.7
Multi-organ PointNet	held-out test	0.780	12.3

4.9 Shape as an auto-labeling tool

The same predictive signal can constructively enrich incomplete metadata in research archives of unlabeled segmentations. Sex predictions are well calibrated. The expected calibration error is 0.045, the bin-weighted mean gap between predicted confidence and observed accuracy, so a subject predicted at 90% confidence is right close to 90% of the time. Ranking subjects by out-of-fold confidence and labeling only the most confident fraction therefore gives reliable selective auto-labeling. Accuracy rises from 78% at full coverage to 94% over the most-confident half and to 98.3% on the top quartile (Fig. 4). In total, 43% of sub-

jects can be auto-sexed at ≥95% accuracy. The approach is also label-efficient: 82 labeled subjects already give AUC 0.758. We target sex, where confidence is meaningful. Age is too coarse for confident per-subject labels (MAE 9.3 yr).

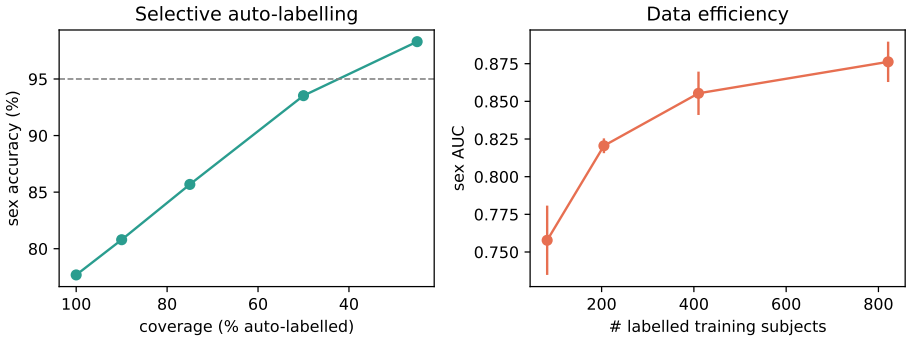

Fig. 4. Auto-labeling sex from shape. Left: out-of-fold accuracy versus coverage; restricting to the most confident subjects raises accuracy from 78% (all) to 98.3% (top quartile). Right: learning curve; 82 labeled subjects already give AUC 0.758.

4.10 Privacy: scale-free shape is highly distinctive

Fig. 5 quantifies distinctiveness as a function of the number of organs used. Signatures concatenate the 8 scale-free shape descriptors per organ, each quantile-binned into 10 levels. Organs are added in order of decreasing availability, the most often fully captured first. Each k-organ estimate uses only subjects with all k organs present, so n falls from 669 at k=1 to 177 at k=16 and the trend is therefore indicative rather than exact. The most frequently captured organ, the heart, already gives a unique signature for 93% of the subjects in which it is fully captured, and adding the second organ raises this to 100%. Distinctiveness also depends on the quantization: at a coarse 3-bin quantization only 18% are unique from that single organ, rising to 99% at 20 bins. Any single uniqueness value is therefore one point on this curve, set by the chosen resolution as much as by the anatomy. These figures are upper bounds on linkability, not re-identification rates (Sec. 5). Even so, combined with the demographic results, they show that sharing organ meshes stripped of intensity and metadata is far from anonymous: shape leaks attributes and is near-unique per person.

5 Discussion

What is in shape. Sex and age are substantially encoded in organ shape, concentrated in the skeleton and great vessels (Table 1). The pelvis result reproduces

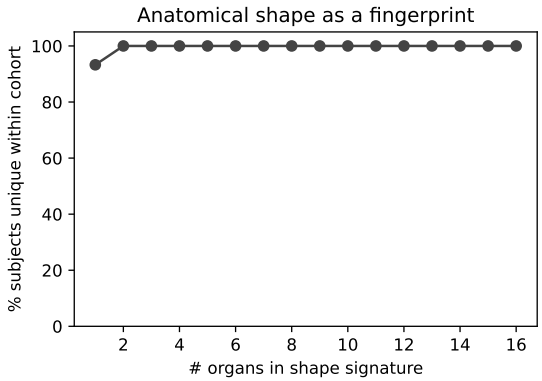

Fig. 5. Fraction of subjects with a unique within-cohort signature as a function of the number of organs used (10-bin quantization; organs added in order of decreasing availability; each point uses only subjects with all k organs present).

established forensic knowledge without anatomical priors. The size-versus-shape analysis cautions against equating "shape-based" sex prediction with morphology per se, since much of the signal is scale (Table 2). Models built on organ shape may therefore inherit demographic information even when trained for unrelated tasks, which matters for fairness audits of shape-based pipelines.

Age. Age is harder than sex from shape (MAE 9.3 yr, $R^2 = 0.34$). Intensity carries part of what shape misses: adding Hounsfield statistics lowers MAE from 9.3 to 8.8 years (Table 3), which suggests that aging alters tissue composition more than outer organ geometry. The shape signal is also diffuse across organs. Even the best single organ, the aorta, reaches only $R^2 = 0.30$ (Table 1). Chronological age is itself a noisy proxy for biological aging, which plausibly sets an MAE floor for any predictor.

Enrichment versus leakage. The auto-labeling and privacy results are the same fact seen from opposite sides. Because shape predicts demographics, it can fill in missing sex and age metadata in consented, internal research datasets (Fig. 4), and it also makes shared "de-identified" meshes non-anonymous (Fig. 5). We advocate the first use only for internal enrichment, never for inference on released data.

When is removing demographic signal desirable. In many clinical studies, age and sex are variables of interest rather than confounders, and pathology itself alters shape. Removing demographic signal from shared shapes is therefore not a universal goal. It is warranted when shapes leave their governance context, for example in public mesh releases, where demographic leakage serves no analytic purpose. Any removal method should be judged on how much demographic signal it removes and, equally, on how much clinically relevant variation it preserves.

Privacy. The distinctiveness in Fig. 5 is an upper bound on linkability, not a measured re-identification rate, for two reasons. First, we count unique signatures within one cohort and do not match individuals across separate acquisitions, as in BrainPrint's longitudinal design [26]. Second, with 10 bins over 8 descriptors per organ, the signature space (10^8) far exceeds the cohort size ($n \approx 1174$), so high uniqueness is partly expected from the feature-to-subject ratio alone. An entropy or k-anonymity style analysis under attacker-realistic quantization remains future work, and the bin sweep in Sec. 4.10 is a first step in that direction. Because the signature space grows with the number of organs, distinctiveness will plausibly persist in larger cohorts, but confirming a practical privacy risk requires repeat scans. With these caveats, our findings corroborate evidence that medical images retain biometric information after metadata and intensity are removed: brain MRI surfaces enable face-recognition-based re-identification [22], chest radiographs carry patient-specific signatures [19], and brain morphometry suffices for longitudinal re-identification [26]. We extend this evidence to internal organ shape. Until repeat-scan validation exists, we recommend treating shared organ meshes as potentially linkable [24] rather than anonymous.

Ethics: race and ethnicity. We deliberately do not predict race. Race is a social rather than biological construct, and models that "recognize" it from medical data [8] risk conflating ancestry with socioeconomic confounders and enabling discriminatory profiling. We raise the point to motivate caution, and argue that shape-based demographic inference should be audited for fairness, not deployed for identification.

Limitations. The available labels bound what can be claimed. The pathology label is coarse and scan-level, and sex is treated as binary throughout, reflecting the labels available in TotalSegmentator, so intersex individuals are not represented. The cohort is a convenience sample from one population with unknown ancestry, since TotalSegmentator does not report ethnicity, so the observed distinctiveness should not be read as cross-population re-identifiability. The cohort is also adult (ages 16–99 years, mean 63). Whether shape dominance over intensity holds in pediatric populations, where intensity is strongly age-dependent through, e.g., ossification, is unknown. Per-organ analyses use field-of-view-dependent subsets (Table 1). Scan protocols differ across institutions, which drives the organ-specific truncation rates and the per-organ sample sizes, from 427 for the aorta to 669 for the heart. The truncation-bias check in Sec. 3.2 shows this selection is approximately random with respect to sex and age, but the varying sample sizes still imply different statistical power per organ, which could influence the ranking. Median imputation of missing organs can also encode field-of-view patterns implicitly in the pooled models. The per-organ results in Table 1 use only subjects with the organ fully captured and are free of this effect, but a visibility-only baseline remains future work. Size, age-related development, and pathology can confound one another, and the size-versus-shape decomposition separates scale without adjusting for these covariates. Our descriptors come from the released TotalSegmentator segmentations, which we use without further manual correc-

tion. The strong cross-validated performance suggests the demographic signal is robust to typical segmentation variability, though a formal sensitivity analysis remains future work. The MRI cohort is small and abdominally focused, so MRI age does not replicate and the cross-modality numbers are indicative rather than definitive. Both CT and MRI segmentations come from the TotalSegmentator ecosystem, so an external population with an independent segmentation method remains untested.

6 Conclusion

Anatomical shape, with intensity removed, still carries substantial demographic signal and shows high within-cohort distinctiveness. One pipeline over 16 organs attributes that signal to individual organs (Table 1), separates it into size and scale-free components (Table 2), and shows that it survives both a change of site and a change of modality (Tables 4 and 5). The same signal is useful and dangerous: it can auto-label incomplete metadata (Fig. 4), and it makes quantized scale-free shape near-unique within the cohort (Fig. 5). That distinctiveness is an upper bound on linkability rather than a measured re-identification rate. Even so, stripping metadata alone may not constitute effective de-identification when organ shape itself carries biometric signal. The full pipeline is released with the paper.

Code and data availability

The full pipeline (extraction, training, evaluation, and result generation) is publicly available at `https://github.com/TIO-IKIM/shape-mi-demographics`, including the complete experimental configuration. All imaging data are from the publicly released TotalSegmentator CT and MRI datasets [28] (CC BY 4.0); no new patient data were acquired.

Acknowledgments. This research was supported by the REACT-EU funded project KITE, Plattform für KI-Translation Essen (Grant No. EFRE-2920801977), `https://kite.ikim.nrw`. 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 to declare that are relevant to the content of this article.

References

1. Bhalodia, R., Elhabian, S., Adams, J., Tao, W., Kavan, L., Whitaker, R.: DeepSSM: A blueprint for image-to-shape deep learning models. Medical Image Analysis **91**, 103034 (2024). `https://doi.org/10.1016/j.media.2023.103034`

2. 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
3. Chauvin, L., Kumar, K., Wachinger, C., Vangel, M., de Guise, J., Desrosiers, C., Wells, W., Toews, M.: Neuroimage signature from salient keypoints is highly specific to individuals and shared by close relatives. NeuroImage **204**, 116208 (2020). https://doi.org/10.1016/j.neuroimage.2019.116208
4. Chen, T., Guestrin, C.: XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 785–794 (2016). https://doi.org/10.1145/2939672.2939785
5. d'Oliveira Coelho, J., 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
6. 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
7. Cole, J.H., Poudel, R.P.K., Tsagkrasoulis, D., Caan, M.W.A., Steves, C., Spector, T.D., Montana, G.: Predicting brain age with deep learning from raw imaging data results in a reliable and heritable biomarker. NeuroImage **163**, 115–124 (2017). https://doi.org/10.1016/j.neuroimage.2017.07.059
8. 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
9. Gutiérrez-Becker, B., Gatidis, S., Gutmann, D., Peters, A., Schlett, C., Bamberg, F., Wachinger, C.: Deep shape analysis on abdominal organs for diabetes prediction. In: Shape in Medical Imaging (ShapeMI 2018). Lecture Notes in Computer Science, vol. 11167, pp. 223–231. Springer International Publishing (2018). https://doi.org/10.1007/978-3-030-04747-4_21
10. Gutiérrez-Becker, B., Sarasua, I., Wachinger, C.: Discriminative and generative models for anatomical shape analysis on point clouds with deep neural networks. Medical Image Analysis **67**, 101852 (2021). https://doi.org/10.1016/j.media.2020.101852
11. Hanocka, R., Hertz, A., Fish, N., Giryes, R., Fleishman, S., Cohen-Or, D.: MeshCNN: A network with an edge. ACM Transactions on Graphics **38**(4), 90:1–90:12 (2019). https://doi.org/10.1145/3306346.3322959
12. 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
13. 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**, 203–211 (2021). https://doi.org/10.1038/s41592-020-01008-z
14. 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
15. Li, J., Zhou, Z., Yang, J., Pepe, A., Gsaxner, C., Luijten, G., Qu, C., Zhang, T., Chen, X., Li, W., Wodzinski, M., Friedrich, P., Xie, K., Jin, Y., Ambigapathy,

N., Nasca, E., Solak, N., Melito, G.M., Vu, V.D., Memon, A.R., Schlachta, C., De Ribaupierre, S., Patel, R., Eagleson, R., Chen, X., Mächler, H., Kirschke, J.S., de la Rosa, E., Christ, P.F., Li, H.B., Ellis, D.G., Aizenberg, M.R., Gatidis, S., Küstner, T., Shusharina, N., Heller, N., Andrearczyk, V., Depeursinge, A., Hatt, M., Sekuboyina, A., Löffler, M.T., Liebl, H., Dorent, R., Vercauteren, T., Shapey, J., Kujawa, A., Cornelissen, S., Langenhuizen, P., Ben-Hamadou, A., Rekik, A., Pujades, S., Boyer, E., Bolelli, F., Grana, C., Lumetti, L., Salehi, H., Ma, J., Zhang, Y., Gharleghi, R., Beier, S., Sowmya, A., Garza-Villarreal, E.A., Balducci, T., Angeles-Valdez, D., Souza, R., Rittner, L., Frayne, R., Ji, Y., Ferrari, V., Chatterjee, S., Dubost, F., Schreiber, S., Mattern, H., Speck, O., Haehn, D., John, C., Nürnberger, A., Pedrosa, J., Ferreira, C., Aresta, G., Cunha, A., Campilho, A., Suter, Y., Garcia, J., Lalande, A., Vandenbossche, V., Van Oevelen, A., Duquesne, K., Mekhzoum, H., Vandemeulebroucke, J., Audenaert, E., Krebs, C., van Leeuwen, T., Vereecke, E., Heidemeyer, H., Röhrig, R., Hölzle, F., Badeli, V., Krieger, K., Gunzer, M., Chen, J., van Meegdenburg, T., Dada, A., Balzer, M., Fragemann, J., Jonske, F., Rempe, M., Malorodov, S., Bahnsen, F.H., Seibold, C., Jaus, A., Marinov, Z., Jaeger, P.F., Stiefelhagen, R., Santos, A.S., Lindo, M., Ferreira, A., Alves, V., Kamp, M., Abourayya, A., Nensa, F., Hörst, F., Brehmer, A., Heine, L., Hanusrichter, Y., Weßling, M., Dudda, M., Podleska, L.E., Fink, M.A., Keyl, J., Tserpes, K., Kim, M.S., Elhabian, S., Lamecker, H., Zukić, D., Paniagua, B., Wachinger, C., Urschler, M., Duong, L., Wasserthal, J., Hoyer, P.F., Basu, O., Maal, T., Witjes, M.J.H., Schiele, G., Chang, T.C., Ahmadi, S.A., Luo, P., Menze, B., Reyes, M., Deserno, T.M., Davatzikos, C., Puladi, B., Fua, P., Yuille, A.L., Kleesiek, J., Egger, J.: MedShapeNet - a large-scale dataset of 3D medical shapes for computer vision. Biomedizinische Technik. Biomedical Engineering **70**(1), 71–90 (Feb 2025). https://doi.org/10.1515/bmt-2024-0396

16. Lorensen, W.E., Cline, H.E.: Marching cubes: A high resolution 3D surface construction algorithm. In: Proceedings of the 14th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH). pp. 163–169 (1987). https://doi.org/10.1145/37401.37422
17. Mitteroecker, P., Gunz, P.: Advances in geometric morphometrics. Evolutionary Biology **36**(2), 235–247 (2009). https://doi.org/10.1007/s11692-009-9055-x
18. Moscoloni, B., Beeche, C., Chirinos, J.A., Segers, P., Peirlinck, M.: Unveiling sex dimorphism in the healthy cardiac anatomy: Fundamental differences between male and female heart shapes. The Journal of Physiology **604**(13), 5512–5534 (2026). https://doi.org/10.1113/JP288667
19. Packhäuser, K., Gündel, S., Münster, N., Syben, C., Christlein, V., Maier, A.: Deep learning-based patient re-identification is able to exploit the biometric nature of medical chest X-ray data. Scientific Reports **12**, 14851 (2022). https://doi.org/10.1038/s41598-022-19045-3
20. 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
21. Qi, C.R., Yi, L., Su, H., Guibas, L.J.: PointNet++: Deep hierarchical feature learning on point sets in a metric space. In: Advances in Neural Information Processing Systems 30 (NeurIPS 2017). pp. 5099–5108 (2017), https://proceedings.neurips.cc/paper/2017/hash/d8bf84be3800d12f74d8b05e9b89836f-Abstract.html

22. Schwarz, C.G., Kremers, W.K., Therneau, T.M., et al.: Identification of anonymous MRI research participants with face-recognition software. New England Journal of Medicine **381**(17), 1684–1686 (2019). https://doi.org/10.1056/NEJMc1908881
23. Seyyed-Kalantari, L., Zhang, H., 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
24. Sweeney, L.: k-anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems **10**(5), 557–570 (2002). https://doi.org/10.1142/S0218488502001648
25. Umeyama, S.: Least-squares estimation of transformation parameters between two point patterns. IEEE Transactions on Pattern Analysis and Machine Intelligence **13**(4), 376–380 (1991). https://doi.org/10.1109/34.88573
26. 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
27. 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
28. 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
29. Zhao, H., Jiang, L., Jia, J., Torr, P.H.S., Koltun, V.: Point transformer. In: IEEE/CVF International Conference on Computer Vision (ICCV). pp. 16239–16248 (2021). https://doi.org/10.1109/ICCV48922.2021.01595