Radiology: Artificial Intelligence

# Segmenting Whole-Body MRI and CT for Multiorgan Anatomic Structure Delineation

Hartmut Häntze, MSc*[1,2]

Lina Xu, MD*[1]

Christian J. Mertens, MD[3]

Felix J. Dorfner[1,4]

Leonhard Donle, MSc[1]

Felix Busch, MD[3]

Avan Kader, MSc[3]

Sebastian Ziegelmayer, MD[3]

Nadine Bayerl, MD[5]

Nassir Navab, PhD[6]

Daniel Rueckert, PhD[7,8,9]

Julia Schnabel, PhD[9]

Hugo J. W. L. Aerts, PhD[10,11,12]

Daniel Truhn, MSc, MD[13]

Fabian Bamberg, MD, MPH[14]

Jakob Weiss, MD[14]

Christopher L. Schlett, MD[14]

Steffen Ringhof, PhD[14]

Thoralf Niendorf, PhD[15]

Tobias Pischon, PhD[15]

Hans-Ulrich Kauczor, MD[16]

Tobias Nonnenmacher, MD[16]

Thomas Kröncke, MD, MBA[17]

Henry Völzke, MD[18]

Jeanette Schulz-Menger, MD[19]

Klaus Maier-Hein, PhD[20]

Alessa Hering, PhD[2]

Mathias Prokop, MD, PhD[2]

Bram van Ginneken, PhD[2]

Marcus R. Makowski, MD, PhD[3]

Lisa C. Adams, MD**[3]

Keno K. Bressem, MD**[3,21]

* H.H. and L.X. contributed equally to this work.

** L.C.A. and K.K.B. are co–senior authors.

Author affiliations, funding, and conflicts of interest are listed at the end of this article.

https://doi.org/10.1148/ryct.240777

**Purpose:** To develop and validate MRSegmentator, a retrospective cross-modality deep learning model for multiorgan segmentation of MRI scans.

**Materials and Methods:** This retrospective study trained MRSegmentator on 1,200 manually annotated UK Biobank Dixon MRI sequences (50 participants), 221 in-house abdominal MRI sequences (177 patients), and 1228 CT scans from the TotalSegmentator-CT dataset. A human-in-the-loop annotation workflow leveraged cross-modality transfer learning from an existing CT segmentation model to segment 40 anatomic structures. The model's performance was evaluated on 900 MRI sequences from 50 participants in the German National Cohort (NAKO), 60 MRI sequences from AMOS22 dataset, and 29 MRI sequences from TotalSegmentator-MRI. Reference standard manual annotations were used for comparison. Metrics to assess segmentation quality included Dice Similarity Coefficient (DSC). Statistical analyses included organ-and sequence-specific mean ± SD reporting and two-sided *t* tests for demographic effects.

**Results:** 139 participants were evaluated; demographic information was available for 70 (mean age 52.7 years ± 14.0 [SD], 36 female). Across all test datasets, MRSegmentator demonstrated high class wise DSC for well-defined organs (lungs: 0.81–0.96, heart: 0.81–0.94) and organs with anatomic variability (liver: 0.82–0.96, kidneys: 0.77–0.95). Smaller structures showed lower DSC (portal/splenic veins: 0.64–0.78, adrenal glands: 0.56–0.69). The average DSC on the external testing using NAKO data, ranged from 0.85 ± 0.08 for T2-HASTE to 0.91 ± 0.05 for in-phase sequences. The model generalized well to CT, achieving mean DSC of 0.84 ± 0.12 on AMOS CT data.

**Conclusion:** MRSegmentator accurately segmented 40 anatomic structures on MRI and generalized to CT; outperforming existing open-source tools.

MRSegmentator accurately segmented 40 anatomical structures on MRI and CT across three external datasets.

## Abbreviations

AI = Artificial Intelligence, AMOS = Abdominal Multi-Organ Segmentation, DSC = Dice Similarity Coefficient, HD = Hausdorff Distance, NAKO = German National Cohort (Nationale Kohorte), T2 HASTE = T2-weighted Half-Fourier Acquisition Single-shot Turbo spin Echo, UKBB = UK Biobank, VC = Vessel Consistency

## Key Points:

- In this retrospective study of 50 participants with whole-body MRI, MRSegmentator achieved mean Dice similarity coefficients (DSC) of 0.91 ± 0.05 for Dixon in-phase and 0.85 ± 0.08 for T2-HASTE sequences.

- Across 60 Multimodality Abdominal Multi-Organ Segmentation Challenge MRI and 300 CT scans, mean DSCs were 0.79 ± 0.11 and 0.84 ± 0.12, respectively, demonstrating cross-modality generalizability.

- Class-wise performance peaked in well-defined organs (DSC, lungs: 0.96; heart: 0.94) and varied for small structures (DSC, portal/splenic veins: 0.64; adrenal glands: 0.56).

Automated segmentation of anatomic structures in medical imaging enables precise organ volumetry, facilitates anatomic context for AI-based diagnosis, and supports quantitative imaging biomarker extraction (1). Recent deep learning advances have led to robust CT segmentation models like TotalSegmentator-CT, which segments 104 anatomic structures (2), however, comparable whole-body tools for MRI segmentation remain limited.

MRI segmentation poses distinct technical challenges. Unlike CT's standardized Hounsfield units, MRI signal intensities vary across scanners and protocols, and motion or field inhomogeneities artifacts are common (3). Anisotropic voxels and lower spatial resolution further complicate consistent segmentation. Despite these challenges, MRI's higher soft-tissue contrast and absence of ionizing radiation make it preferred for longitudinal studies and tissue characterization.

Current MRI segmentation solutions are predominantly organ-specific, focusing on individual structures like the kidneys, prostate, or spleen (4–6). While these specialized models achieve high accuracy for their target organs, the need to run multiple models for different structures limits clinical workflow integration and adds computational overhead. Recent multiorgan approaches have been proposed (7–9), but their generalizability across MRI sequences and external datasets remains unclear. The growing availability of large-scale MRI repositories, such as the UK Biobank Imaging Study, creates opportunities for developing more comprehensive segmentation tools that work across multiple protocols and anatomic regions.

To address this need, we have developed MRSegmentator, an nnU-Net based (10), cross modality image segmentation model that segments 40 anatomic structures in both MRI and CT images. Our approach combines cross-modality learning from CT data (11) with a human-in-the-loop annotation workflow to accelerate image annotation and address MRI-specific challenges. We evaluate the model's performance across diverse MRI sequences using three external datasets and demonstrate its ability to handle anatomic variants.

# Materials and Methods

This retrospective study was conducted in accordance with the Declaration of Helsinki and received approval from the local ethics committee (EA4/062/20) with a waiver of patient consent.

## Datasets

Six datasets of 3D MRI or CT images were used (Fig 1). Training data included UK Biobank (UKBB), an in-house dataset, and the TotalSegmentator-CT dataset. For testing, we use MRI scans from the National German Cohort (NAKO), the Multimodality Abdominal Multi-Organ Segmentation Challenge (AMOS22), and the TotalSegmentator-MRI dataset. We created our own annotations for the in-house, UKBB, and NAKO datasets, selecting as many participants as we could annotate within a two-month timeframe each. The AMOS and both TotalSegmentator datasets included their own annotations.

## UK Biobank Dataset

The UKBB is a large biomedical database containing genetic and health information from half a million UK participants (12). We accessed the MRI subset (datafield 20201-2, October 2023), which contains whole-body MRI scans from 69,571 participants. For our study, we selected 1,200 sequences from 50 randomly selected participants. Each participant's MRI data are divided into six 3D sections from the shoulders to the knees, with in-phase (IN), opposed-phase (OPP), fat-only (F), and water-only (W) images obtained using the Dixon technique for each segment.

This dataset is highly standardized, with consistent image size, voxel spacing, and subject positioning within each region. Each sequence consists of 44 to 72 axial slices spaced 3 to 4.5 mm along the z-axis. Access to the UKBB dataset for scientific research is available upon request at https://www.ukbiobank.ac.uk/.

## In-house Dataset

We screened our in-house data for participants with kidney tumors or cysts. From the resulting 690 participants, we randomly selected 180. We also excluded scans from three participants due to poor image quality. Some participants had multiple sequence types, resulting in a final dataset of 221 axial abdominal MRI sequences from 177 participants. Tumor size is less than 7 cm in 213 scans and up to 13 cm in the remaining eight scans. The dataset has an approximately equal distribution of T1, T2-weighted fat-saturated (T2fs), and postcontrast T1-weighted fat-saturated (T1fs) sequences. Images were acquired on Siemens Magnetom Avanto 1.5T and Vida 3T scanners with different signal intensity distributions and matrix sizes.

After acquisition, the sequences were exported and resized to a uniform voxel spacing of (1) mm, resulting in sequences consisting of 100 to 450 slices along the z-axis.

## TotalSegmentator-CT Dataset

TotalSegmentator-CT (2) is a whole-body CT segmentation model that was released together with a publicly available dataset of 1,228 CT examinations with a wide range of different pathologies, scanners and acquisition protocols. To enable cross modality segmentation capabilities, we included the TotalSegmentator-CT dataset to the training data. From the 117 segmented structures, we selected a subset of 40 classes, consistent with the classes annotated for MR images (Fig 2). The dataset is available at https://zenodo.org/records/10047292.

## NAKO Dataset

The German National Cohort (NAKO) is a population-based prospective cohort study investigating the causes of the development of major chronic diseases in the German population (13). The study includes whole-body MRI scans of 30,868 participants from the general population (14). We obtained a subset of 900 MRI sequences from 50 participants (25 women, 25 men). For each subject, the data included T1-weighted 3D VIBE two-point Dixon sequences (in phase (IN), opposed phase (OPP), water only (W), fat only (F)) and T2-weighted HASTE sequences (Table 1).

The Dixon sequences consist of four 3D sections from the shoulders to the knees, and the T2 HASTE scans consist of two sections from the shoulder to the sacrum. We stitched the MRI stations for each participant and modality using a self-developed open-source utility (https://github.com/ai-assisted-healthcare/AIAH_utility) based on SimpleITK, resulting in 250 whole-body MRI sequences across contrast types. The stitched Dixon sequences have a matrix size of (320,260,316) with voxel spacing of (1.4,1.4,3.0) mm and the stitched T2-weighted HASTE sequences have a matrix size of (320,260,80) with voxel spacing of (1.4,1.4,6.0) mm. Note that stitching is only done to increase the interpretability of results and is not required for inference (see also Fig S1).

## AMOS22 Dataset

The Multimodality Abdominal Multi-Organ Segmentation Challenge (AMOS22) was held at the MICCAI conference in 2022 (11). The accessible training and validation sections include 300 CT and 60 MRI sequences/patients from multicenter, multivendor, multimodality and multidisease patients, each with voxel-level annotations of 15 abdominal organs. We excluded the classes 'prostate', as it is not part of our target classes, and 'bladder', which we believe to be incorrectly annotated in the AMOS scans. (Fig S5). The specific sequence and scanner types are not disclosed in the AMOS paper. The dataset is available here: https://zenodo.org/records/7262581.

## TotalSegmentator-MRI Dataset

TotalSegmentator-MRI (9) is a whole-body MRI segmentation model developed by D’Antonoli et al and is the successor of TotalSegmentator for CT images. Given the similar names of the two models, the image modality will be specified when referring to both models. The model was released together with an openly available dataset that consists of 298 MRI sequences, of which D’Antonoli et al marked 30 as test data. We excluded one test sequences (focused on the brain) due to incomplete header information. This dataset includes sequences from different MRI scanners, covering multiple body regions, and including different sequence types in axial, sagittal, and coronal planes. It contains annotations for 59 anatomic structures; for our analysis, we selected the 40 structures that correspond to the segmentation targets of MRSegmentator. The dataset is publicly available at https://zenodo.org/doi/10.5281/zenodo.11367004.

## Target Anatomical Structures

Our datasets comprise MR images spanning from the shoulders to the knees. The anatomic structures selected for segmentation were chosen based on their inclusion in the TotalSegmentator-CT target classes, visibility and delineability in MRI, clinical relevance, and consistent presence across the dataset. To improve segmentation robustness, anatomically similar structures were consolidated where MRI differentiation was challenging. For example, vertebrae were grouped into a single "spine" class, and lung lobes were merged due to poor tissue contrast and anisotropic voxel spacing. After this consolidation, 40 structures across five regions were finalized: the chest (heart, lungs, esophagus), gastrointestinal tract (liver, spleen, pancreas, gallbladder, stomach, intestines, colon), retroperitoneum (kidneys, adrenal glands, urinary bladder), musculoskeletal system (spine, sacrum, hips, femurs, gluteal and iliopsoas muscles), and vessels (aorta, vena cava, portal/splenic vein, iliac arteries/veins) (Fig 2).

## Annotation Strategy

We developed a four-stage human-in-the-loop annotation workflow to create high-quality segmentations.

1. Presegmentation: First, we generated initial segmentations by applying the TotalSegmentator-CT model (2) to the MRI scans. To improve performance of the TotalSegmentator-CT model on MRI scans, we used preprocessing steps including intensity inversion and histogram equalization (15). For UK Biobank data, we segmented water-only sequences and propagated labels to the remaining Dixon sequences. While some structures like kidneys required only minor corrections, others such as muscles and bones needed complete reannotation due to poor initial segmentation quality.

2. Manual annotation: One radiology resident with one (LX) and two radiologists with eight (KKB, LCA) years of experience in diagnostic radiology, refined and reviewed the presegmentations using MONAI Label (16) and 3D Slicer (17). Overall, 40 different classes were created, which are detailed in Figure 2.

3. Model training: An nnU-Net (10) model was repeatedly trained, each time 50 new MRI sequences were annotated, enabling the generation of more refined labels, which were reviewed and refined again by the radiologists. This was repeated until the full training dataset of 1,200 UKBB sequences and 221 in-house sequences was annotated. Once the annotation process was complete, we trained the final nnU-Net with fivefold cross-validation on the fully annotated images, resulting in the final MRSegmentator model. This final training was performed using the 3 d_fullres_no_flipping configuration of nnU-Net V2 (https://github.com/MIC-DKFZ/nnUNet) with an increased batch size of eight. Other training parameters were kept at default values.

4. Test data annotation: After the MRSegmentator model was trained, the test data were manually annotated by the radiology resident and quality checked and refined by one of two board-certified radiologists.

## Statistical Analysis

We assessed segmentation performance using three metrics: Dice Similarity Coefficient (DSC), 95th percentile Hausdorff Distance (HD), and a novel vessel consistency (VC) metric. DSC and HD were calculated per structure and sequence type, comparing model output against manual annotations. For thin, elongated structures like blood vessels, where DSC and HD can be misleading, we introduced VC as a complementary measure. VC is defined as the proportion of segmentations containing exactly one single connected component for a given class c:

$$VC(c) \equiv \frac{number\ of\ images\ with\ exactly\ one\ connected\ component\ c}{number\ of\ all\ images\ with\ c}$$

The VC provides additional information about anatomic plausibility of vessel segmentations, putting emphasis on the continuity of the vessel segmentation (Fig S2). We further quantified under-and over-segmentation relative to the reference standard for each NAKO MRI sequence. To assess the impact of including the CT scans to the training pool we trained a second model on MR images only and compare the results. To evaluate potential demographic effects on segmentation quality, we analyzed the relationship between DSC and participant characteristics in the NAKO dataset, which offered balanced demographics (25 men, 25 women). Sex-based differences in segmentation performance were assessed using independent two-sided *t* tests, while age-related effects were evaluated using Pearson's correlation coefficient, both implemented in SciPy version 1.10.0 (18). With 50 participants (1:1 male-to-female ratio), our study achieved a power of 0.41 to detect medium-sized sex-based differences (19).

At the time of writing only one openly available whole body MRI segmentation algorithm exists (TotalSegmentator-MRI), therefore, for model comparison, we evaluated MRSegmentator against TotalSegmentator-MRI on matching anatomic structures across all test datasets. All metrics are reported as mean ± SD unless otherwise specified. To adjust for multiple comparisons, we reported the false discovery rate (FDR) using the Benjamini-Hochberg procedure and an FDR < 0.05 was considered significant.

## Data Availability

To facilitate research applications and further development, we have made MRSegmentator's code and trained weights are publicly available at https://github.com//hhaentze/MRSegmentator.

# Results

## Participant Characteristics

We used three datasets for testing. Exclusion criteria are illustrated in Figure 1. The NAKO dataset included 30,868 participants, from which we randomly selected 50 (median age, 52.5 years; interquartile range (IQR): 12.75 years; 25 men, 25 women), yielding a total of 900 MRI sequences. The AMOS22 dataset included 60 MRI sequences and 300 CT scans from diverse patient populations (MRI: median age, 50 years; age range, 22–85 years; male-to-female ratio 1.2:1; CT: median age, 54 years; age range, 14–95 years; male-to-female ratio, 1.7:1). Individual-level demographic data were not available for AMOS22 so we report aggregated statistics as published in the original dataset description. The TotalSegmentator-MRI dataset contained 298 MRI sequences, from which we selected 30 sequences assigned to the test set by D'Antonoli et al.; one was excluded due to missing header information, resulting in 29 MRI sequences (median age, 60.5 years; IQR, 23.75 years; 9 men, 11 women, age and gender unknown for 9 scans). Detailed participant demographics are provided in Table 1.

## Segmentation Performance

On NAKO data, MRSegmentator achieved averaged DSCs ranging from 0.85 ± 0.08 for T2-HASTE sequences to 0.91 ± 0.05 for T1-weighted Dixon in-phase sequences (Table 2). Performance on the AMOS dataset yielded mean DSC scores of 0.79 ± 0.11 for MRI and 0.84 ± 0.12 for CT. The highest DSCs were observed for well-defined organs (lungs: 0.96, heart: 0.94, Fig 3) and organs with anatomic variability (liver: 0.96, right kidney: 0.95, left kidney: 0.93). Small structures proved most challenging, particularly the portal/splenic vein (0.64) and adrenal glands (0.56 for AMOS).

Vessel Consistency analysis of NAKO GRE examinations demonstrated high reliability for major vessels. The aorta and inferior vena cava were consistently segmented as single connected structures (VC: 100% (200/200) and 92% (184/200), respectively; Table 3). Iliac veins showed higher consistency (left: 0.95 (189/200), right: 0.85(170/200)) than arteries (left: 0.69 (137/200), right: 0.60 (119/200)), while the portal/splenic vein typically appeared as multiple components (VC: 0.40 (79/200)). A second model, which we trained without CT scans, showed equal performance for NAKO MRI but reduced DSC scores in the AMOS CT set (0.59 vs 0.84). Demographic analysis revealed superior segmentation quality in males (DSC = 0.89 ± 0.02) compared with females (DSC = 0.87 ± 0.02) in NAKO GRE sequences (*P* = .009), with the largest differences in adrenal glands (ΔDSC = 0.13/0.10) and duodenum (ΔDSC = 0.11). Participant age positively correlated with DSC (r = 0.37, FDR = 0.009; Fig S6).

## Comparison with TotalSegmentator-MRI

Across NAKO and AMOS22 datasets, MRSegmentator outperformed TotalSegmentator-MRI (NAKO: 0.91 ± 0.05 vs 0.83 ± 0.07; AMOS: 0.79 ± 0.11 vs 0.75 ± 0.12; FDR < 0.001 each) and lower HD values (NAKO: 7.5 ± 14.8 mm vs 15.1 ± 20.3 mm; AMOS: 8.4 ± 7.0 mm vs 10.0 ± 8.7 mm; FDR < 0.001 each) for all classes. The performance difference was most pronounced in abdominal organs (liver: 0.96 vs 0.89, spleen: 0.91 vs 0.84, pancreas: 0.82 vs 0.73) and blood vessels (aorta: 0.93 vs 0.85, inferior vena cava: 0.86 vs 0.78) (Fig S4).

On the TotalSegmentator-MRI test set, overall DSCs were comparable (0.74 ± 0.21 vs 0.75 ± 0.22, FDR = 0.017), with MRSegmentator excelling in abdominal organ and vessel segmentation, while TotalSegmentator-MRI demonstrated better performance in musculoskeletal structures.

## Failure Cases and Sequence-specific Performance

Despite robust overall performance, we could identify specific failure patterns. Left-right confusion occurred occasionally in the pelvic region, evidenced by large HDs for femurs (left: 37.8 mm, right: 18.9 mm) and left iliopsoas muscle (left: 19.6 mm, right 6.0 mm). In the kidney tumor subset, MRSegmentator accurately segmented kidneys with tumors larger than 7 cm (*n* = 8, Fig 4), though some oversegmentation occurred with irregular tumor borders. The model showed inconsistent performance in postnephrectomy cases, correctly identifying single kidneys in validation data but occasionally misclassifying colon as kidney in test data. On average, the model's errors were more often due to undersegmentation than oversegmentation (Table 4). Segmentation failures were least frequent in in-phase sequences. In water-only sequences, the model tended to oversegment, while in fat-only sequences, it often failed to capture structures completely. The highest variability was observed in T2-Haste, where undersegmentation fraction quartiles ranged from 0.07 to 0.29 (Fig 5).

## Discussion

This retrospective study addressed the lack of comprehensive whole-body MRI segmentation tools by developing MRSegmentator, an nnU-Net-based cross-modality model trained on CT and MRI to segment 40 anatomic structures. Evaluated on three external datasets, MRSegmentator achieved mean Dice Similarity Coefficients (DSCs) of 0.91 ± 0.05 on NAKO Dixon in-phase sequences and 0.79 ± 0.11 on AMOS22 MRI, outperforming the only other available whole-body MRI model (TotalSegmentator-MRI DSC 0.83 ± 0.07 on NAKO and 0.75 ± 0.12 on AMOS22; FDR < 0.001). Vessel consistency exceeded 92% for major vessels, and objective cross-modality training yielded robust generalizability.

Comparison with existing methods shows that MRSegmentator achieves competitive performance against specialized single-organ models. Our spleen and liver segmentation accuracy (DSC: 0.95, 0.96) matches dedicated models (0.96, 0.95) (20,21), while multiorgan performance for abdominal structures performance (DSC for liver, spleen, kidneys, pancreas: 0.96, 0.91, 0.94, 0.82 for NAKO MRI; 0.96, 0.95, 0.95, 0.81 for AMOS MRI) approaches that of recent organ-specific approaches, such as reported by Kart et al (0.98, 0.96, 0.98, 0.89) (5). Given the widespread adoption of the nnU-Net architecture, observed performance variations are primarily attributable to differences in training data. Incorporating CT scans into the training pipeline improved segmentation accuracy on CT images (DSC 0.59 to 0.84) but did not demonstrably affect MR image segmentation quality. We could not observe evidence that multimodality training improves single-modality inference, as reported in studies like the AMOS challenge (11). Notably, MRSegmentator's ability to process both MRI and CT images with a single model (CT DSC: 0.84), achieving segmentation quality on par with TotalSegmentator-CT, the current state of the art in multiorgan segmentation in CT, offers practical advantages over modality-specific solutions, potentially streamlining clinical and research workflows. This versatility may outweigh slightly inferior performance compared with specialized models.

Since the initial submission of this work, new whole-body MRI segmentation models have emerged, including TotalSegmentator-MRI. On the NAKO dataset, MRSegmentator achieves higher DSC values across all structures and sequences, though annotation bias may contribute to this advantage, as the same radiologists annotated both the training and test data. However, on the AMOS dataset, where no such bias exists, MRSegmentator still outperforms across all classes. When tested on the TotalSegmentator-MRI dataset, where annotation practices may favor their model, MRSegmentator maintains a comparable overall DSC (0.74 vs 0.75) while demonstrating superior segmentation performance for 17 structures, particularly abdominal organs and blood vessels.

That said, TotalSegmentator-MRI supports a broader range of structures, particularly femoral muscle groups such as the quadriceps femoris and sartorius. For studies focused on musculoskeletal anatomy, it thus provides valuable additional segmentations. Rather than declaring one model superior, these results highlight that MRSegmentator excels in abdominal segmentation, while TotalSegmentator-MRI offers a broader selection of muscle classes. Both models have distinct strengths depending on the research application.

To address the limitations of the voxel-based metrics DSC and HD, we introduced the vessel consistency metric, which revealed that large vessels such as the aorta are consistently segmented as single structures (VC: 1.0), whereas smaller vessels such as the portal/splenic vein often appear fragmented (VC: 0.40). This metric provides insights not captured by traditional DSC measurements, particularly relevant for elongated structures. We observed gender differences in performance, with higher accuracy in male subjects (ΔDSC = 0.02, FDR = 0.009), likely reflecting anatomic differences in muscle mass and organ positioning rather than technical limitations. Both the TotalSegmentator-CT and in-house datasets are male-dominated and have a more diverse participant pool than our UKBB data. On the other hand, metrics like DSC tend to be less accurate for smaller structures, and women generally have smaller body volumes, which may also explain the observed differences. While we highlight these metric differences, further research is needed to assess their clinical relevance. The model showed robust performance on pathologic cases, successfully segmenting kidneys with large tumors, while occasionally struggling with postoperative anatomy. Training models to detect rare pathologies and missing structures is challenging due to limitations in available data. One potential solution is to integrate synthetic MRI data designed to include these specific structures (22). The model occasionally confused left and right structures, likely due to nnU-Net's patch-wise design limiting global context. A whole-image postprocessing step could help mitigate this but would need to address edge cases and metadata inconsistencies.

MRSegmentator will allow researchers to obtain biomarkers relevant for various research questions and clinical tasks. For instance, total kidney volume has been shown to correlate with glomerular filtration rate, a key indicator in polycystic kidney disease (23). Additionally, iliac artery tortuosity may serve as a predictor of biologic age (24), while the fat fraction within the autochthonous muscles can assist in stratifying the risk of incidental, non-traumatic vertebral fractures in the lower thoracic spine among elderly patients (25).

Our study had several limitations. First, the human in the loop may have introduced annotation bias, however strong performance on fully independent external datasets suggests minimal impact. Second, the UK Biobank training data, while numerous in sequences (1,200), represents only 50 unique participants, potentially limiting anatomic variety. This limitation is partially mitigated by including diverse in-house and TotalSegmentator-CT data. Third, the observed gender-based performance differences highlight the need for more balanced training datasets.

In conclusion, objective cross-modality segmentation with MRSegmentator provided reproducible, high-accuracy delineation of whole-body anatomy on MRI. The capability of segmenting CT and MR images, makes it a valuable tool for researchers and clinicians. We will continue to work on MRSegmentator by focusing on expanding the range of supported anatomic structures and pathologic conditions while maintaining the model's cross-modality capabilities.

**Author affiliations:**

[1] Department of Radiology, Charité - Universitätsmedizin Berlin corporate member of Freie Universität Berlin and Humboldt Universität zu Berlin, Berlin, Germany

[2] Diagnostic Image Analysis Group, Radboud University Medical Center, Nijmegen, the Netherlands

[3] Department of Diagnostic and Interventional Radiology, School of Medicine and Health, Klinikum rechts der Isar, TUM University Hospital, Technical University of Munich, Munich, Germany

[4] Athinoula A. Martinos Center for Biomedical Imaging, Massachusetts General Hospital and Harvard Medical School, Charlestown, Mass

[5] Institute of Radiology, University Hospital Erlangen, Friedrich-Alexander-Universität (FAU) Erlangen-Nürnberg, Erlangen, Germany

[6] Laboratory for Computer Aided Medical Procedures, Technical University of Munich, Munich, Germany

[7] Chair for AI in Medicine and Healthcare, Klinikum rechts der Isar, Technical University Munich, Munich, Germany

[8] Department of Computing, Imperial College London, London, UK

[9] Institute for Advanced Study, Technical University Munich, Munich, Germany

[10] Artificial Intelligence in Medicine (AIM) Program, Mass General Brigham, Harvard Medical School, Boston, Mass

[11] Departments of Radiation Oncology and Radiology, Dana-Farber Cancer Institute and Brigham and Women's Hospital, Boston, Mass

[12] Radiology and Nuclear Medicine, CARIM & GROW, Maastricht University, Maastricht, the Netherlands

[13] Department of Diagnostic and Interventional Radiology, University Hospital Aachen, Aachen, Germany

[14] Department of Diagnostic and Interventional Radiology, Medical Center - University of Freiburg, Faculty of Medicine, University of Freiburg, Freiburg, Germany

[15] Berlin Ultrahigh Field Facility (B.U.F.F.), Max Delbrück Center for Molecular Medicine in the Helmholtz Association, Berlin, Germany

[16] Clinic for Diagnostic and Interventional Radiology, Heidelberg University Hospital, Heidelberg, Germany

[17] Department of Diagnostic and Interventional Radiology and Neuroradiology, Universitätsklinikum Augsburg, Augsburg, Germany

[18] Institute for Community Medicine, University Medicine Greifswald, Greifswald, Germany

[19] Experimental Clinical Research Center, Charité - Universitätsmedizin Berlin corporate member of Freie Universität Berlin and Humboldt Universität zu Berlin, Berlin, Germany

[20] Division of Medical Image Computing, German Cancer Research Center, Heidelberg, Germany

[21] Department of Cardiovascular Radiology and Nuclear Medicine, School of Medicine and Health, German Heart Center, TUM University Hospital, Technical University of Munich, Lazarettstr 36, 80636 Munich, Germany

Received XXX; revision requested XXX; revision received XXX; accepted XXX.

**Address correspondence to:** K.B. (email: keno.bressem@tum.de).

**Funding:** This work was in large parts funded by the Wilhelm Sander Foundation. It was also 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. The authors acknowledge the Scientific Computing of the IT Division at the Charité - Universitätsmedizin Berlin for providing computational resources that have contributed to the research results reported in this paper.

**Author contributions:** Guarantors of integrity of entire study, **H.H., L.X., F. Busch, A.K., M.R.M., L.C.A., K.K.B.**; study concepts/study design or data acquisition or data analysis/interpretation, all authors; manuscript drafting or manuscript revision for important intellectual content, all authors; approval of final version of submitted manuscript, all authors; agrees to ensure any questions related to the work are appropriately resolved, all authors; literature research, **H.H., L.X., F. Busch, N.B., N.N., D.T., F. Bamberg, S.R., T. Niendorf, T. Nonnenmacher, J.S.M., K.M.H., B.v.G., L.C.A., XXX**; clinical studies, **L.X., F. Busch, A.K., S.Z., T.P., H.U.K., T.K., H.V., J.S.M., L.C.A., XXX**; experimental studies, **H.H., L.D., F. Busch, F. Bamberg, J.W., B.v.G., L.C.A., XXX**; statistical analysis, **H.H., L.X., F. Busch, J.W., XXX**; and manuscript editing, **H.H., L.X., C.J.M., L.D., F. Busch, S.Z., N.B., N.N., D.R., J.S., H.J.W.L.A., D.T., F. Bamberg, J.W., C.L.S., S.R., T. Niendorf, T.P., H.U.K., H.V., J.S.M., K.M.H., A.H., M.P., M.R.M., L.C.A., K.K.B.**

**Acknowledgments:** This research has been conducted using the UK Biobank Resource under Application Number 105529. 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 (https://www.charite.de/en/research/research_support_services/research_infrastructure/science_it/#c30646061). This work was in large parts funded by the Wilhelm Sander Foundation. 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.

**Disclosures of conflicts of interest: H.H.** No relevant relationships. **L.X.** No relevant relationships. **C.J.M.** No relevant relationships. **F.J.D.** No relevant relationships. **L.D.** No relevant relationships. **F. Busch** No relevant relationships. **A.K.** No relevant relationships. **S.Z.** No relevant relationships. **N.B.** No relevant relationships. **N.N.** No relevant relationships. **D.R.** Research grants for unrelated projects, EPSRC, Wellcome Trust, ERC, H2020, DFG, BMBF, Alexander von Humboldt Foundation, Bavarian Ministry for Arts and Sciences, Roche, Siemens. **J.S.** No relevant relationships. **H.J.W.L.A.** No relevant relationships. **D.T.** No relevant relationships. **F. Bamberg** Unrestricted research grants, Bayer Healthcare, Siemens Healthineers; consulting fees, Bayer Healthcare; speakers bureau, Bayer Healthcare, Siemens Healthineers; board member, German Roentgen Ray Society. **J.W.** No relevant relationships. **C.L.S.** Grants or contracts, Siemens Healthineers; honoraria, Bayer Healthcare, Siemens Healthineers. **S.R.** No relevant relationships. **T. Niendorf** No relevant relationships. **T.P.** Member of the NAKO board of directors (unpaid position). **H.U.K.** Grants to institution, Siemens, Philips, Boehringer Ingelheim; honoraria, Siemens, Philips, Boehringer Ingelheim; DSMB or advisory board, Median, ContextFlow. **T. Nonnenmacher** No relevant relationships. **T.K.** Research support, Siemens Healthineers; honoraria, Abbott Vascular, Sirtex Medical, Canon Medical Systems, AstraZeneca, BRACCO Suisse; support for meeting/travel, Cardiovascular and Interventional Radiological Society of Europe (CIRSE), German Local Radiological Societies; CIRSE Executive Committee member. **H.V.** No relevant relationships. **J.S.M.** No relevant relationships. **K.M.H.** No relevant relationships. **A.H.** No relevant relationships. **M.P.** Grants or contracts to institution, Koningin Wilhelmina Fonds (KWF) AMARA (KWF 2021-14113), NELSON POP (KWF 2021- 9037), European Commission SOLACE (EU4 Health Programme 101101187), Dutch Government COVID-Climate (ZonMW 10430102110004); royalties to institution, Canon Medical Systems; speakers bureau, Canon Medical Systems, Siemens Healthineers, Bracco SA; patents planned issued or pending, Radboud University Medical Center Similarity filter (US 10,902,562 B2), Radboudumc, Canon Medical Systems Multiphase filter (US 12,086,979 B2); vice chair and advisory board, Fraunhofer Mevis, Bremen, Germany; ESR, 2nd Vice Chairman since 2024, Vice Chairman since 2025; recipient of equipment, Prototype photon counting CT scanner. **B.v.G.** No relevant relationships. **M.R.M.** No relevant relationships. **L.C.A.** Former member of the *Radiology: Artificial Intelligence* trainee editorial board. **K.K.B.** Grants or contracts, Bayern Innovativ, German Federal Ministry of Education and Research, Max Kade Foundation and Wilhelm-Sander Foundation; honoraria, Canon Medical Systems Corporation and GE HealthCare; advisor for the EU Horizon 2020 LifeChamps project (875329) and the EU IHI Project IMAGIO (101112053); member of the *Radiology: Artificial Intelligence* trainee editorial board.

## References

1. Gillies RJ, Kinahan PE, Hricak H. Radiomics: Images Are More than Pictures, They Are Data. Radiology 2016;278(2):563–577.

2. Wasserthal J, Breit HC, Meyer MT, et al. Totalsegmentator: Robust segmentation of 104 anatomic structures in CT images. Radiol Artif Intell 2023;5(5):e230024.

3. Litjens G, Kooi T, Bejnordi BE, et al. A survey on deep learning in medical image analysis. Med Image Anal 2017;42:60–88.

4. Langner T, Hedström A, Mörwald K, et al. Fully convolutional networks for automated segmentation of abdominal adipose tissue depots in multicenter water–fat MRI. Magn Reson Med 2019;81(4):2736–2745.

5. Kart T, Fischer M, Küstner T, et al. Deep learning-based automated abdominal organ segmentation in the UK Biobank and German National Cohort Magnetic Resonance Imaging Studies. Invest Radiol 2021;56(6):401–408.

6. Adams LC, Makowski MR, Engel G, et al. Prostate158-An expert-annotated 3T MRI dataset and algorithm for prostate cancer detection. Comput Biol Med 2022;148:105817.

7. Zhou A, Liu Z, Tieu A, et al. MRAnnotator: multi-Anatomy and many-Sequence MRI segmentation of 44 structures. arXiv 2024. Preprint posted online February 1, 2024; doi:10.48550/arxiv.2402.01031.

8. Zhuang Y, Mathai TS, Mukherjee P, et al. MRISegmentator-Abdomen: A Fully Automated Multi-Organ and Structure Segmentation Tool for T1-weighted Abdominal MRI. arXiv 2024. Preprint posted online May 9, 2024; doi:10.48550/arxiv.2405.05944.

9. D’Antonoli TA, Berger LK, Indrakanti AK, et al. TotalSegmentator MRI: Robust Sequence-independent Segmentation of Multiple Anatomic Structures in MRI. arXiv 2024. Preprint posted online May 29, 2024; doi:10.48550/arxiv.2405.19492.

10. Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat Methods 2021;18(2):203–211.

11. Ji Y, Bai H, Ge C, et al. AMOS: A Large-Scale Abdominal Multi-Organ Benchmark for Versatile Medical Image Segmentation. arXiv 2022. Preprint posted online June 16, 2022; doi:10.48550/arxiv.2206.08023.

12. Littlejohns TJ, Holliday J, Gibson LM, et al. The UK Biobank imaging enhancement of 100,000 participants: rationale, data collection, management and future directions. Nat Commun 2020;11(1):2624.

13. German National Cohort (GNC) Consortium. The German National Cohort: aims, study design and organization. Eur J Epidemiol 2014;29(5):371–382.

14. Bamberg F, Kauczor HU, Weckbach S, et al; German National Cohort MRI Study Investigators. Whole-body MR imaging in the German National Cohort: rationale, design, and technical background. Radiology 2015;277(1):206–220.

15. Häntze H, Xu L, Donle L, et al. Improve Cross-Modality Segmentation by Treating MRI Images as Inverted CT Scans. arXiv 2024. Preprint posted online May 4, 2024; doi:10.48550/arxiv.2405.03713.

16. Diaz-Pinto A, Alle S, Nath V, et al. Monai label: A framework for ai-assisted interactive labeling of 3d medical images. Med Image Anal 2024;95:103207.

17. Fedorov A, Beichel R, Kalpathy-Cramer J, et al. 3D Slicer as an image computing platform for the Quantitative Imaging Network. Magn Reson Imaging 2012;30(9):1323–1341.

18. Virtanen P, Gommers R, Oliphant TE, et al; SciPy 1.0 Contributors. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods 2020;17(3):261–272.

19. Faul F, Erdfelder E, Lang AG, Buchner A. G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behav Res Methods 2007;39(2):175–191.

20. Sharbatdaran A, Romano D, Teichman K, et al. Deep learning automation of kidney, liver, and spleen segmentation for organ volume measurements in autosomal dominant polycystic kidney disease. Tomography 2022;8(4):1804–1819.

21. Hossain MSA, Gul S, Chowdhury MEH, et al. Deep Learning Framework for Liver Segmentation from T 1-Weighted MRI Images. Sensors (Basel) 2023;23(21):8890.

22. Lei W, Chen H, Zhang Z, et al. A Data-Efficient Pan-Tumor Foundation Model for Oncology CT Interpretation. arXiv 2025. Preprint posted online February 10, 2025; doi:10.48550/arxiv.2502.06171.

23. Jo WR, Kim SH, Kim KW, et al. Correlations between renal function and the total kidney volume measured on imaging for autosomal dominant polycystic kidney disease: A systematic review and meta-analysis. Eur J Radiol 2017;95:56–65.

24. Mach M, Poschner T, Hasan W, et al. The Iliofemoral tortuosity score predicts access and bleeding complications during transfemoral transcatheter aortic valve replacement: Data from the VIenna Cardio Thoracic aOrtic valve registrY (VICTORY). Eur J Clin Invest 2021;51(6):e13491.

25. Backhauß JC, Jansen O, Kauczor HU, Sedaghat S. Fatty Degeneration of the Autochthonous Muscles Is Significantly Associated with Incidental Non-Traumatic Vertebral Body Fractures of the Lower Thoracic Spine in Elderly Patients. J Clin Med 2023;12(14):4565.

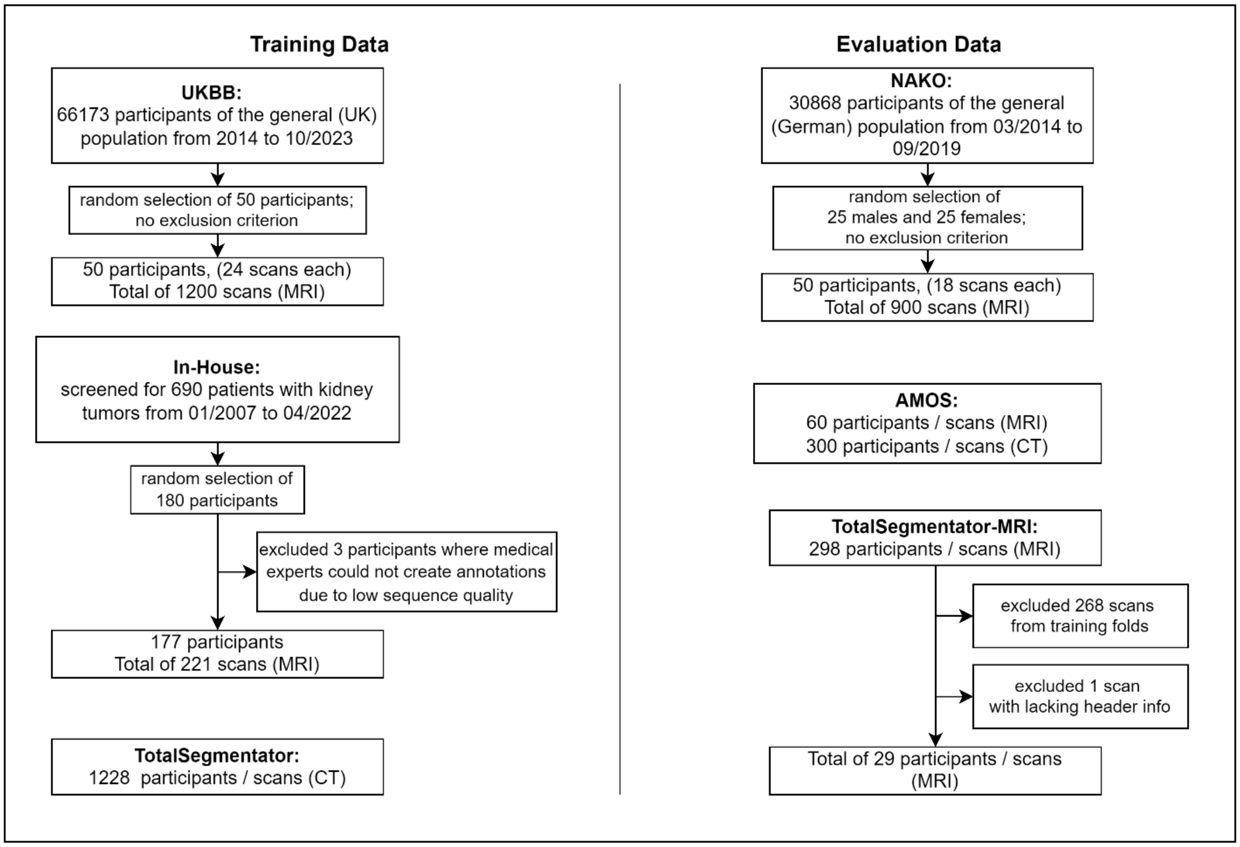

Training Data
UKBB:
66173 participants of the general (UK) population from 2014 to 10/2023
random selection of 50 participants; no exclusion criterion
50 participants, (24 scans each) Total of 1200 scans (MRI)
In-House:
screened for 690 patients with kidney tumors from 01/2007 to 04/2022
random selection of 180 participants
excluded 3 participants where medical experts could not create annotations due to low sequence quality
177 participants Total of 221 scans (MRI)
TotalSegmentator:
1228 participants / scans (CT)
Evaluation Data
NAKO:
30868 participants of the general (German) population from 03/2014 to 09/2019
random selection of 25 males and 25 females; no exclusion criterion
50 participants, (18 scans each) Total of 900 scans (MRI)
AMOS:
60 participants / scans (MRI)
300 participants / scans (CT)
TotalSegmentator-MRI:
298 participants / scans (MRI)
excluded 268 scans from training folds
excluded 1 scan with lacking header info
Total of 29 participants / scans (MRI)

**Figure 1:** Study workflow with training and evaluation data. For training, data were obtained from the UK Biobank (UKBB), an in-house dataset and the TotalSegmentator-CT dataset. For evaluation data were used from the National German Cohort (NAKO), the Abdominal Multi-Organ Segmentation challenge (AMOS) and the TotalSegmentator-MRI dataset.

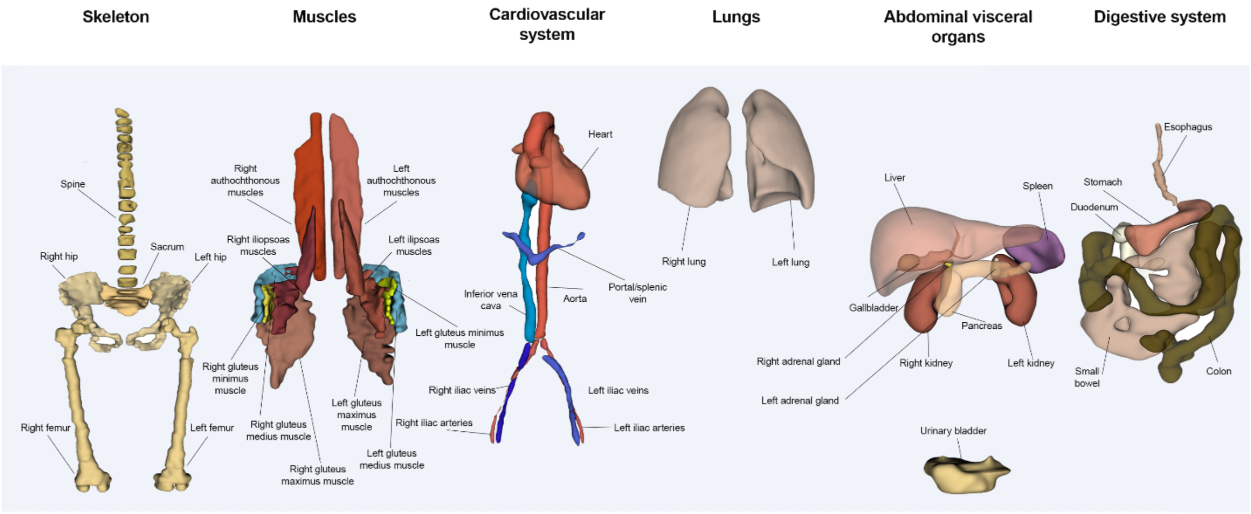


**Figure 2:** Sample segmentation output of MRSegmentator for all 40 classes. This includes spine, sacrum, hips, femurs, heart, aorta, inferior vena cava, portal/splenic vein, iliac arteries and veins, left and right lungs, liver, spleen, pancreas, gallbladder, stomach, duodenum, small bowel, colon, left and right kidneys, adrenal glands, urinary bladder, and muscles, specifically, gluteal muscles, autochthonous muscles, iliopsoas muscles. The model was trained on diverse datasets including UK Biobank, in-house clinical data, and CT scans, using a human-in-the-loop annotation approach. It demonstrates robust performance across various MRI sequences and can also segment CT images.

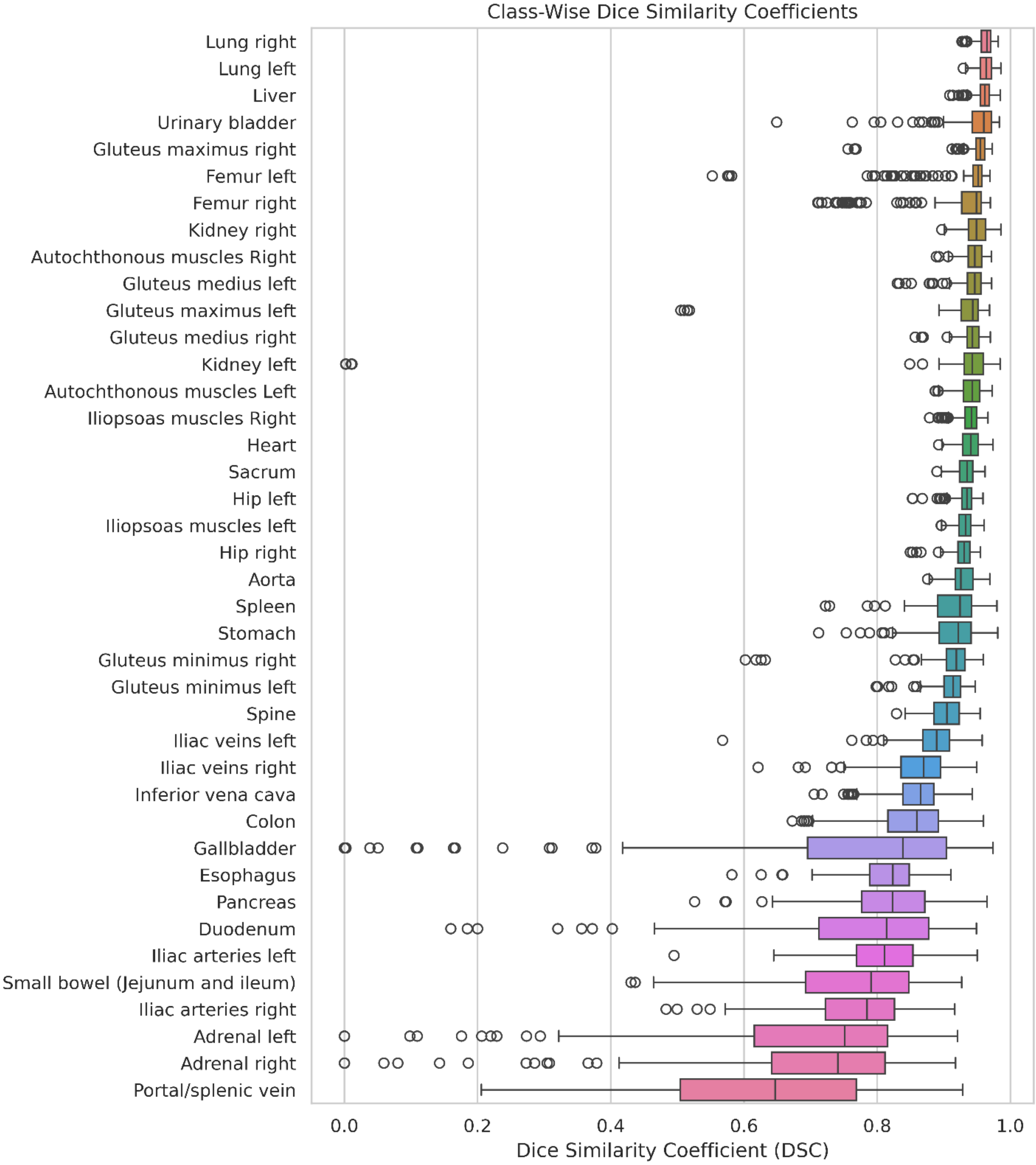


**Figure 3:** Class-wise Dice Similarity Coefficients (DSC) obtained by the MRSegmentator on the four gradient echo modalities of the dataset from the German National Cohort (NAKO). The box plots show the distribution of DSC values for each of the 40 anatomic structures segmented by the model. The boxes represent the interquartile range (IQR), with the median DSC marked by the horizontal line within each box. The whiskers extend to the minimum and maximum

values within 1.5 times the IQR, and any outliers beyond this range are represented by single points. The left kidney segmentations with a DSC of zero are false positives for a single patient who's left kidney was removed.

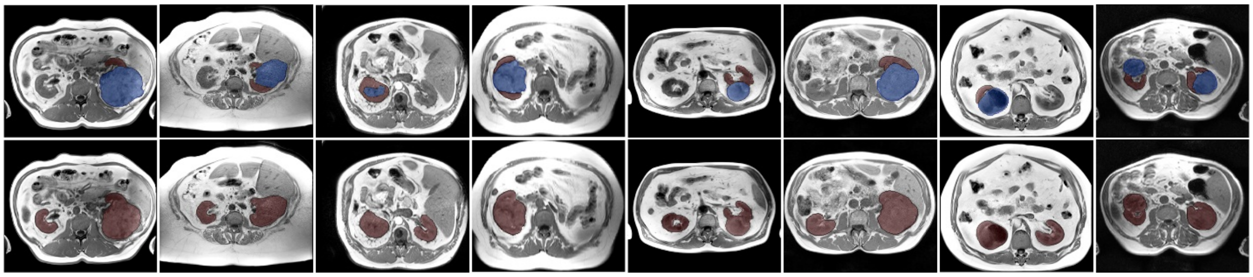


**Figure 4:** The first row depicts manually annotated kidneys (red) and tumors (blue) in eight different MRI scans. The second row shows the corresponding kidney segmentations generated by MRSegmentator. The model accurately localizes and segments the kidneys even in the presence of large tumors, demonstrating its robustness in handling pathologic cases. In the fourth sample, the missing right kidney due to a previous nephrectomy is correctly not segmented by the model.

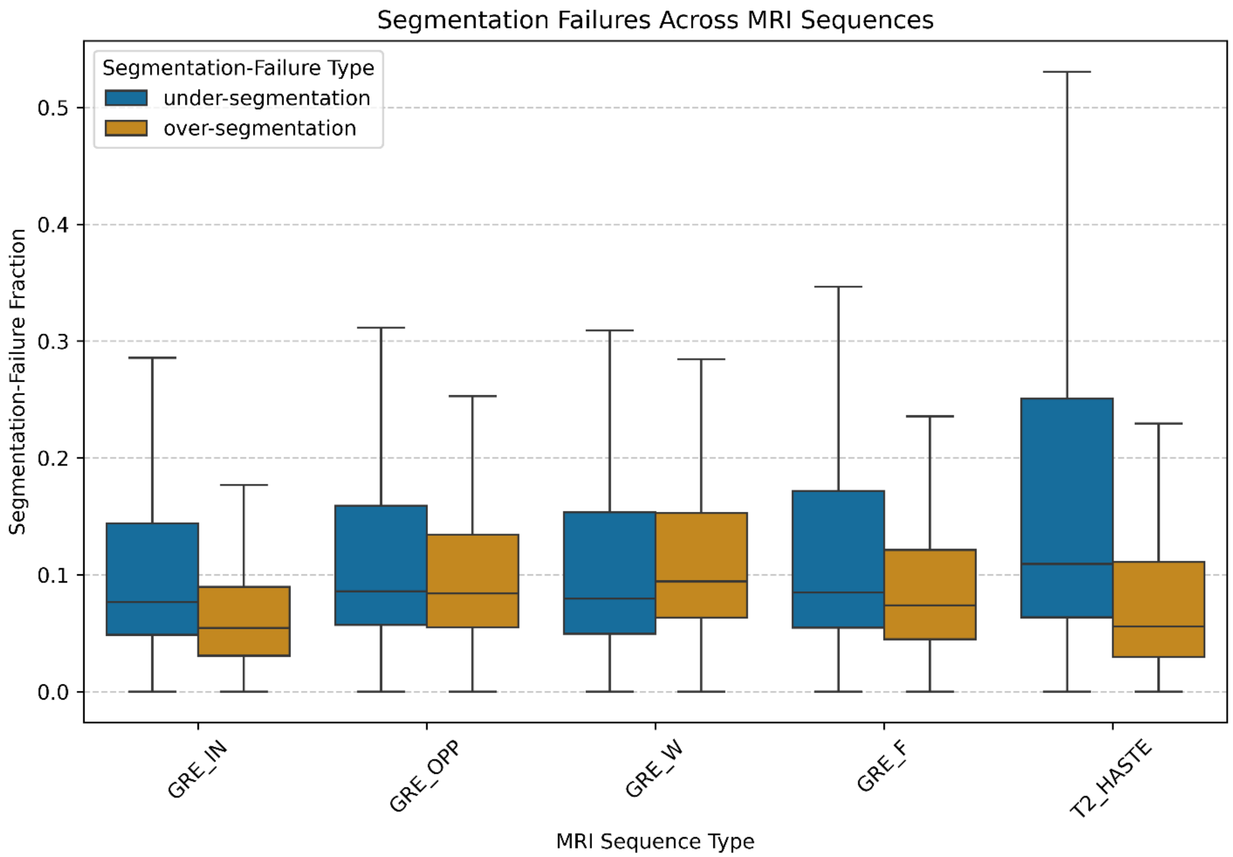


**Figure 5:** Segmentation failure types across different MRI sequences from the German National Cohort (NAKO). The box plots illustrate the distribution of failure fractions for 40 target structures across 50 selected participants from the NAKO dataset, categorized by MRI sequence type. Segmentation errors are measured relative to the reference standard, distinguishing between under-segmentation and over-segmentation. Under-segmentation is generally more prevalent, except in water-only sequences, where over-segmentation is more common. The highest under-segmentation rates are observed in T2-HASTE and fat-only sequences. Abbreviations: GRE: Gradient Echo, T2 HASTE: T2-weighted Half-Fourier Acquisition Single-shot Turbo spin Echo.

**Table 1: Data Composition of Training and Test Sets**

| | In-House | UKBB | TotalSegmentator-CT |
|---|---|---|---|
| Nr. Participants | 177 (M:121, F:56) | 50 (M:19, F: 31) | 1228 (M: 716, F:510) |
| Nr. Scans | 221(M:150, F:71) | 1200 (M:456, F: 744) | 1228 (M: 716, F:510) |
| Age [years] | 37–83 (median = 62) | 40-69[1] | 15–98 (median = 65) |
| Scanner Types | 1.5 and 3 Tesla MRI | 1.5 Tesla MRI | 20 different models |

| | | | |
|---|---|---|---|
| Sequences<br>Test Data | T1 ($n = 90$)<br>T2fs ($n = 64$)<br>T1fs ($n = 67$) | IN ($n = 300$)<br>OPP ($n = 300$)<br>W ($n = 300$)<br>F ($n = 300$) | CT ($n = 1228$) |
| | NAKO (MRI) | AMOS[2] | TotalSegmentator-MRI[3] |
| Nr. Participants | 50 (M:25, F:25) | MRI: 60 (M: 55, F: 45) CT: 300 (M: 314, F: 186) | - |
| Nr. Scans | 900 (M:450, F:450) | MRI: 60 (M: 55, F: 45) CT: 300 (M: 314, F: 186) | 29 (M: 9, F:11, n.a.: 9) |
| Age [years] | 26–69 (median = 52.5) | MRI: 22–85 (median = 50) CT: 14–94 (median = 54) | 14–78 (median 60.5) |
| Scanner Types | 3 Tesla MR | MRI: 3 different models CT: 5 different models | 13 different models |
| Sequences | T1 GRE IN ($n = 200$)<br>T1 GRE OPP ($n = 200$)<br>T1 GRE W ($n = 200$)<br>T1 GRE F ($n = 200$)<br>T2 HASTE ($n = 100$) | MRI ($n = 60$)<br>CT ($n = 300$) | MRI ($n = 29$) |

| | Training Data | | | Test Data | | |
|---|---|---|---|---|---|---|
| | In-House | UKBB | TotalSegmentator-CT | NAKO (MRI) | AMOS[2] | TotalSegmentator-MRI[3] |
| Nr. Participants | 177 (M:121, F:56) | 50 (M:19, F: 31) | 1228 (M: 716, F:510) | 50 (M:25, F:25) | MRI: 60 (M: 55, F: 45) CT: 300 (M: 314, F: 186) | NA |
| Nr. Scans | 221(M:150, F:71) | 1200 (M:456, F: 744) | 1228 (M: 716, F:510) | 900 (M:450, F:450) | MRI: 60 (M: 55, F: 45) CT: 300 (M: 314, F: 186) | 29 (M: 9, F:11, n.a.: 9) |
| Age [years] | 37–83 (median = | 40-69[1] | 15–98 (median = 65) | 26–69 (median = | MRI: 22–85 (median = 50) | 14–78 (median 60.5) |

| | | | | | | |
|---|---|---|---|---|---|---|
| | 62) | | | 52.5) | CT: 14–94 (median = 54) | |
| Scanner Types | 1.5 and 3 Tesla MRI | 1.5 Tesla MRI | 20 different models | 3 Tesla MR | MRI: 3 different models CT: 5 different models | 13 different models |
| Sequences | T1 (*n* = 90)<br>T2fs (*n* = 64)<br>T1fs (*n* = 67) | IN (*n* = 300)<br>OPP (*n* = 300)<br>W (*n* = 300)<br>F (*n* = 300) | CT (*n* = 1228) | T1 GRE IN (*n* = 200)<br>T1 GRE OPP (*n* = 200)<br>T1 GRE W (*n* = 200)<br>T1 GRE F (*n* = 200)<br>T2 HASTE (*n* = 100) | MRI (*n* = 60)<br>CT (*n* = 300) | MRI (*n* = 29) |

Note.—(1) We did not have the social-demographics metadata of the UK Biobank data. Generally, the UK Biobank targets healthy individuals between the ages of 40 and 69. (2) Age and sex of the participants of the dataset from the Abdominal Multi-Organ Segmentation challenge (AMOS) are given for all 600 images, of which 360 are publicly available with manual annotations. Specific sequence and scanner types are not disclosed in the AMOS paper. (3) The TotalSegmentator-MRI paper reports their demographic information on an examination level, not a participant level. Abbreviations: UKBB: UK Biobank, NAKO: German National Cohort, GRE: Gradient Echo, T2 HASTE: T2-weighted Half-Fourier Acquisition Single-shot Turbo spin Echo.

**Table 2: Comparison of DSC and HD between MRSegmentator and TotalSegmentator-MRI**

| Dataset | | DSC | HD [mm] | |
|---|---|---|---|---|
| | MRSeg[1] | TotalSegMRI[2] | MRSeg[1] | TotalSegMRI[2] |
| NAKO in-phase (40 classes) | 0.91 ± 0.05 | 0.83 ± 0.07 | 7.5 ± 14.8 | 15.1 ± 20.3 |
| NAKO opposed-phase (40 classes) | 0.88 ± 0.05 | 0.82 ± 0.06 | 8.4 ± 15.0 | 12.7 ± 17.4 |
| NAKO water only (40 classes) | 0.87 ± 0.05 | 0.81 ± 0.06 | 8.5 ± 15.0 | 13.1 ± 16.3 |

Radiology: Artificial Intelligence

Just Accepted papers have undergone full peer review and have been accepted for publication. This article will undergo copyediting, layout, and proof review before it is published in its final version. Please note that during production of the final copyedited article, errors may be discovered which could affect the content.

| | | | | |
|---|---|---|---|---|
| NAKO fat only (40 classes) | 0.88 ± 0.06 | 0.80 ± 0.07 | 8.3 ± 14.8 | 15.6 ± 20.6 |
| TotalSegmentator-MRI data (40 classes) | 0.74 ± 0.21 | 0.75 ± 0.22 | 16.1 ± 21.7 | 15.9 ± 21.5 |
| NAKO T2 HASTE (24 classes) | 0.85 ± 0.08 | 0.75 ± 0.10 | 8.5 ± 7.0 | 14.1 ± 9.5 |
| NAKO in-phase (24 classes)[3] | 0.90 ± 0.06 | 0.82 ± 0.08 | 6.2 ± 7.2 | 11.3 ± 8.1 |
| AMOS MRI (13 classes) | 0.79 ± 0.11 | 0.75 ± 0.12 | 8.4 ± 7.0 | 10.0 ± 8.7 |
| AMOS CT (13 classes)[4] | 0.84 ± 0.12 | 0.84 ± 0.11 | 8.6 ± 11.7 | 8.1 ± 10.2 |
| NAKO in-phase (13 classes)[3] | 0.88 ± 0.09 | 0.78 ± 0.12 | 4.3 ± 3.6 | 10.1 ± 6.0 |

We compared (1) MRSegmentator and (2) the TotalSegmentator-MRI model on data from the German National Cohort (NAKO), the Abdominal Multi-Organ Segmentation challenge (AMOS) and the TotalSegmentator-MRI dataset. We reported the Dice Similarity Coefficient (DSC), where larger values indicate better performance, and Hausdorff-distance (HD), where smaller values indicate better performance. Both metrics are accompanied by their standard deviation. Significant higher values are highlighted in bold. All false discovery rates (FDR) were smaller than 0.001 except the comparisons on the TotalSegmentator-MRI data with FDR = 0.017 for the DSC and FDR = 0.787 for the HD. (3) The NAKO T2 and the AMOS data have fewer annotated classes. To allow a fair comparison to the NAKO GRE sequences we additionally report the results for the classes-subset on the in-phase sequences, as a representative for the GRE scans. (4) For the AMOS CT dataset we use TotalSegmentator-CT instead of TotalSegmentator-MRI as a baseline model. Abbreviations: GRE: Gradient Echo, T2 HASTE: T2-weighted Half-Fourier Acquisition Single-shot Turbo spin Echo.

**Table 3: Vessel Consistency**

| Vessel | VC | Fraction Of Samples With More Than One Component | Average Number Of Components If Segmented Incorrectly |
|---|---|---|---|
| Aorta | 1.00 (200/200) | 0.00 (0/200) | — |
| Inferior vena cava | 0.92 (184/200) | 0.08 (16/200) | 2.06 ± 0.25 |
| Portal/splenic vein | 0.40 (79/200) | 0.60 (121/200) | 2.60 ± 0.87 |
| Iliac arteries left | 0.69 (137/200) | 0.32 (63/200) | 2.60 ± 0.85 |
| Iliac arteries right | 0.60 (119/200) | 0.41 (81/200) | 2.54 ± 0.88 |
| Iliac veins left | 0.95 (189/200) | 0.06 (11/200) | 2.18 ± 0.40 |
| Iliac veins right | 0.85 (170/200) | 0.15 (30/200) | 2.13 ± 0.43 |

Vessel consistency (VC) refers to the proportion of segmentations where a given vessel class is represented by a single connected component. A higher VC indicates that the model effectively treats the vessel as a unified entity. For the gradient echo (GRE) sequences of the German National Cohort (NAKO), MRSegmentator did not fail to segment any structure; therefore, the fraction of samples with multiple components is directly related to the inverse of the VC. The final column presents the average number of components detected when multiple components were identified. For example, segmentations of the portal/splenic vein exhibit a VC of 40% (79/200), indicating that, in 60% (121/200) of cases, multiple components are detected. The average number of components observed in these cases is 2.6, suggesting a high degree of fragmentation in the segmentations for this vessel class.

**Table 4: Segmentation Failure Types in External NAKO Data**

| MRI Sequence | Under-segmentation Fraction | Over-segmentation Fraction |
|---|---|---|
| GRE in-phase | 0.08 (0.10) | 0.07 (0.06) |
| GRE opposed-phase | 0.09 (0.11) | 0.10 (0.06) |
| GRE water-only | 0.08 (0.11) | 0.10 (0.07) |
| GRE fat-only | 0.09 (0.13) | 0.09 (0.06) |
| T2 HASTE | 0.11 (0.21) | 0.07 (0.10) |

Segmentation failure types across different MRI sequences from the German National Cohort (NAKO). Segmentation failures are measured relative to the reference standard, distinguishing between under-and over-segmentation. The table reports the median and interquartile range of segmentation failures across 40 target structures across all participants. Under-segmentation is generally more prevalent, except in water-only sequences, where over-segmentation is more common. The highest under-segmentation rates are observed in T2-HASTE and fat-only sequences. Abbreviations: GRE: Gradient Echo, T2 HASTE: T2-weighted Half-Fourier Acquisition Single-shot Turbo spin Echo.

# Segmenting Whole-Body MRI and CT for Multi-Organ Anatomical Structure Delineation

### Key Result

MRSegmentator accurately segmented 40 anatomical structures on MRI and CT across three external datasets.

### Patients:

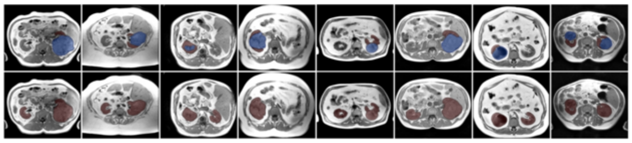
- 1,894 healthy participants and patients with various pathologies

### Methods:

- MRSegmentator was trained on a total of 2,649 scans and tested on 1,289 scans.
- Segmentation quality was assessed by Dice similarity coefficient (DSC).

### Results:

- In 50 participants with whole-body MRI scans, MRSegmentator achieved DSCs of 0.91 ± 0.05 for Dixon in-phase and 0.85 ± 0.08 for T2-HASTE sequences.
- Across 60 MRI and 300 CT scans, MRSegmentator achieved DSCs of 0.79 ± 0.11 and 0.84 ± 0.12, respectively.



*The first row depicts manually annotated kidneys (red) and tumors (blue) in eight different MRI scans. The second row shows the corresponding kidney segmentations generated by MRSegmentator. The model accurately localizes and segments the kidneys even in the presence of large tumors, demonstrating its robustness in handling pathological cases. In the fourth sample, the missing right kidney due to a previous nephrectomy is correctly not segmented by the model.*

**Häntze H and Xu L et al. Published Online:** August 6, 2025
https://doi.org/10.1148/ryct.240777