



Advanced Automated Model for Robust Bone Marrow Segmentation in Whole-body MRI

Fabian Bauer¹, Jessica Kächele, Juliane Bernhard, Marina Hajjiyanni, Niels Weinhold, Sandra Sauer, Martin Grözinger, Marc-Steffen Raab, Elias K. Mai, Tim F. Weber, Hartmut Goldschmidt, Heinz-Peter Schlemmer, Klaus Maier-Hein, Stefan Delorme, Peter Neher, Markus Wennmann

Rationale and Objectives: To establish an advanced automated bone marrow (BM) segmentation model on whole-body (WB-)MRI in monoclonal plasma cell disorders (MPCD), and to demonstrate its robust performance on multicenter datasets with severe myeloma-related pathologies.

Materials and Methods: The study cohort comprised multi-vendor, multi-protocol imaging data acquired with varying field strength across 8 different centers. In total, 210 WB-MRIs of 207 MPCD patients were included. An nnU-Net algorithm was established for segmenting the individual bone marrow spaces (BMS) of the spine, pelvis, humeri and femora (advanced segmentation model). For this task, 186 T1-weighted (T1w) WB-MRIs from center 1 were used in the training set. Test sets included 12 T1w WB-MRIs from center 2 (I) and 9 T1w WB-MRIs from centers 3–8 (II). Example cases were included to showcase segmentation performance on T1w WB-MRIs with extensive tumor load. The segmentation accuracy of the advanced segmentation model was compared to a prior established basic segmentation model by calculating Dice scores and using the Wilcoxon signed-rank test.

Results: The mean Dice score on the individual BMS was 0.89 ± 0.13 (test set I) and 0.88 ± 0.11 (test set II), significantly higher than the Dice scores of a prior basic model ($p < 0.05$). Dice scores for the BMS of the individual bones ranged from 0.77 to 0.96 (test set I), and 0.81 to 0.95 (test set II). BM altered by myeloma-relevant pathologies, artifacts or low imaging quality was precisely segmented.

Conclusion: The advanced model performed reliable, automated segmentations, even on heterogeneously acquired multicenter WB-MRIs with severe pathologies.

Key Words: Multiple Myeloma; Deep Learning; Segmentation; Magnetic Resonance Imaging.

© 2025 The Association of Academic Radiology. Published by Elsevier Inc. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

Abbreviations: **BM** bone marrow, **BMS** bone marrow space, **FL** focal lesion, **GMMG** German-Speaking Myeloma Multicenter Group, **IMWG** International Myeloma Working Group, **MGUS** monoclonal gammopathy of unknown significance, **MM** multiple myeloma, **m-protein** monoclonal protein, **MPCD** monoclonal plasma cell disorder, **PCI** plasma cell infiltration, **SFLCR** serum free light chain ratio, **SMM** smoldering multiple myeloma, **WB** whole-body

INTRODUCTION

Patients with monoclonal plasma cell disorders (MPCD) exhibit considerable variability in clinical course and outcome (1). Genetic aberrations (1,2)

and tumor load (3,4) vary both among different patients and within different tumor sites in the same patient. The assessment of progression risk in smoldering multiple myeloma (SMM) and staging of multiple myeloma (MM) both include

Acad Radiol 2025; 32:2824–2835

From the Division of Radiology, German Cancer Research Center (DKFZ), 69120 Heidelberg, Germany (F.B., M.G., H.P.S., S.D.); Division of Musculoskeletal Imaging and Intervention, Massachusetts General Hospital and Harvard Medical School, Boston, MA 02114 (F.B.); Division of Medical Image Computing, German Cancer Research Center (DKFZ), 69120 Heidelberg, Germany (J.K., J.B., K.M.H., P.N.); German Cancer Consortium (DKTK), Partner Site Heidelberg, 69120 Heidelberg, Germany (J.K., K.M.H., P.N.); Internal Medicine V, Hematology, Oncology and Rheumatology, Heidelberg University Hospital, 69120 Heidelberg, Germany (M.H., N.W., S.S., M.S.R., E.K.M., H.G.); Diagnostic and Interventional Radiology, University Hospital Heidelberg, 69120 Heidelberg, Germany (T.F.W., M.W.); National Center for Tumor Diseases (NCT), University Hospital Heidelberg, 69120 Heidelberg, Germany (H.G., K.M.H., P.N.); Pattern Analysis and Learning Group, Department of Radiation Oncology, Heidelberg University Hospital, 69120 Heidelberg, Germany (K.M.H., P.N.). Received September 5, 2024; revised December 19, 2024; accepted December 27, 2024. **Address correspondence to:** F.B. e-mail: fabian.bauer@dkfz-heidelberg.de

¹ ORCID ID: 0000-0002-0896-5734

plasma cell infiltration (PCI), monoclonal protein (m-protein), and cytogenetic aberrations (5–7). Whole-body MRI (WB-MRI) has proven valuable in assessing bone marrow (BM) involvement in MPCD. The presence and number of focal lesions (FLs) are particularly informative (8–11), while the prognostic value of the different patterns of diffuse infiltration remains inconclusive (8–10,12,13). The current diagnostic guideline of MM includes more than one FL as a myeloma-defining criterion (14). Consequently, the International Myeloma Working Group (IMWG) recommends WB-MRI for diagnosis and monitoring in SMM as well as in MM without osteolytic lesions in CT imaging (15). Recent advancements suggest that assessing FLs beyond counting them provides additional prognostic information. This includes evaluating their size (2,16), volume (13,17,18), growth patterns (13), and radiological features (19,20).

The integration of automatic computational approaches, such as deep learning, allows for the analysis of even large and complex imaging datasets with minimal to no time constraint (21,22). Imaging features extracted from relevant segmented structures can be incorporated in automated models, predicting disease and patient characteristics (23). In an approach to automatize the underlying segmentation of the regions of interest needed for subsequent machine-learning analysis in patients with MPCD, several automated segmentation models were presented for the bone and BM in different sequences in MRI (24–32), including a prior work of our group on whole-body BM segmentation on MRI (33). Previous segmentation models on MRI face several challenges that hinder their clinical application. These include the following: (1) limited model performance on external, multicenter imaging data, (2) segmentation performance that is inferior to radiologists, (3) BM segmented of a limited number of bones that is unrepresentative of the overall tumor load, (4) segmented bones lacking individual labeling, or (5) models that lack robust segmentation performance in datasets affected by severe pathologies.

In this study, we aimed to establish an advanced segmentation model for BM from WB-MRIs with corresponding MPCD diagnosis, which segments the BM of the whole body with individual labeling of all bones. The model was intended to perform robustly particularly in multicenter data with severe pathologies, aiming for a performance comparable to the interrater variability between radiologists.

MATERIALS AND METHODS

This retrospective study was conducted in accordance with the Declaration of Helsinki and approved by the local institutional review board (S-537/2020) with waiver of informed consent.

Study Cohort and Data Annotation

Imaging and clinical data were acquired at eight different centers between 2015 and 2021 and subsequently processed

within the study between 2021 and 2024. Data were split by center into training (center 1) and test sets (centers 2–8) based on the respective data available.

The training set consisted of three different datasets. Dataset 1 consecutively included SMM patients from the Transregio-79B8 study (ClinicalTrials.gov identifier: NCT01374412) (34) with WB-MRIs performed at center 1 between 2014 and 2019. Dataset 1 was defined in a previous study with exclusion criteria being the usage of an older MRI scanner at center 1, secondary malignancy, or the diagnosis of MGUS or MM at the time or the MRI scan (33). In dataset 2, patients with newly diagnosed MM enrolled in the German-speaking Myeloma-Multicenter Group (GMMG) HD7 study (EudraCT identifier: 2017–004768–37) and baseline WB-MRI consecutively performed at center 1 between 2018 and 2020 were included (35). For this previously defined dataset, unavailability of the composed WB-MRI or the presence of a pelvic implant were applied as exclusion criteria (25). Dataset 3 included by consecutive selection the WB-MRIs of patients with MPCD acquired between 2015 and 2020 at center 1. This previously established dataset was defined by the following exclusion criteria: (1) scan performed at a PET-MRI scanner, (2) scan performed after start of MM systemic therapy, (3) exam performed at a 3 Tesla MRI scanner, or (4) inclusion in one of the two aforementioned datasets (25).

The model was tested on two previously defined external, independent test sets. Test set 1 included consecutively selected patients with SMM and WB-MRI performed at center 2 between February 2021 and March 2021. Exclusion criteria were systemic treatment started before MRI scan or the diagnosis of MGUS, MM or solitary plasmacytoma at the time of the MRI scan. Test set 2 included a consecutive selection of patients with newly diagnosed MM enrolled in the GMMG HD7 study and baseline WB-MRI performed between 2018 and 2020. Exclusion criteria were (1) imaging performed at center 1 or 2, (2) unavailable coronal MRI scan, (3) unavailable composed coronal WB-MRI scan, (4) missing information in the DICOM header, or (5) imaging data not transferred to center 1. For this dataset, the two WB-MRI scans from each imaging center with available data were selected based on the lowest visually assessed tumor load and minimal imaging artifacts (33).

Three WB-MRIs performed at center 2 between 2015 and 2021 were included by convenience selection as example cases. Due to the moderate number and size of focal lesions in the test sets, these cases served as additional examples to showcase BM segmentation performance on WB-MRIs with the utmost visual tumor load in WB-MRI. Inclusion criteria were (1) presence of the most severe myeloma-related pathologic changes in the BM in the WB-MRI, and (2) an available composed T1w WB-MRI. Exclusion criteria were (1) inclusion in the training or test sets, or (2) prior myeloma therapy.

Clinical data collected within 90 days of WB-MRI exams were obtained and included age, sex, body-mass-index

TABLE 1. List of Scanners with Respective MRI Sequence Parameters Used for the Data Sets

Center (Dataset) Scanner (s) (Field Strength)	Coronal T1-Weighted Turbo Spin Echo		
	In-Plane Resolution	Slice Thickness, Distance Factor	Repetition Time, Echo Time
Center 1, Dataset 1			
Siemens Magnetom Aera (1.5 T)	0.6 mm × 0.6 mm	5 mm, 10%	550 ms/9.5 ms
Siemens Magnetom Symphony (1.5 T)	1.3 mm × 1.3 mm	5 mm, 20%	739 ms/12 ms
Center 1, Dataset 2			
Siemens Magnetom Aera (1.5 T)	1.3 mm × 1.3 mm	5 mm, 10%	528 ms/8.4 ms
Center 1, Dataset 3			
Siemens Magnetom Aera (1.5 T)	0.6 mm × 0.6 mm or 1.3 mm × 1.3 mm	5 mm, 10%	528 ms–876 ms / 8.4 ms–9.5 ms
Center 2			
Siemens Magnetom Avanto ^{fit} (1.5 T)	0.7 mm × 0.7 mm to 1.3 mm × 1.3 mm	5 mm, 10–20%	438 ms–627 ms / 6 ms–11 ms
Center 3			
GE Optima MR450w (1.5 T)	0.9 mm × 0.9 mm	8 mm, 0%	429 ms/9.0 ms
Center 4			
Siemens Magnetom Aera	0.9 mm × 0.9 mm	5 mm, 25%	650 ms/8.7 ms
Center 5			
Philips Ingenia (1.5 T)	1.0 mm × 1.0 mm	6 mm, 10%	351 ms/17.5 ms
Center 6			
Siemens Magnetom Avanto ^{fit} (1.5 T)	1.3 mm × 1.3 mm	6 mm, 0%	439 ms/11 ms
Siemens Prisma ^{fit} (3.0 T)	1.3 mm × 1.3 mm	6 mm, 0%	439 ms/11 ms
Center 7			
Philips Ingenia (1.5 T)	1.0 mm × 1.0 mm	6 mm, 17%	412 ms/4.0 ms
Center 8			
Siemens Magnetom Area (1.5 T)	1.2 mm × 1.2 mm	5 mm, 10%	720 ms/8.2 ms

T, Tesla

(BMI), disease stage, serum free light chain ratio (SFLCR), m-protein, and PCI.

Imaging

In this study, non-contrast coronal T1w turbo spin echo (tse) images of WB-MRI scans were used. Datasets with WB-MRI of patients with MPCD were acquired at center 1 to 8 using seven different MRI scanner models from three different vendors with diverse sequence parameters. Details on the MRI acquisition parameters used at each center are presented in the [Table 1](#).

For establishing the advanced version of the automated segmentation model, the bone marrow spaces (BMS) containing the BM of the humeri, femora, pelvis, sacrum, lumbar vertebrae (L1–L5), thoracic vertebrae (T1–T12) and cervical vertebrae (C2–C7) were manually delineated in each image of the composed T1w WB-MRIs and individually labeled using the Medical Imaging Interaction Toolkit (MITK) ([36](#)). Dataset 1 comprised WB-MRIs with manual segmentation of 30 BMSs that had been previously segmented in a separate study. The basic segmentation algorithm established in this preceding study was employed to automatically segment the BMS in the WB-MRIs of datasets 2 and 3 ([33](#)). The automated BM segmentation for datasets 2 and 3 was manually refined

through manual correction by a medical doctor with four years of experience in image segmentation as well as imaging of the BM in MM (F.B.). BMS in test set 1 and 2 were manually segmented in a prior study ([33](#)).

Algorithmic Architecture

The advanced automated segmentation model was trained on T1w WB-MRIs with 30 individually labeled BMSs using the 3D nnU-Net segmentation framework ([37](#)). Within the preprocessing of the nnU-Net framework, WB-MRIs of the training set were resampled to the median spacing of 5.5 mm (z-axis), 0.5859375 mm (y-axis), and 0.5859375 mm (x-axis). Test imaging data was resampled accordingly.

In addition to the 3D nnU-Net segmentation framework, two consecutive automated postprocessing steps were implemented. The right and left femora, humeri, and pelvic bones were automatically relabeled to ensure correct assignment to the respective sides (step 1) ([33](#)). This step was established to reduce mislabeling of symmetrical BMS. This relabeling process was based on the relative position of the sacral bone centroid and the centroids of the segmented bones. Specifically, the algorithm automatically verified whether the position (left/right of the sacral bone) matched the assigned label. If the ratio of the volumes of the right and left pelvic

BMS was 0.75 or less, an additional automatic postprocessing step was initiated, as this condition indicates that parts of one bone may have been incorrectly labeled as belonging to the opposite side (step 2). In this step, labels were automatically reassigned based on a sagittal plane defined by the geometrical center of the sacral BMS. Model training was performed using PyTorch, version 1.12, in Python, version 3.10. To assess real-world clinical applicability, the model was evaluated on a NVIDIA GeForce RTX 4070 Ti.

Comparison with Basic Segmentation Model

The advanced model was compared to a previously established basic model for segmentation accuracy. In comparison to the basic model, the training dataset for the advanced model was more than three times the size of the earlier training set. Furthermore, the advanced model was trained on a high proportion of WB-MRIs from patients with MM, with the objective of including a substantial number of cases with severe pathologies in the training dataset. The training set of the basic model included manual segmentations of 30 BMS from 50 WB-MRIs of patients diagnosed with SMM (33). Therefore, the training data of the basic model only included WB-MRIs with lower tumor burden, as defined by the IMWG criteria for diagnosing SMM (14). Training data of the advanced segmentation model also included the training data of the basic model (dataset 1). Comparability between the advanced and basic models was ensured by using the identical test sets, as well as an interrater reference established in the prior study, based on segmentations of three WB-MRIs by two radiologists (33).

Statistical Analysis

The accuracy of the automated segmentations in the test sets was evaluated by calculating the Dice score for every

individual BMS. Additionally, the mean Dice score of the individual BMS was calculated. The Dice scores of the basic and advanced segmentation model were compared using the Wilcoxon signed-rank test. P-values < 0.05 were considered statistically significant. All statistical metrics were generated using scikit-learn and SciPy packages in Python, version 3.10.

RESULTS

Descriptive Statistics

Patient and disease characteristics of the training and test sets of the advanced segmentation model are provided in Table 2. Figure 1 provides an overview on the datasets utilized for the training and testing of the advanced segmentation model. The advanced automated segmentation model was trained and tested using 186 WB-MRIs from 182 patients (106 of whom had corresponding MM diagnoses) acquired at centers 1 and 21 WB-MRIs from 21 patients with (S)MM acquired at centers 2–8. Three example cases were included to showcase the efficacy of the segmentation models in the context of extensively myeloma-altered BM. The advanced model mostly included WB-MRIs from patients diagnosed with MM (106/186 patients), followed by SMM (75/186 patients), whereas the basic model only included WB-MRIs with corresponding SMM diagnosis (50/50 patients) (Supplementary Table 1).

Performance of the Advanced Automated Whole-Body Bone Marrow Segmentation Model

For the fully automated segmentation of the individual BMSs, interference time for each WB-MRI was 18 min. Table 3 presents the mean Dice score for each BMS, as well

TABLE 2. Patient and Disease Characteristics				
	Training Set	Test Set 1	Test Set 2	Exemplary Cases
WB-MRIs	186	12	9	3
Center	Center 1	Center 2	Center 3 - 8	Center 2
Patient Characteristics				
Age in Years	62; 54–67	55; 49–61	54; 50–60	55; 47–62
Male sex	114; 61%	9; 75%	7; 78%	2; 67%
BMI in kg/m ² (n missing values)	27; 25–27	28; 25–32	28; 25–33 (5)	31; 29–39
Disease Characteristics				
MGUS	5	0	0	0
SMM	75	12	0	0
MM	106	0	9	3
SFLCR (n missing values)	22; 5–133 (13)	6; 2–17	36; 10–323	19; 14–109
M-Protein in g/l (n missing values)	21; 11–36 (30)	15; 11–22 (5)	35; 24–44 (5)	34; 24–61
PCI in % (n missing values)	20; 12–39 (43)	10; 9–14 (6)	28; 11–50 (1)	19; 12–38

Data presented as median with interquartile range for continuous variables and n (%) for categorical variables. Data corresponding to each WB-MRI utilized in this study is given for the training and test sets of the advanced automated segmentation model. The test sets of the advanced segmentation model were defined in an earlier study and used to compare the results with the earlier version (33). MGUS, monoclonal gammopathy of unknown significance; MM, multiple myeloma; m-protein, monoclonal protein; PCI, plasma cell infiltration; SFLCR, serum free light chain ratio; SMM, smoldering multiple myeloma; WB-MRI, whole-body MRI

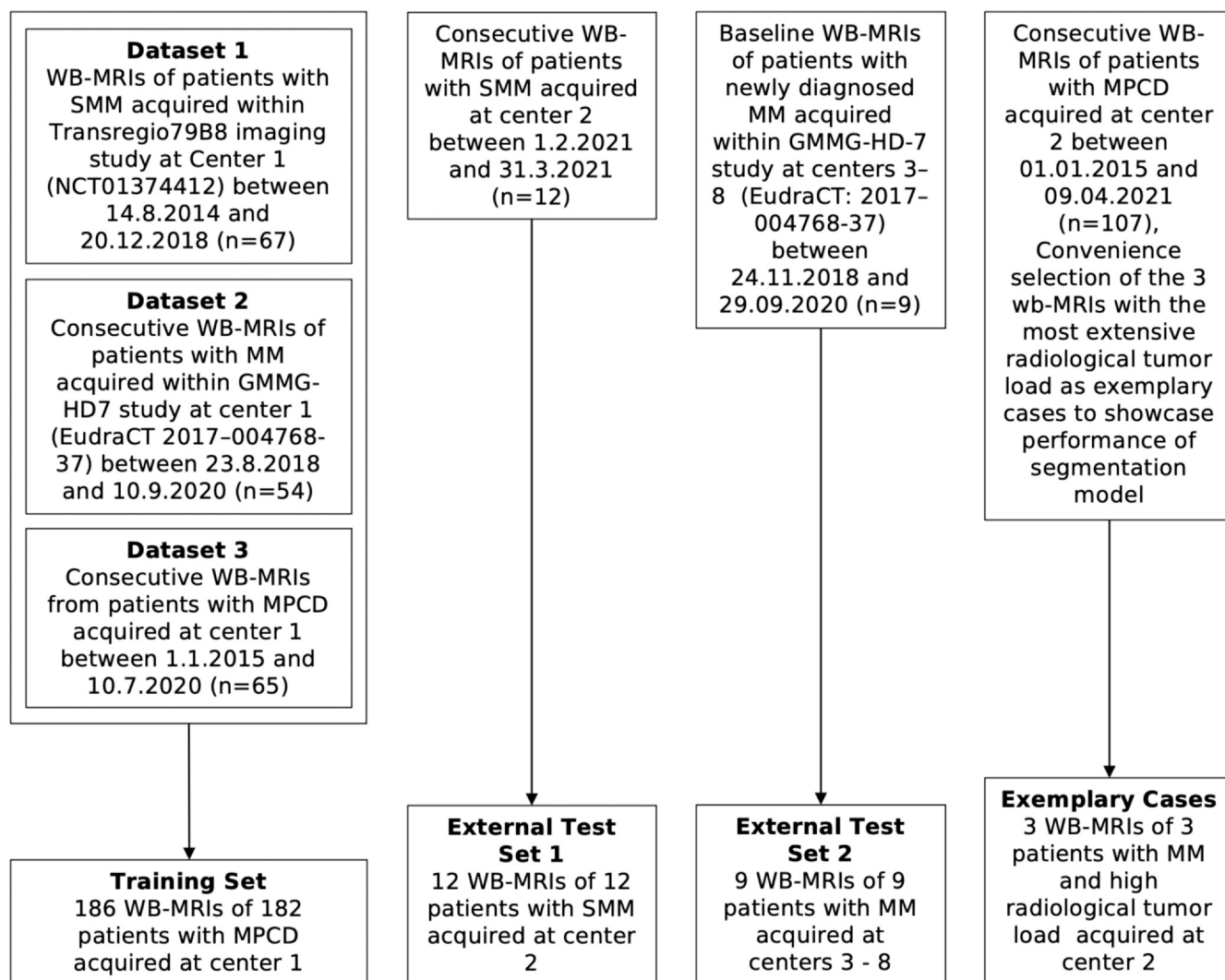


Figure 1. Overview of the datasets used for training and testing of the segmentation model. Datasets were used as defined in two prior studies (25,33). GMMG, German-speaking Myeloma-Multicenter Group; MM, multiple myeloma; SMM, smoldering multiple myeloma; WB, whole-body.

as the overall mean Dice score of all individual BMSs. In test set 2, the mean segmentation performance (Dice score of 0.88) was nearly equivalent to that observed in test set 1 (Dice score of 0.89). Test set 1 contained a smaller number of pathologies that were less severe compared to those observed in test set 2, which was characterized by a higher tumor load, more diverse acquisition parameters and the presence of implants that affected image quality.

Comparison of the Advanced Segmentation Model to a Prior Established Basic Model and Interrater Reference

Figure 2 depicts delta Dice scores, which are used to evaluate the segmentation accuracy of the established model in comparison to a prior basic model and an interrater reference. In both test sets, the mean performance of the advanced segmentation model across all individual BMS (test set 1: Dice score 0.89 ± 0.13 , test-set 2: dice score

0.88 ± 0.11) was found to be equivalent to the interrater-variability (dice score 0.88 ± 0.01) previously established by two radiologists (33). In test set 1, the improvement observed in the advanced model's performance in comparison to the basic model was relatively minor (0.89 ± 0.13 vs. 0.87 ± 0.06 ; $p = 0.021$). This was likely due to the fact that test set 1 only included WB-MRIs of SMM patients with good image quality, which is similar to the training set of the basic model (33). However, in test set 2, which included multicentric imaging data with variable quality and severe pathologies, the mean Dice score for the BMSs segmented by the advanced model exhibited a marked improvement compared to the basic model (0.88 ± 0.11 vs. 0.83 ± 0.11 ; $p = 0.039$). The advanced segmentation model demonstrated inferior performance to the interrater reference standard for some individual thoracic BMSs.

Problems of mislabeling of individual thoracic vertebrae appeared less frequent for the advanced model compared to

TABLE 3. Segmentation Performance

Bone Marrow Space	Test Set 1	Test Set 2
Mean of All Individual Bones	0.89 ± 0.13	0.88 ± 0.11
<i>Individual Bones</i>		
C2	0.87 ± 0.06	0.86 ± 0.09
C3	0.84 ± 0.09	0.86 ± 0.06
C4	0.85 ± 0.06	0.86 ± 0.06
C5	0.86 ± 0.05	0.82 ± 0.08
C6	0.82 ± 0.10	0.85 ± 0.04
C7	0.85 ± 0.06	0.83 ± 0.08
T1	0.87 ± 0.03	0.81 ± 0.12
T2	0.88 ± 0.03	0.88 ± 0.03
T3	0.87 ± 0.06	0.86 ± 0.04
T4	0.83 ± 0.19	0.88 ± 0.05
T5	0.78 ± 0.29	0.88 ± 0.09
T6	0.77 ± 0.34	0.88 ± 0.10
T7	0.85 ± 0.26	0.84 ± 0.19
T8	0.86 ± 0.26	0.84 ± 0.15
T9	0.90 ± 0.10	0.81 ± 0.29
T10	0.94 ± 0.02	0.83 ± 0.29
T11	0.94 ± 0.02	0.91 ± 0.04
T12	0.93 ± 0.01	0.92 ± 0.02
L1	0.93 ± 0.02	0.90 ± 0.04
L2	0.93 ± 0.02	0.91 ± 0.05
L3	0.94 ± 0.02	0.90 ± 0.06
L4	0.93 ± 0.02	0.90 ± 0.07
L5	0.93 ± 0.02	0.87 ± 0.08
Sacral Bone	0.88 ± 0.03	0.84 ± 0.04
Right Pelvis	0.92 ± 0.01	0.91 ± 0.01
Left Pelvis	0.92 ± 0.02	0.90 ± 0.02
Right Femur	0.95 ± 0.01	0.95 ± 0.02
Left Femur	0.96 ± 0.01	0.95 ± 0.01
Right Humerus	0.94 ± 0.03	0.90 ± 0.07
Left Humerus	0.94 ± 0.02	0.92 ± 0.03

Dice score and standard deviation are given for the established advanced segmentation model of this study on predefined external, independent test sets (33). C, cervical vertebra; T, thoracic vertebra; L, lumbar vertebra

the basic model. However, mislabeled vertebrae were still observed in a small number of cases. Figure 3 illustrates segmentations of both models for a representative WB-MRI with high tumor load and artifacts, providing a realistic example for assessing the segmentation performance. A comparison of the segmentations performed by the two models revealed that the advanced model achieved a mean Dice score of 0.88 for the individual BMSs of the WB-MRI, as reported for test set 2 in total. The basic model (mean Dice of the individual BMS: 0.75) demonstrated a markedly inferior segmentation performance, with a Dice score below 0.9 achieved for 20 out of 30 BMS segmentations (advanced model: 12 out of 30). In eight cases out of 30 BMS segmentations, the calculated Dice scores were below 0.8 (advanced model: 2/30). For the advanced model, the low Dice scores for cervical vertebra 7 and thoracic vertebra 1 were due to partial mislabeling, with each being incorrectly

assigned as the other by the advanced automated segmentation model.

To examine the model's performance in the context of BM altered by severe pathologies in WB-MRI, three exemplary cases with an exceptionally high tumor load were additionally segmented by both the basic and advanced models. A comprehensive visual analysis revealed that the advanced model successfully segmented BM with large FLs, ensuring that these areas are fully included in the segmentation without being left out, whereas the basic model exhibited limitations. Figure 4 illustrates representative MR images alongside the automated segmentations performed by the advanced model for the three example cases.

DISCUSSION

In this study, we established an advanced automated segmentation model for BM in T1w WB-MRIs, demonstrating robust performance even in the presence of pathological alterations to the BM, as well as varying imaging quality and acquisition parameters. The segmentation performance on a multicenter, independent test set was comparable to that of a radiologist's interrater reference, representing a significant improvement over a prior automated segmentation model presented by our group (33). Upon visual examination of example cases, it was observed that the advanced segmentation model appeared to have a greater capacity to encompass myeloma-altered BM than an earlier established model.

Despite advances in machine-learning-assisted image analysis, existing models utilizing different underlying algorithm concepts have only presented precise segmentation performance for individual BMSs or bones on MRI (24,25,29,30,31). By focusing only on a small group of bones, further overall application of the segmentation model is limited. For WB-MRI application, Arabi et al. presented an atlas-based bone segmentation model on Dixon MRI for attenuation correction in PET/MRI, achieving the best result with a Dice coefficient of 0.76 (28). Ceranka et al. developed a multi-atlas bone segmentation model on 10 T1w WB-MRIs of five healthy volunteers and five prostate cancer patients with bone metastases, evaluating different label-fusing strategies. The mean Dice score for the binary classification of bone versus background was 0.89 in a leave-one-out cross-validation setting (32). Candito et al. recently presented a multistep, atlas-based registration model for bone segmentation in diffusion-weighted WB-MRI, achieving mean Dice scores of 0.77 to 0.68 for different spinal regions, 0.65 for long bones, and 0.70 for the pelvis (26). In the respective study, bones were not individually labeled, and intervertebral discs were not excluded from the spinal segmentations (26). In comparison to atlas-based and classification forest-based models, Lavdas et al. found a convolutional neural network trained on manually annotated T2w WB-MRI to achieve superior performance in terms of

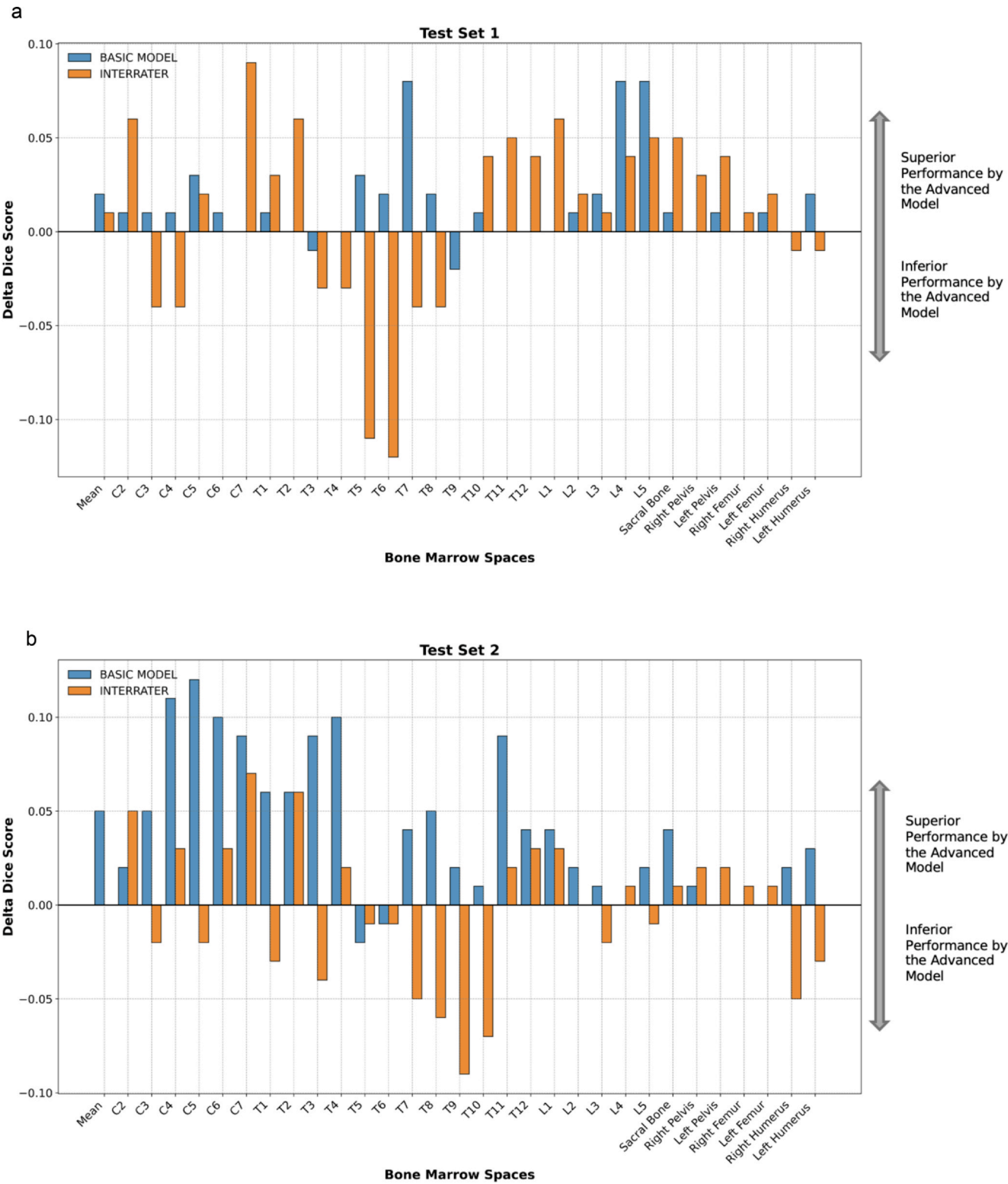


Figure 2. Delta Dice Scores of the established advanced segmentation model compared to the basic model and interrater reference. The difference in the respective mean Dice scores are presented for the advanced segmentation model compared to the basic segmentation model (blue color) or compared to the interrater reference (orange color) for the bone marrow spaces (33). Positive delta Dice scores resulted from a higher Dice Score of the advanced model for the respective bone marrow space in a test set compared to the Dice scores of the basic model or an interrater reference. Delta Dice scores are displayed for test set 1 (a) and 2 (b) individually. C, cervical vertebra; T, thoracic vertebra; L, lumbar vertebra.

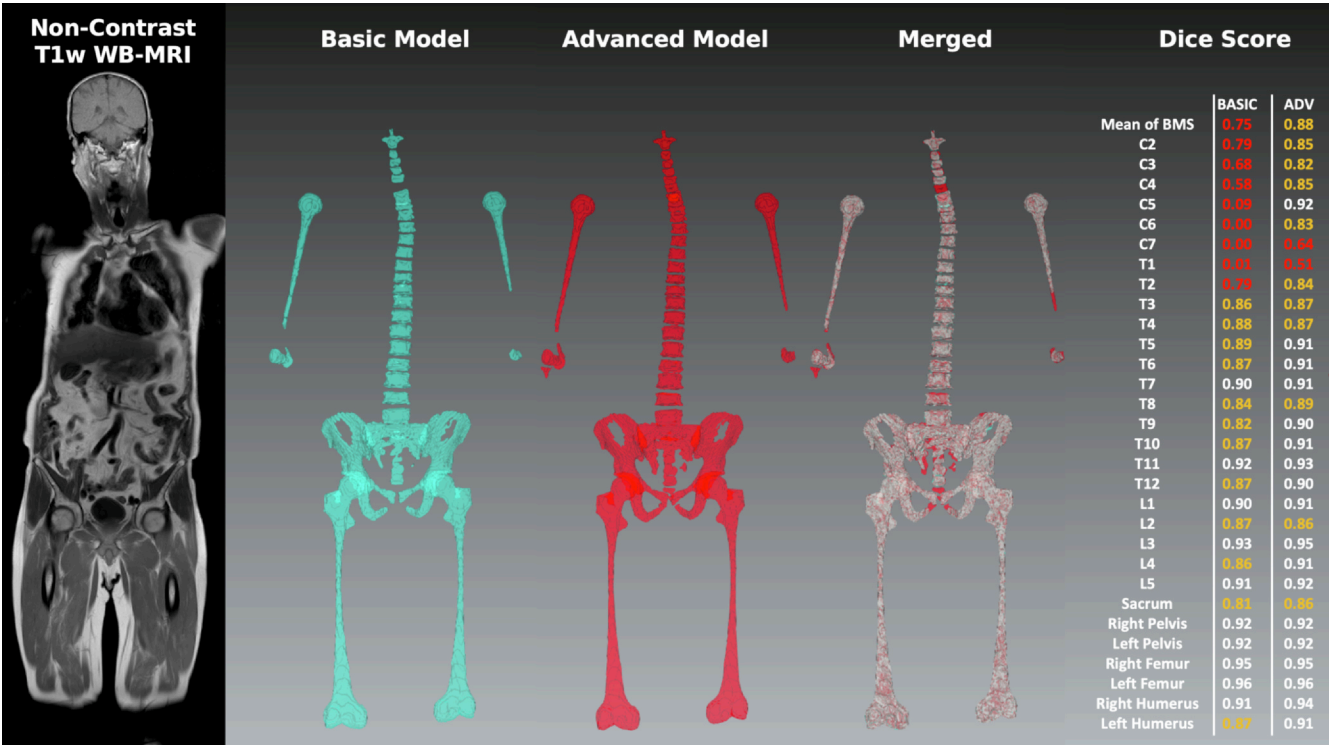


Figure 3. Exemplary visual comparison of both segmentation model versions. On the left, an exemplary non-contrast T1- weighted WB-MRI scan of external test set 2 is shown with lower imaging quality due to motion and composing artifacts, particularly in the cervical MRI section, presenting challenging conditions for segmentation. The 69-year-old male subject was newly diagnosed with multiple myeloma International Staging System (ISS) II at the time of the MRI scan. In the center, the automated segmentations performed by both the basic and advanced model are displayed in 3D in turquoise and red, respectively. For reasons of clarity and comprehensibility, individual segmentations of each version are displayed in a single color. A merged 3D model illustrates the overlap in segmented voxels of both model versions. Despite lower imaging quality, both model versions segmented most of the bone marrow accurately. On the right, Dice scores for the automated segmentations of both the basic and advanced (ADV) model are given. Dice scores < 0.8 and < 0.9 are highlighted in red and orange color. Compared to the extensive mislabeling of segmented cervical BMS by the basic model, the advanced model segmented every individual vertebra and performed correct labeling. In the automated segmentation by the basic model, cervical and thoracic vertebrae have been partly mislabeled, leading to a very low Dice score. BMS, bone marrow space; C, cervical vertebra; T, thoracic vertebra; T1w, T1-weighted; L, lumbar vertebra; WB, whole-body.

pooled Dice score across various segmented structures. In their convolutional neural network model, the automatic segmentation of the spine and pelvic bones, including femurs, resulted in a Dice score of 0.87 and 0.81, respectively. These scores were based on fivefold cross-validation experiments that included 34 artifact-free MRI datasets (27). While all these models undoubtedly represent substantial contributions to this research field, these prior models were established on relatively small datasets and tested using cross-validation only. As a result, there was a lack of external test sets for a model evaluation regarding a potential real-world clinical application. Additionally, the models did not provide individual labeling of each bone, and the reported performance of most of the models was inferior to the interrater variability between radiologists. In a previous work, our group presented a basic automated segmentation model for 30 BMS on T1w WB-MRI based on the nnU-Net architecture (37). This model was trained on a relatively limited training set of 50 WB-MRIs from patients with SMM (33), which are defined as patients with low tumor load and few

pathologies in the BM. Although the basic model demonstrated satisfactory performance in an external test set comprising cases with high imaging quality and low tumor load, comparable to a radiologist (test set 1), it exhibited a notable decline in segmentation accuracy on a multicenter test set 2, which included WB-MRIs of MM patients with high tumor load and heterogeneous imaging quality (33). As demonstrated in this preceding study, a very good performance in relatively straightforward cases does not necessarily guarantee that a model will also perform well in real-world application, where cases with high tumor load or discrepancies in technical imaging parameters are encountered. Consequently, it is imperative to exercise caution when interpreting high performance metrics reported in other studies, particularly if it is unclear whether the model has been tested in cases with high tumor load and heterogeneous, multicenter imaging data, or whether the high metrics result from testing in cases with mainly physiological findings. In our present study, the advanced segmentation model demonstrated a markedly improved performance when trained on a larger quantity of

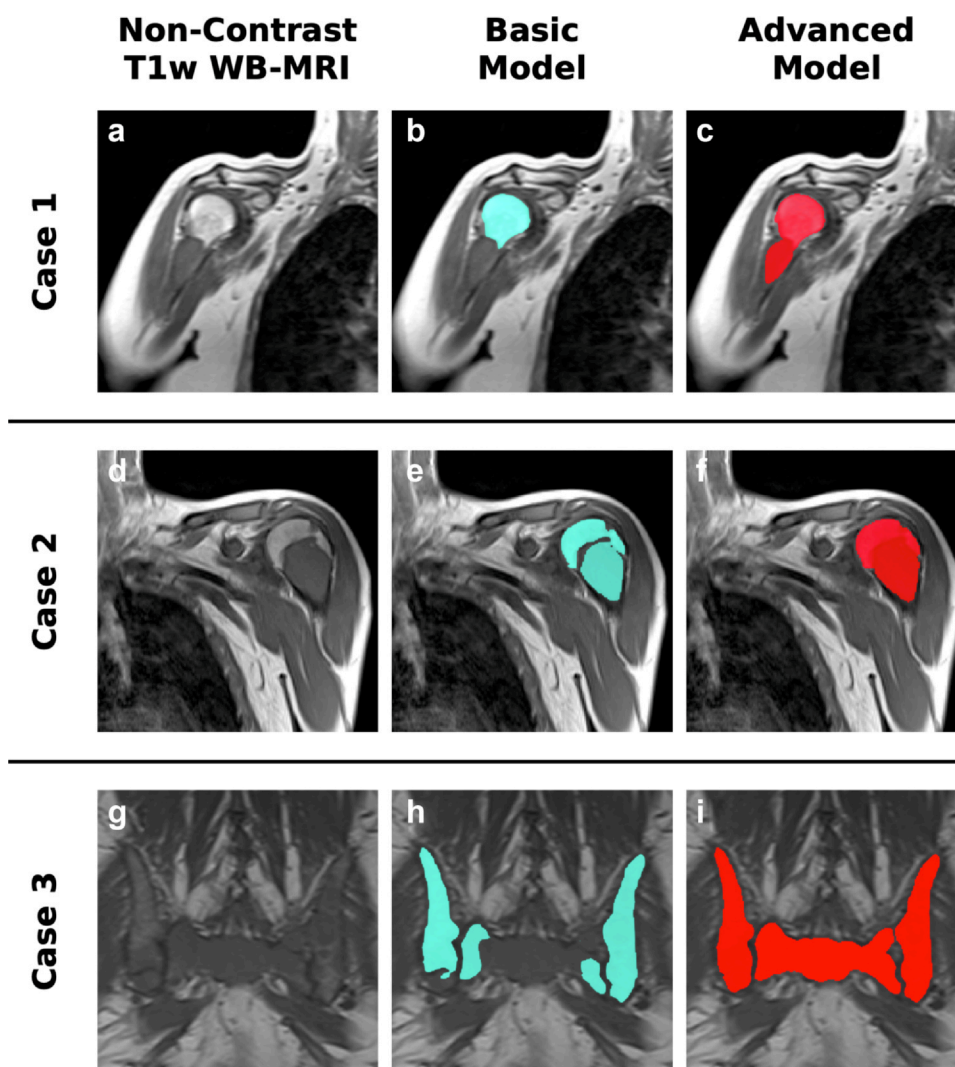


Figure 4. Differences in the automated bone marrow segmentations altered by focal lesions performed by both model versions. For the exemplary case 1 (a,b,c), case 2 (d,e,f), and case 3 (g,h,i) with radiologically assessed high tumor load and multiple myeloma diagnosis, bone marrow spaces containing prominent focal lesions were segmented incompletely by the basic model (33). The advanced segmentation model segmented the focal lesions in full. T1w, T1-weighted; WB, whole-body.

data, particularly cases with severe pathologies. This resulted in a significant enhancement in its performance on both external, independent test sets. Notably, the segmentation performance on the multicenter test set 2 was quite comparable to the interrater reference with a mean Dice score of 0.88, as previously established. This level of performance was not achieved by the earlier basic model (mean Dice score of 0.83) of our group (33). Additionally, the segmentation model presented herein demonstrated a reduction in outliers, contributing to more reliable automated segmentation. Further studies are required to ascertain whether the model is capable of accurately segmenting all relevant pathologic BM in other diseases where WB-MRI is a valuable tool for disease evaluation, such as in prostate and breast cancer with metastatic bone disease (38–40). Further applications of the segmentation model could include automated assessment of vertebral fractures, and automated tracking of disease

trajectory. The presented segmentation model could also be a crucial first step in a two-step approach for FL detection or segmentation, whereby the search volume for FLs is pre-defined.

On MRI, histogram parameters of the BM have been reported to serve as potential predictors for high-risk patients (41), defined by the revised international staging system (7), and high-risk cytogenetic aberrations (42) in MM. Studies on radiomics in (S)MM have demonstrated that BM contains predictive information on several clinical aspects, including PCI (25), high-risk cytogenetic status (43,44), treatment assessment (45–47), and survival (48). Within these studies, radiomics models were established based on a limited number of selected BMS, which were segmented from MRIs with corresponding MPCD diagnoses. Consequently, these models were unable to leverage the information from the entire BM for their predictions, instead including information from a

very limited and potentially non-representative sample. Conversely, our automated advanced segmentation model precisely encompasses most of the myeloma-relevant BM, thereby highlighting its capacity to capture the heterogeneous distribution of tumor load in MM (3,4). Future radiomics models may incorporate whole-body BM in a fully automated concept, paving the way for more comprehensive radiomics models. Additionally, the automated and precise segmentation of whole-body BM could allow to compute whole-body novel myeloma imaging markers.

This study has several limitations, including its retrospective design. Despite achieving state-of-the-art segmentation results for 30 BMS in T1w MRI, the current model does not yet encompass BM segmentation of the lower legs, skull, shoulder girdle, or rib cage. The automated segmentations of individual thoracic BMSs did not consistently align with the interrater reference. The anatomical proximity and similar shape of the thoracic BMS and their corresponding thoracic bones may contribute to this challenge. However, the mean Dice score of the individual BMSs was comparable to the reference, with the automatic segmentations successfully incorporating myeloma-relevant pathologies. While the presence of implants may influence segmentation accuracy, this was not systematically evaluated in our study. This study was limited to T1w coronal MRI data, which is currently not part of the WB-MRI protocol according to MY-RADS (49), but is a central standard sequence included in the core clinical protocol of our centers and in ongoing multicentric trials in our country. BM segmentation might benefit from integrating multi-sequence MRIs or even other imaging modalities like CT for better anatomical detail. Future studies will aim to validate the established model prospectively.

In conclusion, this study presents an automated segmentation model with very high segmentation performance for the BM from T1w WB-MRIs in MM, including individual labeling of each bone. Most importantly, the automated segmentation performance remained consistently high even across multicenter test cases with high tumor load, performing at a level comparable to the interrater variability between radiologists.

FUNDING SOURCES

The authors state that this work has not received any funding. M.W. has been funded by a fellowship of the DKFZ Clinician Scientist Program, supported by the Dieter Morszeck Foundation. F.B. was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)—project number 538209205.

DATA SHARING STATEMENT

The data used for this study is not publicly available, because it contains protected patient health information and is subject to EU data protection laws.

DECLARATION OF COMPETING INTEREST

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Marina Hajianni reports a relationship with Sanofi that includes: speaking and lecture fees. Sandra Sauer reports a relationship with Celgene, BMS, Janssen, Takeda, and Amgen that includes: speaking and lecture fees and travel reimbursement. Marc-Steffen Raab reports a relationship with BMS, Amgen, GSK, Janssen, Sanofi, Pfizer, AbbVie, Novartis, and Roche that includes: consulting or advisory. Marc-Steffen Raab reports a relationship with Sanofi that includes: funding grants. Marc-Steffen Raab reports a relationship with BMS, AbbVie, Janssen, Sanofi, and GSK that includes: travel reimbursement. Marc-Steffen Raab reports a relationship with BMS, Janssen, GSK, AbbVie, and Sanofi that includes: speaking and lecture fees. Marc-Steffen Raab reports a relationship with Novartis and Sanofi that includes: non-financial support. Elias K. Mai reports a relationship with Amgen, Bristol Myers Squibb Celgene, GlaxoSmithKline, Janssen-Cilag, Pfizer, Sanofi, Stemline, and Takeda that includes: consulting or advisory. Elias K. Mai reports a relationship with Amgen, Bristol Myers Squibb Celgene, GlaxoSmithKline, Janssen-Cilag, Pfizer, Sanofi, Stemline, and Takeda that includes: speaking and lecture fees. Elias K. Mai reports a relationship with Bristol Myers Squibb Celgene, GlaxoSmithKline, Janssen-Cilag, Sanofi, and Takeda that includes: funding grants. Elias K. Mai reports a relationship with Bristol Myers Squibb Celgene, GlaxoSmithKline, Janssen-Cilag, Sanofi, Stemline, and Takeda that includes: travel reimbursement. Hartmut Goldschmidt reports a relationship with Amgen, Array Biopharma Pfizer, BMS, Celgene, Chugai, Dietmar-Hopp-Foundation, Janssen, Johns Hopkins University, Mundipharma GmbH, Sanofi, GlycoMimetics Inc., GSK, Heidelberg Pharma, Hoffmann-La Roche, Karyopharm, Incyte, millennium Pharmaceuticals Inc., Molecular Partners, Merck Sharp and Dohme, MorphoSys AG, Takeda, and Novartis that includes: funding grants and non-financial support. Hartmut Goldschmidt reports a relationship with Adaptive Biotechnology, Amgen, BMS, Janssen, and Sanofi that includes: consulting or advisory. Hartmut Goldschmidt reports a relationship with Amgen, BMS, Chugai, GlaxoSmithKline, Janssen, Novartis, Sanofi, and Pfizer that includes: speaking and lecture fees. Heinz-Peter Schlemmer reports a relationship with Siemens, Curagita, Profound, and Bayer that includes: consulting or advisory, travel reimbursement, and speaking and lecture fees. Heinz-Peter Schlemmer reports a relationship with Curagita that includes: board membership. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

ACKNOWLEDGMENTS

The authors thank Dr Oyunbileg von Stackelberg, Dr Ekaterina Menis, Barbara Wagner, and Kerstin Scherbaum-

Lawrenz for administrative support. The authors thank the German-speaking Myeloma Multicenter Group (GMMG), as well as Prof Dr Doris Kraemer, Prof Dr Yon-Dschun Ko, Prof Dr Ullrich Graeven, Dr Britta Besemer, Prof Dr Marius Stefan Horger, Dr Thomas Geer, and Dr Jörg Thomas Bittenbring for providing data acquired within the GMMG-HD-7 study (EudraCT: 2017–004768–37) at centers 3 to 8 for the multicentric testing performed in this study.

APPENDIX A. SUPPORTING INFORMATION

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.acra.2024.12.060.

REFERENCES

- Kumar SK, Rajkumar V, Kyle RA, et al. Multiple myeloma. *Nat Rev Dis Primers* 2017; 3:17046.
- Rasche L, Chavan SS, Stephens OW, et al. Spatial genomic heterogeneity in multiple myeloma revealed by multi-region sequencing. *Nat Commun* 2017; 8(1):268.
- Bauer F, Sauer S, Weinhold N, et al. (Smoldering) multiple myeloma: mismatch between tumor load estimated from bone marrow biopsy at iliac crest and tumor load shown by MRI. *Skeletal Radiol* 2023; 52(12):2513–2518.
- Latifoltojar A, Boyd K, Riddell A, et al. Characterising spatial heterogeneity of multiple myeloma in high resolution by whole body magnetic resonance imaging: towards macro-phenotype driven patient management. *Magn Reson Imaging* 2021; 75:60–64.
- Mateos MV, Kumar S, Dimopoulos MA, et al. International Myeloma Working Group risk stratification model for smoldering multiple myeloma (SMM). *Blood Cancer J* 2020; 10(10):102.
- D'Agostino M, Cairns DA, Lahuerta JJ, et al. Second Revision of the International Staging System (R2-ISS) for Overall Survival in Multiple Myeloma: A European Myeloma Network (EMN) Report Within the HARMONY Project. *J Clin Oncol* 2022; 40(29):3406–3418.
- Palumbo A, Avet-Loiseau H, Oliva S, et al. Revised international staging system for multiple myeloma: a report from international myeloma working group. *J Clin Oncol* 2015; 33(26):2863–2869.
- Walker R, Barlogie B, Haessler J, et al. Magnetic resonance imaging in multiple myeloma: diagnostic and clinical implications. *J Clin Oncol* 2007; 25(9):1121–1128.
- Hillengass J, Fechtner K, Weber MA, et al. Prognostic significance of focal lesions in whole-body magnetic resonance imaging in patients with asymptomatic multiple myeloma. *J Clin Oncol* 2010; 28(9):1606–1610.
- Mai EK, Hielscher T, Kloth JK, et al. A magnetic resonance imaging-based prognostic scoring system to predict outcome in transplant-eligible patients with multiple myeloma. *Haematologica* 2015; 100(6):818–825.
- Hillengass J, Landgren O. Challenges and opportunities of novel imaging techniques in monoclonal plasma cell disorders: imaging "early myeloma". *Leuk Lymphoma* 2013; 54(7):1355–1363.
- Mouloupoulos LA, Dimopoulos MA, Christoulas D, et al. Diffuse MRI marrow pattern correlates with increased angiogenesis, advanced disease features and poor prognosis in newly diagnosed myeloma treated with novel agents. *Leukemia* 2010; 24(6):1206–1212.
- Wennmann M, Hielscher T, Kintzele L, et al. Analyzing longitudinal wb-MRI data and clinical course in a cohort of former smoldering multiple myeloma patients: connections between MRI findings and clinical progression patterns. *Cancers* 2021; 13(5):961.
- Rajkumar SV, Dimopoulos MA, Palumbo A, et al. International Myeloma Working Group updated criteria for the diagnosis of multiple myeloma. *Lancet Oncol* 2014; 15(12):e538–e548.
- Hillengass J, Usmani S, Rajkumar SV, et al. International myeloma working group consensus recommendations on imaging in monoclonal plasma cell disorders. *Lancet Oncol* 2019; 20(6):e302–e312.
- Rasche L, Angtuaco EJ, Alpe TL, et al. The presence of large focal lesions is a strong independent prognostic factor in multiple myeloma. *Blood* 2018; 132(1):59–66.
- Wennmann M, Kintzele L, Piraud M, et al. Volumetry based biomarker speed of growth: Quantifying the change of total tumor volume in whole-body magnetic resonance imaging over time improves risk stratification of smoldering multiple myeloma patients. *Oncotarget* 2018; 9(38):25254–25264.
- Piraud M, Wennmann M, Kintzele L, et al. Towards quantitative imaging biomarkers of tumor dissemination: a multi-scale parametric modeling of multiple myeloma. *Med Image Anal* 2019; 57:214–225.
- Latifoltojar A, Hall-Craggs M, Bainbridge A, et al. Whole-body MRI quantitative biomarkers are associated significantly with treatment response in patients with newly diagnosed symptomatic multiple myeloma following bortezomib induction. *Eur Radiol* 2017; 27(12):5325–5336.
- Merz M, Hielscher T, Mai EK, et al. Cystic transformation of focal lesions after therapy is associated with remission but adverse outcome in myeloma. *Blood Cancer J* 2019; 9(9):71.
- Aerts HJ, Velazquez ER, Leijenaar RT, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nat Commun* 2014; 5:4006.
- LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature* 2015; 521(7553):436–444.
- Hosny A, Parmar C, Quackenbush J, et al. Artificial intelligence in radiology. *Nat Rev Cancer* 2018; 18(8):500–510.
- Wennmann M, Neher P, Stanczyk N, et al. Deep learning for automatic bone marrow apparent diffusion coefficient measurements from whole-body magnetic resonance imaging in patients with multiple myeloma: a retrospective multicenter study. *Invest Radiol* 2023; 58(4):273–282.
- Wennmann M, Ming W, Bauer F, et al. Prediction of bone marrow biopsy results from MRI in multiple myeloma patients using deep learning and radiomics. *Invest Radiol* 2023; 58(10):754–765.
- Candito A, Holbrey R, Ribeiro A, et al. Deep learning assisted atlas-based delineation of the skeleton from whole-body diffusion weighted MRI in patients with malignant bone disease. *Biomed Signal Process Control* 2024; 92:106099.
- Lavdas I, Glocker B, Kamnitsas K, et al. Fully automatic, multiorgan segmentation in normal whole body magnetic resonance imaging (MRI), using classification forests (CFs), convolutional neural networks (CNNs), and a multi-atlas (MA) approach. *Med Phys* 2017; 44(10):5210–5220.
- Arabi H, Zaidi H. Whole-body bone segmentation from MRI for PET/MRI attenuation correction using shape-based averaging. *Med Phys* 2016; 43(11):5848.
- Liu X, Han C, Wang H, et al. Fully automated pelvic bone segmentation in multiparametric MRI using a 3D convolutional neural network. *Insights Imaging* 2021; 12(1):93.
- Pang S, Pang C, Zhao L, et al. SpineParseNet: spine parsing for volumetric MR image by a two-stage segmentation framework with semantic image representation. *IEEE Transac Med Imaging* 2021; 40(1):262–273.
- van der Graaf JW, van Hooff ML, Buckens CFM, et al. Lumbar spine segmentation in MR images: a dataset and a public benchmark. *Sci Data* 2024; 11(1):264.
- Ceranka J, Verga S, Kvasnytsia M, et al. Multi-atlas segmentation of the skeleton from whole-body MRI—Impact of iterative background masking. *Magn Reson Med* 2020; 83(5):1851–1862.
- Wennmann M, Klein A, Bauer F, et al. Combining deep learning and radiomics for automated, objective, comprehensive bone marrow characterization from whole-body MRI: a multicentric feasibility study. *Invest Radiol* 2022; 57(11):752–763.
- Wennmann M, Goldschmidt H, Mosebach J, et al. Whole-body magnetic resonance imaging plus serological follow-up for early identification of progression in smoldering myeloma patients to prevent development of end-organ damage. *Br J Haematol* 2022; 199(1):65–75.
- Goldschmidt H, Mai EK, Bertsch U, et al. Addition of isatuximab to lenalidomide, bortezomib, and dexamethasone as induction therapy for newly diagnosed, transplantation-eligible patients with multiple myeloma (GMMG-HD7): part 1 of an open-label, multicentre, randomised, active-controlled, phase 3 trial. *Lancet Haematol* 2022; 9(11):e810–e821.
- Nolden M, Zelzer S, Seitel A, et al. The Medical Imaging Interaction Toolkit: challenges and advances: 10 years of open-source development. *Int J Comput Assist Radiol Surg* 2013; 8(4):607–620.

37. Isensee F, Jaeger PF, Kohl SAA, et al. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat Methods* 2021; 18(2):203–211.
38. Padhani AR, Lecouvet FE, Tunariu N, et al. METastasis reporting and data system for prostate cancer: practical guidelines for acquisition, interpretation, and reporting of whole-body magnetic resonance imaging-based evaluations of multiorgan involvement in advanced prostate cancer. *Eur Urol* 2017; 71(1):81–92.
39. Petralia G, Padhani AR, Pricolo P, et al. Whole-body magnetic resonance imaging (WB-MRI) in oncology: recommendations and key uses. *Radiol Med* 2019; 124(3):218–233.
40. Petralia G, Koh DM, Attariwala R, et al. Oncologically relevant findings reporting and data system (ONCO-RADS): guidelines for the acquisition, interpretation, and reporting of whole-body MRI for cancer screening. *Radiology* 2021; 299(3):494–507.
41. Wang Q, Zhang L, Li S, et al. Histogram analysis based on apparent diffusion coefficient maps of bone marrow in multiple myeloma: an independent predictor for high-risk patients classified by the revised international staging system. *Acad Radiol* 2022; 29(6):e98–e107.
42. Li J, Wang Q, Zhou J, et al. Total tumor ADC histogram analysis: a new tool for predicting high-risk cytogenetic abnormalities in multiple myeloma patients. *Acad Radiol* 2024; 31(11):4560–4567.
43. Liu J, Zeng P, Guo W, et al. Prediction of high-risk cytogenetic status in multiple myeloma based on magnetic resonance imaging: utility of radiomics and comparison of machine learning methods. *J Magn Reson Imaging* 2021; 54(4):1303–1311.
44. Van Den Berghe T, Verberckmoes B, Kint N, et al. Predicting cytogenetic risk in multiple myeloma using conventional whole-body MRI, spinal dynamic contrast-enhanced MRI, and spinal diffusion-weighted imaging. *Insights Imaging* 2024; 15(1):106.
45. Ekert K, Hinterleitner C, Baumgartner K, et al. Extended texture analysis of non-enhanced whole-body MRI image data for response assessment in multiple myeloma patients undergoing systemic therapy. *Cancers* 2020; 12(3):761.
46. Li Y, Yin P, Liu Y, et al. Radiomics models based on magnetic resonance imaging for prediction of the response to bortezomib-based therapy in patients with multiple myeloma. *Biomed Res Int* 2022; 2022:6911246.
47. Wu Z, Wang H, Zheng Y, et al. Lumbar MR-based radiomics nomogram for detecting minimal residual disease in patients with multiple myeloma. *Eur Radiol* 2023; 33(8):5594–5605.
48. Li Y, Liu Y, Yin P, et al. MRI-based bone marrow radiomics nomogram for prediction of overall survival in patients with multiple myeloma. *Front Oncol* 2021; 11:709813.
49. Messiou C, Hillengass J, Delorme S, et al. Guidelines for acquisition, interpretation, and reporting of whole-body MRI in myeloma: myeloma response assessment and diagnosis system (MY-RADS). *Radiology* 2019; 291(1):5–13.