

# **Investigation of deep learning approaches for automated analysis in radiological cross-sectional imaging**

Doctoral thesis

to obtain a doctorate (PhD)

from the Faculty of Medicine

of the University of Bonn

**Maike Theis**

from Bonn, Germany

2025

Written with authorization of  
the Faculty of Medicine of the University of Bonn

First reviewer: Univ.-Prof. Dr. med. Ulrike I. Attenberger

Second reviewer: Prof. Dr. rer. nat. Jürgen Hesser

Day of oral examination: 05.03.2025

From the Department of Diagnostic and Interventional Radiology,  
University Hospital Bonn

## Table of Contents

	<b>List of abbreviations</b>	<b>4</b>
<b>1.</b>	<b>Abstract</b>	<b>5</b>
<b>2.</b>	<b>Introduction and aims with references</b>	<b>6</b>
2.1	Artificial intelligence in radiology	6
2.2	Targeted quantitative analysis	8
2.3	Opportunistic quantitative analysis	8
2.4	Direct image-based survival prediction	9
2.5	Aim	10
2.6	References	11
<b>3.</b>	<b>Publications</b>	<b>16</b>
3.1	Publication 1: Deep learning enables automated MRI-based estimation of uterine volume also in patients with uterine fibroids undergoing high-intensity focused ultrasound therapy	16
3.2	Publication 2: End-to-end automated body composition analyses with integrated quality control for opportunistic assessment of sarcopenia in CT	26
3.3	Publication 3: Direct deep learning-based survival prediction from pre-interventional CT prior to transcatheter aortic valve replacement	37
<b>4.</b>	<b>Discussion with references</b>	<b>46</b>
4.1	Discussion and conclusion	46
4.2	References	50
<b>5.</b>	<b>Acknowledgement</b>	<b>53</b>

## List of abbreviations

AI	Artificial intelligence
BCA	Body composition analysis
CNN	Convolutional neural network
CPHM	Cox proportional hazards model
DL	Deep learning
EAT	Epicardial adipose tissue
HIFU	High-intensity focused ultrasound
ML	Machine learning
TAVR	Transcatheter aortic valve replacement

## 1. Abstract

Artificial intelligence (AI)-based methods are nowadays an integral part of medical research and are also applied in radiology, for example, for the automation of quantitative image analysis. Automated evaluation tools are of particular interest in the analysis of radiological images, as quantitative assessment usually requires the segmentation of specific organs and tissues, making manual analysis very time-consuming and therefore difficult to integrate into everyday clinical practice. Moreover, automatic analysis facilitates large-scale assessment and an opportunistic application, which may contribute to the identification of biomarkers that provide new insights into different pathologies. This thesis focuses on the development of deep learning (DL) methods in radiological cross-sectional imaging and presents three studies dealing with the automation of quantitative measurements.

The first study investigates a DL approach for targeted analysis of uterine volume in T2-weighted MRI scans from patients with leiomyomas by automatically segmenting uterine tissue. This method enables monitoring uterine volume before and after high-intensity focused ultrasound (HIFU) treatment to assess the response to therapy. In the second work, a dual-center study is conducted to develop an automatic assessment of body composition based on CT scans. Body composition is performed on 2D abdominal slices at L3/L4 lumbar level, which is why automatic slice extraction is also investigated in the pipeline next to tissue segmentation. Additionally, quality control methods are integrated to ensure fully automatic application in the clinical routine. Lastly, a DL model is developed that directly predicts the survival of patients undergoing transcatheter aortic valve replacement (TAVR) based on quantitative image features extracted from abdominal CT slices.

This thesis provides insight into the achievable robustness of DL models for both automatic tissue segmentation and direct image-based outcome analysis. To this end, several techniques for efficient AI training are employed, such as data augmentation or pre-training strategies. The successfully automated quantitative analyses enable application in a large-scale setting as well as opportunistic analyses for further investigation of quantitative imaging markers in future studies.

## 2. Introduction and aims with references

### 2.1 Artificial intelligence in radiology

The use of AI-based methods in medicine is rapidly increasing and has become an essential part of medical research. In radiology, AI approaches are used in a wide range of applications, e.g. image reconstruction, computer-aided diagnosis, workflow optimization, patient monitoring, or risk modeling (Castiglioni et al., 2021; Hosny et al., 2018; Thrall et al., 2018). One example of AI-based image reconstruction is the AUTOMAP framework presented by Zhu et al. (2018), a deep neural network for image reconstruction that addresses multiple problems, such as artifact reduction. In general, AI approaches can detect complex patterns from biomedical image data and thus not only automate a quantitative image analysis but can also extract new information from the biomedical image that may be related to a specific pathology (Gillies et al., 2016; Hosny et al., 2018). An important goal of computer-aided diagnosis is to identify image features that are associated with a specific disease, thereby enabling a comprehensive characterization of the pathological changes. This has already proven successful in the diagnosis of intestinal polyps or in mammography (Komeda et al., 2017; Rodríguez-Ruiz et al. 2019). AI-based methods thus contribute to data-driven medicine, which enables more individualized patient care (Castiglioni et al., 2021; Hosny et al., 2018).

In 2016, Gillies et al. presented the concept of Radiomics, a classic machine learning (ML) method to automate quantitative image analysis that has especially proven itself in oncological imaging (Colen et al., 2021; Gillies et al., 2016; Kniep et al., 2019; Limkin et al., 2017). In Radiomics analysis, a set of predefined quantitative features, for example, based on image texture are extracted from a region of interest in the image and can then be used to build classification systems for a certain disease (Gillies et al., 2016). However, the main drawback of this method is the need for standardized imaging and the fact that the extracted quantitative features are predefined and therefore may not always represent the most appropriate characteristics for the underlying research question. Also, feature selection and calculation require standardization to enable widespread application (Hosny et al., 2018; Limkin et al., 2017).

In contrast, DL methods are able to autonomously learn abstract data representations and automatically select relevant image features depending on the task at hand. When it comes to processing image data, the convolutional neural network (CNN) in particular has

established itself as a specific DL architecture that can be used to address various tasks such as image classification, semantic segmentation, or object detection (Castiglioni et al., 2021; Hosny et al., 2018; LeCun et al., 2015). To be able to model complex non-linear relationships, DL models require a large amount of data to avoid the risk of overfitting and ensure a well-generalized model. As most CNNs are trained in a supervised fashion, this aspect is a major weakness of DL models as data annotation is expensive, time-consuming, and often requires a high level of expertise, especially in the medical domain. However, there are various techniques to counteract this weakness, such as the use of data augmentation to increase the variation in the training data or the use of transfer learning, where a previously trained model is applied to another related task (Castiglioni et al., 2021). For this, the data does not necessarily have to come from the same domain, e.g. a CNN designed to detect liver cirrhosis from MRI data was successfully pre-trained on the ImageNet dataset (Nowak et al., 2021; Russakovsky et al., 2015). Furthermore, data compression methods, such as autoencoders, can be used to prevent overfitting on high-dimensional medical image data. Autoencoders extract the most essential features from the input data and restore the original input based on this compressed data representation. This unsupervised learning method is particularly efficient as no labeled data is required (Chen et al., 2017; Wolf et al., 2023).

In general, the ability of DL models to independently identify relevant features and recognize complex patterns in the image may allow the extraction of new biomarkers from radiological images, i.e., a previously unrecognized quantitative parameter that correlates with a specific clinical outcome (Hosny et al., 2018; Thrall et al., 2018). They enable, among other things, the automatic extraction of quantitative information from medical image data that otherwise might not be used at all. Such quantitative analyses can either be used in a targeted manner to automatically extract a known quantitative biomarker or opportunistically, where various parameters are collected without a known association to a clinical endpoint. Instead of using DL methods for the extraction of specific quantitative parameters for later outcome analysis, another possibility is to implement direct predictive models. In this process, a DL model directly extracts quantitative information to predict a certain clinical outcome, like survival time (Castiglioni et al., 2021).

## 2.2 Targeted quantitative analysis

In targeted quantitative analysis, a specific parameter is collected that is either an already known biomarker or is used for the investigation of a specific research question. Manually collecting these parameters can be time-consuming, especially if they are to be collected on a large scale, and often requires expert knowledge. The collection of image-based parameters usually requires the segmentation of specific tissues, which limits quantitative analysis in large cohorts. An automatic assessment based on AI methods is therefore desirable and may also ensure a more objective, reproducible evaluation.

One example to be investigated in this work is the measurement of uterine volume in T2-weighted MRI of patients with uterine fibroids who have undergone HIFU therapy. The minimally invasive HIFU treatment of symptomatic fibroids significantly reduces uterine volume which leads to an improvement in leiomyoma-associated symptoms (Kim et al., 2011; Hindley et al., 2004; Marinova et al., 2021). It is thus of great interest to have an objective measure of the response to therapy. However, the exact determination of uterine volume would require manual outlining of the contours in each axial slice and is therefore often only estimated based on diameter measurements using the prolate ellipsoid formula (Kung and Chang, 1996; Marinova et al., 2021). Automated segmentation of uterine tissue using a DL model has the potential to provide rapid and reproducible evaluation, facilitating robust assessment in all patients.

## 2.3 Opportunistic quantitative analysis

AI methods in medicine have also facilitated the automation of opportunistic analyses. Such studies intend to acquire information that is generally not evaluated in routine clinical practice but that may positively influence patient care. However, to find new biomarkers through opportunistic analysis, a survey of large cohorts is essential. One prominent example is body composition analysis (BCA), which has proven its prognostic value in recent years. Research on oncological diseases has demonstrated an association between body composition and both survival and chemotherapy toxicity in cancer patients (Faron et al. 2021; Nowak et al., 2024; Prado et al., 2008). In addition, body composition markers have been identified as predictive values of several cardiovascular events, including heart failure, cardiogenic shock, and severe aortic stenosis (Kenchiah et al., 2002; Luetkens et al., 2020; Salam et al., 2023). In the past, BCA was carried out using



methods such as hydrostatic weighing or air displacement plethysmography. At present, an evaluation based on radiological image data such as CT or MRI has become the gold standard, which enables a precise analysis of fat and muscle distribution in the body (Biaggi et al., 1999; Borga et al., 2018; Cruz-Jentoft et al. 2010, Fields et al. 2002). Since manual BCA takes huge effort due to required tissue segmentation, it is usually not performed in the entire dataset but only on specific single-slice images (Borga et al., 2018; Faron et al., 2020). Assessing BCA from radiological cross-sectional data also enables automatic quantitative image analysis based on automatic tissue segmentation by DL approaches (Castiglione et al. 2021; Magudia et al., 2021; Nowak et al., 2020; Weston et al., 2019). However, previously proposed methods do not yet allow for a fully automated end-to-end analysis that enables opportunistic assessment in routine clinical practice, as either manual selection of a single slice at a certain anatomical landmark is required or manual quality control is needed to identify cases where the DL approach has failed. To achieve this, this dissertation investigates the development of an end-to-end automated pipeline for opportunistic BCA with integrated quality control.

#### 2.4 Direct image-based survival prediction

The features extracted from the quantitative image analysis are then used to evaluate their prognostic value for a specific clinical endpoint. For this purpose, either classical ML models or further DL approaches are developed. However, another possibility is to build models that directly predict the clinical endpoint from medical image data. Using DL models for direct survival prediction based on different clinical and laboratory parameters has already been examined and proved to be advantageous compared to other ML approaches such as Cox proportional hazards models (CPHM) or random survival forests (Katzman et al., 2018, Kim et al., 2019, Vale-Silva and Rohr, 2021). Moreover, Starke et al. (2020) investigated direct image-based prediction of loco-regional tumor control based on CT data using 2D and 3D CNNs. However, the direct DL-based survival prediction based solely on imaging has not yet been investigated. This thesis presents one application example, namely the survival prediction of patients with severe aortic stenosis who have undergone TAVR. In this patient group, body composition markers extracted from an abdominal slice at the L3/L4 lumbar level were shown to have a predictive value for survival prediction (Luetkens et al., 2020). Rather than extracting predefined markers

of body composition, this work explores the use of DL models to autonomously identify relevant image features for direct image-based survival time prediction from abdominal CT slices.

## 2.5 Aim

This thesis aims to investigate the use of DL models to automate quantitative image analysis. To this end, a robust model for targeted quantitative assessment of uterine volume in patients with uterine fibroids and an end-to-end pipeline for automated BCA, allowing for opportunistic analysis in large-scale cohorts, should be developed. Finally, the use of DL methods for direct image-based survival prediction should be investigated.

## 2.6 References

- Biaggi RR, Vollman MW, Nies MA, Brener CE, Flakoll PJ, Levenhagen DK, Sun M, Karabulut Z, Chen KY. Comparison of air-displacement plethysmography with hydrostatic weighing and bioelectrical impedance analysis for the assessment of body composition in healthy adults. *Am J Clin Nutr* 1999; 69(5): 898-903
- Borga M, West J, Bell JD, Harvey NC, Romu T, Heymsfield SB, Dahlqvist Leinhard O. Advanced body composition assessment: from body mass index to body composition profiling. *J Invest Med* 2018; 66(5): 1-9
- Castiglione J, Somasundaram E, Gilligan LA, Trout AT, Brady S. Automated segmentation of abdominal skeletal muscle on pediatric CT scans using deep learning. *Radiol Artif Intel* 2021; 3(2): e200130
- Castiglioni I, Rundo L, Codari M, Di Leo G, Salvatore C, Interlenghi M, Gallivanone F, Cozzi A, D'Amico NC, Sardanelli F. AI applications to medical images: From machine learning to deep learning. *Phys Med* 2021; 83: 9-24
- Chen M, Shi X, Zhang Y, Wu D, Guizani M. Deep feature learning for medical image analysis with convolutional autoencoder neural network. *IEEE Trans Big Data* 2017; 7(4): 750-758
- Colen RR, Rolfo C, Ak M, Ayoub M, Ahmed S, Elshafeey N, Mamindla P, Zinn PO, Ng C, Vikram R, Bakas S, Peterson CB, Rodon Ahnert J, Subbiah V, Karp DD, Stephen B, Hajjar J, Naing A. Radiomics analysis for predicting pembrolizumab response in patients with advanced rare cancers. *J Immunother Cancer* 2021; 9(4): e001752
- Cruz-Jentoft AJ, Baeyens JP, Bauer JM, Boirie Y, Cederholm T, Landi F, Martin FC, Michel JP, Rolland Y, Schneider SM, Topinková E, Vandewoude M, Zamboni M. Sarcopenia: European consensus on definition and diagnosis: Report of the European Working Group on Sarcopenia in Older People. *Age Ageing* 2010; 39(4): 412-423
- Faron A, Sprinkart AM, Kuetting DL, Feisst A, Isaak A, Endler C, Chang J, Nowak S, Block W, Thomas D, Attenberger U, Luetkens JA. Body composition analysis using CT and MRI: intra-individual intermodal comparison of muscle mass and myosteatosis. *Sci Rep* 2020; 10(1): 11765

Faron A, Opheys NS, Nowak S, Sprinkart AM, Isaak A, Theis M, Mesropyan N, Endler C, Sirokay J, Pieper CC, Kuetting D, Attenberger U, Landsberg J, Luetkens JA. Deep learning-based body composition analysis predicts outcome in melanoma patients treated with immune checkpoint inhibitors. *Diagnostics* 2021; 11(12): 2314

Fields DA, Goran MI, McCrory MA. Body-composition assessment via air-displacement plethysmography in adults and children: a review. *Am J Clin Nutr* 2002; 75(3): 453-467

Gillies RJ, Kinahan PE, Hricak H. Radiomics: images are more than pictures, they are data. *Radiology* 2016; 278(2): 563-577

Hindley J, Gedroyc WM, Regan L, Stewart E, Tempany C, Hynnen K, Macdanold N, Inbar Y, Itzchak Y, Rabinovici J, Kim K, Geschwind JF, Hesley G, Gostout B, Ehrenstein T, Hengst S, Sklair-Levy M, Shushan A, Jolesz F. MRI guidance of focused ultrasound therapy of uterine fibroids: early results. *AJR Am J Roentgenol* 2004; 183(6): 1713-1719

Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts HJWL. Artificial intelligence in radiology. *Nat Rev Cancer* 2018; 18(8): 500-510

Katzman JL, Shaham U, Cloninger A, Bates J, Jiang T, Kluger Y. DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC Med Res Methodol* 2018; 18: 24

Kenchiah S, Evans JC, Levy D, Wilson PWF, Benjamin EJ, Larson MG, Kannel WB, Vasan RS. Obesity and the risk of heart failure. *N Engl J Med* 2002; 347(5): 305-313

Kim DW, Lee S, Kwon S, Nam W, Cha IH, Kim HJ. Deep learning-based survival prediction of oral cancer patients. *Sci Rep* 2019; 9(1): 6994

Kim HS, Baik JH, Pham LD, Jacobs MA. MR-guided high-intensity focused ultrasound treatment for symptomatic uterine leiomyomata: long-term outcomes. *Acad Radiol* 2011; 18(8): 970-976

Kniep HC, Madesta F, Schneider T, Hanning U, Schönfeld MH, Schön G, Fiehler J, Gauer T, Werner R, Gellissen S. Radiomics of brain MRI: utility in prediction of metastatic tumor type. *Radiology* 2019; 290(2): 479-487

Komeda Y, Handa H, Watanabe T, Nomura T, Kitahashi M, Sakurai T, Okamoto A, Minami T, Kono M, Arizumi T, Takenaka M, Hagiwara S, Matsui S, Nishida N, Kashida H, Kudo M. Computer-aided diagnosis based on convolutional neural network system for colorectal polyp classification: preliminary experience. *Oncology* 2017; 93(Suppl. 1): 30-34

Kung FT, Chang SY. The relationship between ultrasonic volume and actual weight of pathologic uterus. *Gynecol Obstet Invest* 1996; 42(1): 35-38

LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature* 2015; 521(7553): 436-444

Limkin EJ, Sun R, Dercle L, Zacharaki EI, Robert C, Reuzé S, Schernberg A, Paragios N, Deutsch E, Ferte, C. (2017). Promises and challenges for the implementation of computational medical imaging (radiomics) in oncology. *Ann Oncol* 2017; 28(6): 1191-1206

Luetkens JA, Faron A, Geissler HL, Al-Kassou B, Shamekhi J, Stundl A, Sprinkart AM, Meyer C, Fimmers R, Treede H, Grube E, Nickenig G, Sinning JM, Thomas D. Opportunistic computed tomography imaging for the assessment of fatty muscle fraction predicts outcome in patients undergoing transcatheter aortic valve replacement. *Circulation* 2020; 141(3): 234-236

Magudia K, Bridge CP, Bay CP, Babic A, Fintelman FJ, Troschel FM, Miskin N, Wrobel WC, Brais LK, Andriole KP, Wolpin BM, Rosenthal MH. Population-scale CT-based body composition analysis of a large outpatient population using deep learning to derive age-, sex-, and race-specific reference curves. *Radiology* 2021; 298(2): 319-329

Marinova M, Ghaei S, Recker F, Tonguc T, Kaverina O, Savchenko O, Kravchenko D, Thudium M, Pieper CC, Egger EK, Mustea A, Attenberger U, Conrad R, Hadizadeh DR, Strunk H. Efficacy of ultrasound-guided high-intensity focused ultrasound (USgHIFU) for uterine fibroids: an observational single-center study. *Int J Hyperthermia* 2021; 38(2): 30-38

Nowak S, Faron A, Luetkens JA, Geißler HL, Praktijnjo M, Block W, Thomas D, Sprinkart AM. Fully automated segmentation of connective tissue compartments for CT-

based body composition analysis: a deep learning approach. *Invest Radiol* 2020; 55(6): 357-366

Nowak S, Mesropyan N, Faron A, Block W, Reuter M, Attenberger UI, Luetkens JA, Sprinkart AM. Detection of liver cirrhosis in standard T2-weighted MRI using deep transfer learning. *Eur Radiol* 2021; 31(11): 8807-8815

Nowak S, Kloth C, Theis M, Marinova M, Attenberger UI, Sprinkart AM, Luetkens JA. Deep learning–based assessment of CT markers of sarcopenia and myosteatosi s for outcome assessment in patients with advanced pancreatic cancer after high-intensity focused ultrasound treatment. *Eur Radiol* 2024; 34(1): 279-286

Prado CMM, Lieffers JR, McCargar LJ, Reiman T, Sawyer MB, Martin L, Baracos VE. Prevalence and clinical implications of sarcopenic obesity in patients with solid tumours of the respiratory and gastrointestinal tracts: a population-based study. *Lancet Oncol* 2008; 9(7): 629-635

Rodríguez-Ruiz A, Krupinski E, Mordang JJ, Schilling K, Heywang-Köbrunner SH, Sechopoulos I, Mann RM. Detection of breast cancer with mammography: effect of an artificial intelligence support system. *Radiology* 2019; 290(2): 305-314

Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M, Berg AC, Fei-Fei L. ImageNet large scale visual recognition challenge. *Int J Comput Vis* 2015; 115: 211-252

Salam B, Al Zaidi M, Sprinkart AM, Nowak S, Theis M, Kuetting D, Aksoy A, Nickenig G, Attenberger U, Zimmer S, Luetkens JA. Opportunistic CT-derived analysis of fat and muscle tissue composition predicts mortality in patients with cardiogenic shock. *Sci Rep* 2023; 13(1): 22293

Starke S, Leger S, Zwanenburg A, Leger K, Lohaus F, Linge A, Schreiber A, Kalinauskaite G, Tinhofer I, Guberina N, Guberina M, Balermipas P, von der Grün J, Ganswindt U, Belka C, Peeken JC, Combs SE, Boeke S, Zips D, Richter C, Troost EG, Krause M, Baumann M, Löck S. 2D and 3D convolutional neural networks for outcome modelling of locally advanced head and neck squamous cell carcinoma. *Sci Rep* 2020; 10(1): 15625

Thrall JH, Li X, Li Q, Cruz C, Do S, Dreyer K, Brink J. Artificial intelligence and machine learning in radiology: opportunities, challenges, pitfalls, and criteria for success. *J Am Coll Radiol* 2018; 15(3): 504-508

Vale-Silva LA, Rohr K. Long-term cancer survival prediction using multimodal deep learning. *Sci Rep* 2021; 11(1): 13505

Weston AD, Korfiatis P, Kline TL, Philbrick KA, Kostandy P, Sakinis T, Sugimoto M, Takahashi N, Erickson BJ. Automated abdominal segmentation of CT scans for body composition analysis using deep learning. *Radiology* 2019; 290(3): 669-679

Wolf D, Payer T, Lisson CS, Lisson CG, Beer M, Götz M, Ropinski T. Self-supervised pre-training with contrastive and masked autoencoder methods for dealing with small datasets in deep learning for medical imaging. *Sci Rep* 2023; 13(1): 20260

Zhu B, Liu JZ, Cauley SF, Rosen BR, Rosen MS. Image reconstruction by domain-transform manifold learning. *Nature* 2018; 555(7697): 487-492

### 3. Publications

3.1. Publication 1: Deep learning enables automated MRI-based estimation of uterine volume also in patients with uterine fibroids undergoing high-intensity focused ultrasound therapy

Theis M, Tonguc T, Savchenko O, Nowak S, Block W, Recker F, Essler M, Mustea A, Attenberger U, Marinova M, Sprinkart AM. **Deep learning enables automated MRI-based estimation of uterine volume also in patients with uterine fibroids undergoing high-intensity focused ultrasound therapy.** Insights Imaging 2023; 14(1): 1-9




ORIGINAL ARTICLE

Open Access



# Deep learning enables automated MRI-based estimation of uterine volume also in patients with uterine fibroids undergoing high-intensity focused ultrasound therapy

Maike Theis<sup>1†</sup>, Tolga Tonguc<sup>1†</sup>, Oleksandr Savchenko<sup>1</sup>, Sebastian Nowak<sup>1</sup>, Wolfgang Block<sup>1,2,3</sup>, Florian Recker<sup>4</sup>, Markus Essler<sup>5</sup>, Alexander Mustea<sup>4</sup>, Ulrike Attenberger<sup>1</sup>, Milka Marinova<sup>1,5†</sup> and Alois M. Sprinkart<sup>1\*†</sup> 

## Abstract

**Background:** High-intensity focused ultrasound (HIFU) is used for the treatment of symptomatic leiomyomas. We aim to automate uterine volumetry for tracking changes after therapy with a 3D deep learning approach.

**Methods:** A 3D nnU-Net model in the default setting and in a modified version including convolutional block attention modules (CBAMs) was developed on 3D T2-weighted MRI scans. Uterine segmentation was performed in 44 patients with routine pelvic MRI (standard group) and 56 patients with uterine fibroids undergoing ultrasound-guided HIFU therapy (HIFU group). Here, preHIFU scans ( $n = 56$ ), postHIFU imaging maximum one day after HIFU ( $n = 54$ ), and the last available follow-up examination ( $n = 53$ , days after HIFU:  $420 \pm 377$ ) were included. The training was performed on 80% of the data with fivefold cross-validation. The remaining data were used as a hold-out test set. Ground truth was generated by a board-certified radiologist and a radiology resident. For the assessment of inter-reader agreement, all preHIFU examinations were segmented independently by both.

**Results:** High segmentation performance was already observed for the default 3D nnU-Net (mean Dice score =  $0.95 \pm 0.05$ ) on the validation sets. Since the CBAM nnU-Net showed no significant benefit, the less complex default model was applied to the hold-out test set, which resulted in accurate uterus segmentation (Dice scores: standard group  $0.92 \pm 0.07$ ; HIFU group  $0.96 \pm 0.02$ ), which was comparable to the agreement between the two readers.

**Conclusions:** This study presents a method for automatic uterus segmentation which allows a fast and consistent assessment of uterine volume. Therefore, this method could be used in the clinical setting for objective assessment of therapeutic response to HIFU therapy.

<sup>†</sup>Maike Theis, Tolga Tonguc, Milka Marinova and Alois M. Sprinkart contributed equally to this study.

\*Correspondence: [sprinkart@uni-bonn.de](mailto:sprinkart@uni-bonn.de)

<sup>1</sup> Department of Diagnostic and Interventional Radiology, University Hospital Bonn, Venusberg-Campus 1, 53127 Bonn, Germany  
Full list of author information is available at the end of the article



© The Author(s) 2022. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

### Key points

- Deep learning methods enable accurate segmentation of the uterus in T2-weighted MRI.
- Automatic uterine volumetry is possible in patients with and without leiomyomas.
- Automated volumetry enables an objective assessment of response to high-intensity focused ultrasound therapy.

**Keywords:** Deep learning, Magnetic resonance imaging, Uterus, Leiomyoma

### Background

Uterine fibroids, also known as leiomyomas, are the most common benign pelvic tumors in women of reproductive age. Fibroid-associated symptoms are observed in about one-third of affected patients [1]. Major symptoms are severe and extended menstrual bleeding (hypermenorrhea and dysmenorrhea) that may lead to anemia-associated complications. Depending on size and location, uterine fibroids can also cause pelvic pressure, urinary frequency and even incontinence and can be associated with adverse reproductive outcome. Thus, symptomatic uterine fibroids have a negative impact on daily living and quality of life [2, 3].

Current treatment strategies mainly involve surgical interventions as laparoscopic or hysteroscopic myomectomy and laparoscopic hysterectomy [4–6]. Nowadays, organ-preserving minimally invasive and noninvasive therapies are becoming increasingly important. In recent years, high-intensity focused ultrasound (HIFU), guided by either ultrasound or magnetic resonance tomography, has also emerged as a viable effective and low-risk treatment option for symptomatic uterine fibroids [7–11]. During the HIFU procedure, the uterine fibroids are thermally ablated by concentrating the ultrasound energy inside the fibroid leading to thermal coagulation necrosis and additional cavitation damage [9, 12–16]. Previous studies have shown that HIFU treatment of symptomatic leiomyomas results in a significant reduction in uterine fibroid volume and total uterine volume during follow-up. In addition, a correlation between improvement in fibroid-associated symptoms and reduction in uterine fibroid volume has been demonstrated [7, 10, 11]. Therefore, automation of uterine measurements is highly desirable in order to be able to assess the response to treatment objectively, quickly and reproducibly.

In recent years, the utility of machine learning and, in particular, deep learning methods has been demonstrated for various medical tasks including medical imaging. To date, most deep learning approaches in medical imaging use artificial neural networks trained in a supervised manner, which means that model development requires annotated data with the desired outcome, also known as ground truth. One potential application of such

deep learning models is the automation of quantitative image analysis, which would otherwise require tedious manual effort. In addition, deep learning methods have also shown great potential for detecting and characterizing pathological findings, which could assist radiologists in making the diagnosis [17–21].

Various deep learning methods have also been proposed for volumetry based on medical image segmentation. A successful neural network for various organ and tissue type segmentation is the open-source framework nnU-Net, a self-adapting pipeline based on the U-Net model introduced by Ronneberger et al. [22–24]. To improve the weighting of the feature map signals, convolutional block attention modules (CBAM) have been suggested, which have led to high performance for various classification, object detection, and segmentation tasks, also in combination with a U-Net architecture [25–28].

Very recently, various neural networks have been proposed for uterine segmentation in MRI and ultrasound imaging, where most of them are also based on the U-Net architecture [29–32]. However, none of these has presented a suitable method for accurate automatic volumetry of the uterus, especially when the evaluation of longitudinal data is required to assess treatment response, such as in patients with uterine fibroids undergoing HIFU therapy.

Against this backdrop, the aim of this study was to develop a 3D deep learning method that allows accurate uterine segmentation of patients with and without uterine fibroids and to ensure automatic assessment of changes in uterine volume during HIFU therapy. For this purpose, two neural networks were trained and compared, a standard 3D U-Net and a modified U-Net using additional CBAMs in the encoder, implemented in the nnU-Net framework.

### Methods

#### Dataset

This study was approved by the local Ethics Committees at the Medical Faculty of the Rheinische Friedrich-Wilhelms-Universität Bonn (no. 295/19). Data of 44 consecutive patients without uterine fibroids who

received routine pelvic MRI (standard group, mean age  $38 \pm 13$  years) and of 56 patients with uterine fibroids who underwent ultrasound-guided HIFU therapy (HIFU group, mean age  $43 \pm 6$  years) were included. The only inclusion criteria were the availability of an axial 3D T2-weighted MRI acquired with a turbo spin echo sequence at a 1.5 Tesla scanner (Philips Ingenia) with a slice thickness  $\leq 5$  mm and an in-plane voxel size  $\leq 1$  mm. In the HIFU group three examinations per patient were considered: An examination prior to HIFU intervention (preHIFU,  $n=56$ ), the immediate follow-up examination maximum one day after HIFU therapy (postHIFU\_1,  $n=54$ ), and the last available follow-up examination (postHIFU\_Last,  $n=53$ ). In two cases, there was no early follow-up within one day after HIFU, and in three cases, patients received only one follow-up examination. The mean time interval between HIFU and the last available follow-up examination was  $420 \pm 377$  days (range: [97; 2007]). Overall, a total of 207 scans from 100 patients were used for method development. Additional information on the dataset and the scanning parameters can be found in Table 1.

The ground truth for the preHIFU images was generated by a board-certified radiologist (O.S., 9 years of experience in radiology and 4 years of experience in gynecologic imaging) and additionally by a radiology resident (T.T., in his fourth year of residency with 2 years of experience in gynecologic imaging). Contours of the uterus were outlined in all slices using the open-source software 3D Slicer [33]. To assist the generation of the ground truth for the remaining datasets, a default 3D nnU-Net was trained on the segmentations of the preHIFU dataset from the board-certified radiologist [23]. This early model was applied to all follow-up examinations of the HIFU group and to the standard non-fibroid group, and the predicted segmentations were subsequently adapted manually by the radiology resident or the board-certified radiologist. The board-certified radiologist approved all segmentations of the resident.

**Table 1** Scan and image parameters of the dataset

	Mean	Median	Range
Pixel spacing (mm)	0.39	0.37	[0.33; 0.61]
Spacing between slices (mm)	4.34	4.4	[3.3; 4.95]
Slice thickness (mm)	3.95	4	[3; 4.5]
Matrix size	1004	1024	[704; 1280]
Number of slices	42	40	[40; 60]
Echo time (TE) (ms)	90	90	[90; 90]
Repetition time (TR) (ms)	3922.15	3729.70	[3729.70; 5594.55]

**Table 2** Number of datasets in the training and test sets for the different groups

	Standard group	HIFU group		
		preHIFU	postHIFU_1	postHIFU_Last
Training	36	45	45	43
Test	8	11	9	10

For method development, datasets of both the standard and the HIFU group were randomly divided into 80% training and 20% test cases, where a single patient was included completely either in the training or in the test set, resulting in 169 training and 38 hold-out test cases (see Table 2).

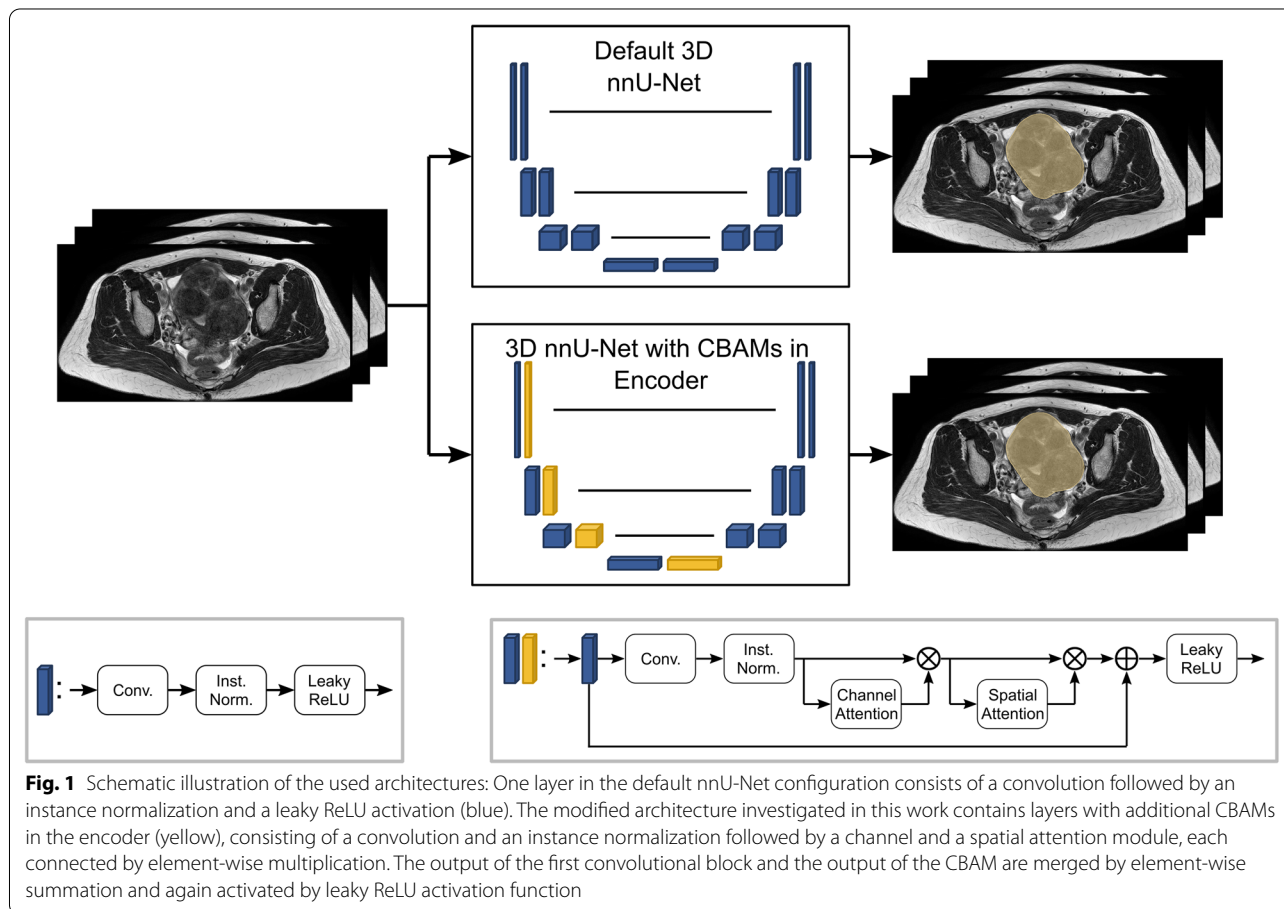
### Model

For automatic uterine segmentation, two different deep learning architectures were trained for 500 epochs based on the 3D nnU-Net framework, one in the default setting and the other with additional CBAMs in the encoder [23, 25]. The default nnU-Net architecture is generated based on the fingerprint of the training dataset, which determines several preferences and parameters, e.g., pre-processing, network depth and the kernel size of the convolutional layers. However, the general architecture of the encoder and decoder always consists of two blocks per resolution step, where one block consists of a convolution, an instance normalization and a leaky rectified linear unit (ReLU) activation [23]. This default architecture was compared to a modified version, where the second block in the encoder was replaced with a CBAM [25]. A CBAM layer returns a weighted feature map, in which important signals should be enhanced and unimportant ones suppressed. In principle, this should improve the focus on the relevant image information and its location [15]. The use of CBAM in combination with a nnU-Net layer is illustrated in Fig. 1.

Using CBAMs in the nnU-Net encoder increases the number of trainable parameters by only 0.12%. This still allows for a relatively fair comparison of both architectures. For more details on the two investigated architectures, see Additional file 1: S1.

### Evaluation

Both models were trained with fivefold cross-validation; thus, the entire training dataset was split into five validation sets and a single model was trained on each of the remaining training data, resulting in five different



models for each of the two investigated methods. The performance of the two architectures was determined based on the mean performance of the five validation sets.

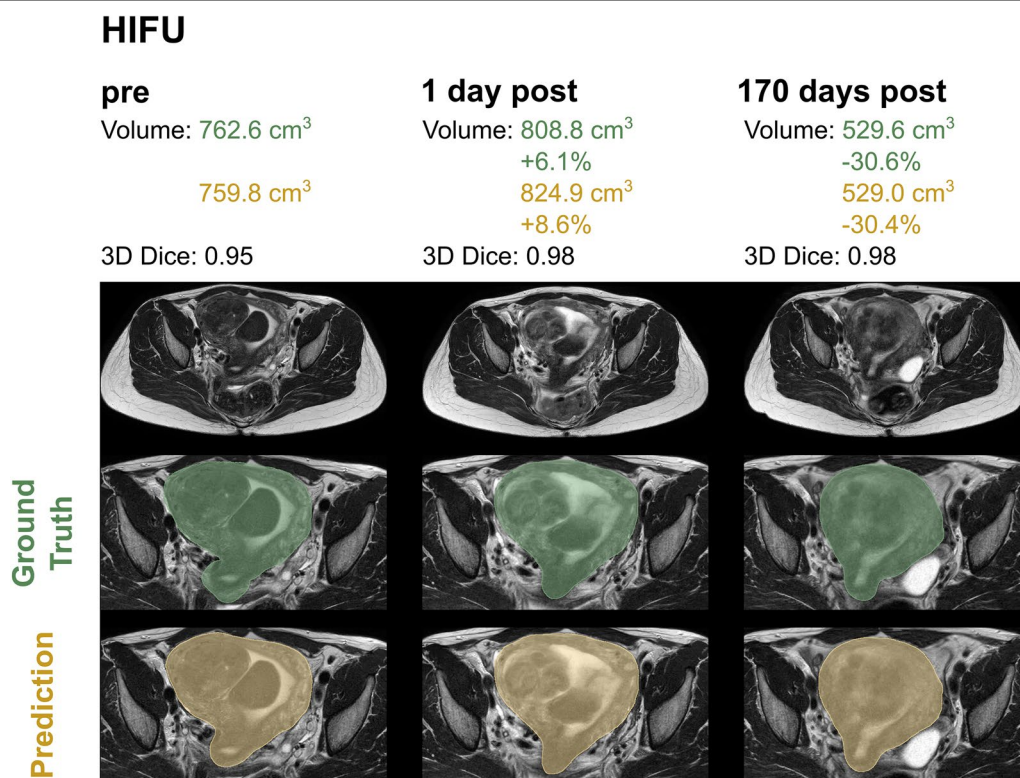
Agreement to the ground truth was measured in terms of Dice score and relative volume difference. The final model was evaluated on the hold-out test data ensembling the predictions of the five individual models from cross-validation. Inter-reader agreement was determined for the preHIFU scans. Model performance was compared to the human inter-reader agreement based on the hold-out test samples from that group.

To investigate whether the model is suitable for post-HIFU treatment follow-up, the automatically determined volume difference between before and after HIFU was compared with the ground truth. All follow-up scans from the HIFU group included in the hold-out test data were considered in this analysis. Pearson correlation coefficient was determined and a Bland–Altman analysis was performed using the Python packages *seaborn* and *pyCompare* [34, 35]

## Results

For both model architectures investigated, an excellent uterine segmentation performance was observed in the standard and all HIFU groups. Figure 2 shows an example for predicted and ground truth uterine segmentation of a patient with uterine fibroids prior to HIFU, short after and 170 days after HIFU treatment. Figure 3 shows three patients of the standard group without uterine fibroids.

The mean performance on each of the five validation sets was very similar for the two models, with a mean relative volume difference of 3.79% for the default nnU-Net and 3.70% for the CBAM nnU-Net and a mean Dice score of 0.95 for both architectures. Since the use of CBAM did not lead to a significant benefit compared to the default nnU-Net configuration for the task of uterus segmentation, the less complex default architecture was chosen as the final segmentation model which was applied to the hold-out test data. A detailed comparison of the two architectures can be found in Additional file 1: S2.



**Fig. 2** Uterine segmentation in a patient undergoing HIFU therapy: before treatment, one day after treatment and 170 days after treatment. Automatic segmentation achieved with the default 3D nnU-Net is shown in yellow and ground truth segmentation validated by a board-certified radiologist in green

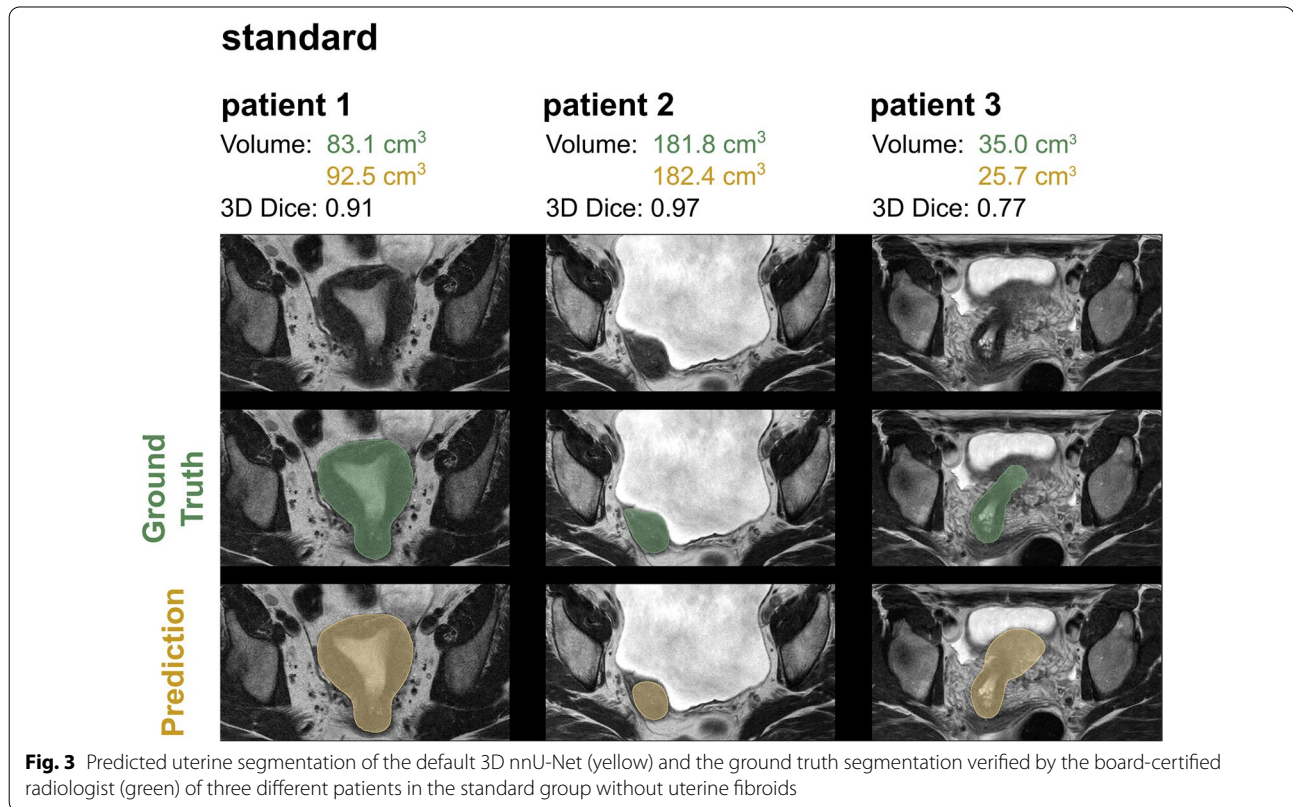
The ensemble of the fivefold cross-validated 3D nnU-Net with default settings applied to the hold-out test data resulted in a mean Dice score of 0.95 and a mean relative volume difference of 4.08% (Table 3).

For the preHIFU dataset, annotated data of both readers were available ( $n = 56$ ). The inter-reader comparison between the board-certified radiologist and the radiology resident on this dataset showed a mean Dice score of 0.92 and a mean relative volume difference of 4.03%. The comparison of the inter-reader agreement and the agreement of both readers with the final segmentation model on the preHIFU test data is listed in Table 4. Comparing the predictions of the neural network with the segmentations of the two readers shows a mean Dice score of 0.91 and higher, indicating a segmentation performance of the neural network similar to human performance.

The agreement between ground truth and automatically determined volume difference before and after HIFU treatment was compared using all follow-up data of the HIFU group included in the hold-out test ( $n = 19$ ). A strong correlation was observed with a Pearson correlation coefficient of 0.99. The Bland–Altman analysis shows a mean difference of  $-1.08 \text{ cm}^3$  (see Fig. 4).

## Discussion

This work presents an innovative method for deep learning-based 3D segmentation of the uterus for automatic and accurate determination of uterine volume in T2-weighted MRI scans with the focus on being applicable in uterine fibroid patients undergoing HIFU treatment. To achieve this, a dataset was used for method development and evaluation that included a relevant portion of examinations prior to HIFU therapy and at different timepoints thereafter. In order to enrich the dataset and to investigate the applicability of uterus segmentation to routine MRI scans, standard examinations from clinical routine were also included. The network performed very well for both extremes, i.e., patients without fibroids and patients with multiple symptomatic fibroids of different sizes who were candidates for HIFU ablation. Thus, it may be assumed (although not proven in this study) that high-quality uterine segmentation can be achieved also in patients with smaller, non-symptomatic fibroids. The performance in the HIFU group was slightly higher than that in the standard group. This may be explained in part by the lower absolute uterine volume in patients without uterine fibroids, which results in lower Dice scores in



**Table 3** Mean Dice scores and mean relative volume difference reported for the entire hold-out test data and separately for the preHIFU dataset, the early (postHIFU\_1) and last follow-up (postHIFU\_Last) after HIFU treatment, as well as for the non-fibroid standard group

Dataset	Dice score	Relative volume difference
All (n = 38)	0.95 ± 0.04	4.08% ± 4.86%
preHIFU (n = 11)	0.94 ± 0.02	3.63% ± 2.91%
postHIFU_1 (n = 9)	0.97 ± 0.01	2.15% ± 1.50%
postHIFU_Last (n = 10)	0.97 ± 0.02	3.12% ± 2.93%
Standard group (n = 8)	0.92 ± 0.07	8.09% ± 8.61%

**Table 4** Comparison of inter-reader agreement between the board-certified radiologist (reader 1) and the intensively trained radiology resident (reader 2) and the respective agreement of these readers with the predicted segmentations of the nnU-Net on the preHIFU data in the hold-out test set (n = 11)

	Dice score	Relative volume difference
Reader 1 vs. Reader 2	0.92 ± 0.02	3.69% ± 3.54%
Reader 1 vs. nnU-Net	0.94 ± 0.02	3.63% ± 2.91%
Reader 2 vs. nnU-Net	0.91 ± 0.03	4.49% ± 5.26%

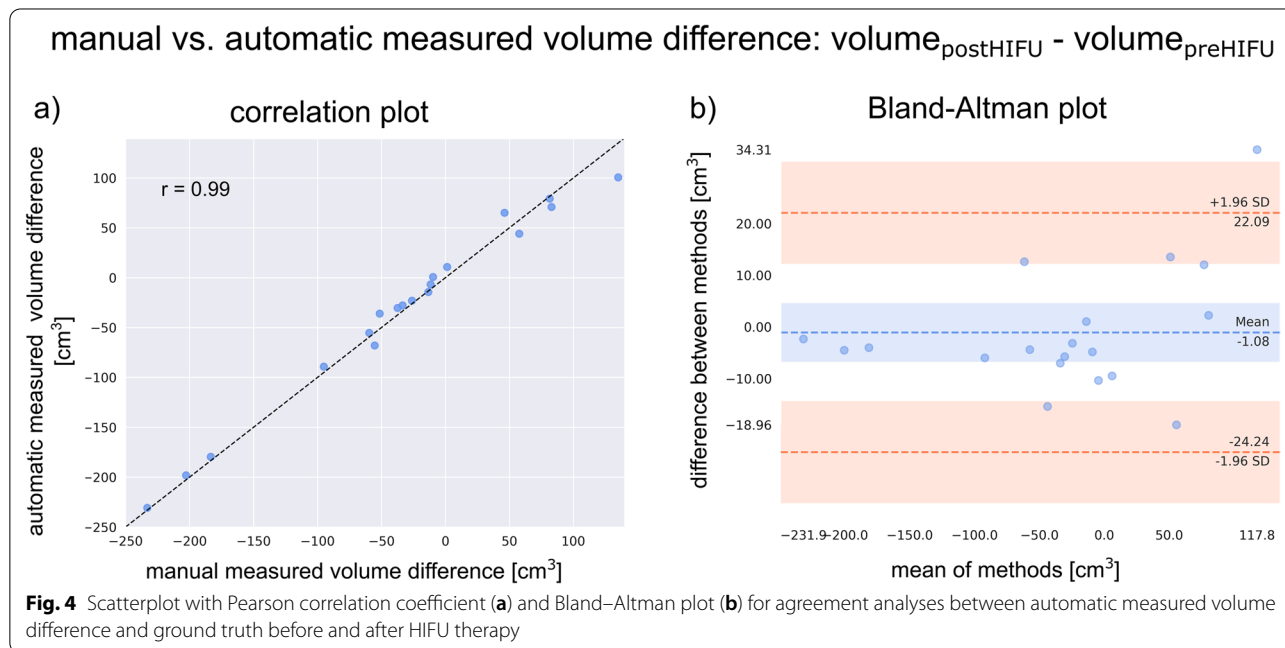
The table provides the mean Dice scores and the mean relative volume difference

areas where partial volume effects make exact delineation of uterine tissue difficult.

In routine clinical practice, the volume of uterus and uterine fibroids has been assessed by magnetic resonance imaging and is currently measured either from the diameters in the anterior–posterior, cranio-caudal and right–left directions using the volume formula of a prolate ellipsoid [7] or by manually drawing the contours on each axial slice [10]. Therefore, an accurate volumetry is very

time-consuming, so that automation of the measurement is very desirable.

Two different approaches for automatic uterine segmentation were compared in this study: A default 3D nnU-Net and a modified version with CBAMs in the encoder. The advantage of CBAMs has already been shown in various works for classification, object detection and segmentation tasks [25–28]. Although the CBAM nnU-Net architecture has shown excellent performance on the five validation sets, it could not outperform the



default architecture overall, which may be attributed to its already high segmentation performance. It may nevertheless be worthwhile to further investigate the combination of CBAM and U-Net models for different medical segmentation tasks, especially in cases where the performance of a standard U-Net is limited.

In order to compare the performance of the nnU-Net model with that of human readers, manual segmentations were generated independently by both annotators and compared to the automatic segmentation on a subset of the data. This comparison showed a similar level of agreement between the board-certified radiologist, the intensively trained resident and the automatic measurements.

Previous research has already investigated deep learning methods for uterine segmentation, where most of the presented approaches are also based on a U-Net architecture [29–31]. In one of these studies, also based on MRI, a 3D U-Net model requiring only minimal user interaction for the segmentation of the uterine cavity and the placenta of pregnant women was presented and evaluated in normal pregnant women and also in women with suspected placental abnormalities. Mean Dice scores of 92% and 88% were achieved for uterine cavity segmentation in the two groups [29].

For ultrasound images, automated segmentation approaches have also been proposed using a modified 2D U-Net architecture for segmentation of the uterus [30]. Patients with uterine fibroids were not specifically considered in that study and although several models were trained at different 2D planes, overall only low Dice

scores were reported. The authors attribute this to issues with slices near the uterine edge, demonstrating the limitation of 2D approaches. To a certain extent, the lower segmentation performance may also be explained by the quality of the available image data, which essentially depends on the sonographic experience of the examiner.

A further approach, also based on MRI, uses a DenseU-Net for segmentation of the uterus on sagittal slices [31]. That work proposed the sharpening of uterine edges in a preprocessing step, which was added as additional input to the network, leading to a mean Dice score of 87.6%. This 2D approach was also not developed in patients with uterine fibroids.

In contrast, Zhang et al. have presented the HIFU-Net, an encoder–decoder network with a pre-trained ResNet101 [36] encoder for segmentation of the uterus, uterine fibroids and the spine on 2D sagittal MRI slices. This study included only preoperative patients and was employed for HIFU therapy planning. Precise segmentation was reported to be difficult at the margins for patients with many fibroids, resulting in a Dice score for uterine segmentation of 82.37%. The authors suggest that direct 3D segmentation may lead to higher segmentation performance [32]. Because post-HIFU image data were not included in this study, information on the applicability of this method for accurate assessment of treatment-related volume changes over time after HIFU ablation is missing. However, a clear advantage of the HIFUNet is the simultaneous segmentation of individual uterine fibroids, which has not been addressed in the current study so far. Direct region detection of uterine fibroids after MRI-fused

ultrasound using a combination of split-and-merge and multi-seed region growing methods was also considered in another work [37]. These studies allow direct segmentation of single fibroids, thus enabling immediate assessment of the treatment response of individual fibroids.

It should be noted that the precise contours, especially of small fibroids, are often difficult to delineate from adjacent uterine tissue, even for human readers. This may also contribute to lower accuracy of segmentation of individual fibroids compared to uterine segmentation [32, 37]. From a clinical perspective, segmentation of the entire uterus may already be a relevant measure for therapy response assessment. For example, when multiple fibroids are treated, the improvement in myoma-associated symptoms is probably primarily related to the reduction in overall uterine volume. Thus, in the post-interventional course, the uterus may lie differently in the pelvic region due to the reduction in its total size, i.e., no longer pressing on the urinary bladder or the Fallopian tubes. The segmentation of the entire uterus provided by our approach is a fast, accurate, and reproducible method that can be applied on MRI image data also in the post-interventional course, even without knowledge of the number and exact location of the treated fibroids. In many cases, the latter is only known to the therapist. Nevertheless, the presented method may also serve as an input for targeted segmentation of single fibroids for assessment of treatment response.

Our study has several limitations. First, part of the labeled dataset was generated semi-manually, in which a network trained on a subset of the data was used for deep learning-assisted annotation. However, all data used as ground truth were finally validated by a board-certified radiologist. In addition, the deep learning model was specifically trained for axial T2-weighted turbo spin echo sequences with explicit specifications regarding the spatial resolution. The study was performed in only one center from a single MRI scanner. Therefore, the generalizability should be further evaluated in a multicenter setting. The use of the algorithm will be enabled for collaborative studies on reasonable request.

## Conclusion

This study provides a method for automatic segmentation of the uterus from patients with and without uterine fibroids with a performance similar to human readers, enabling fast, easy and reproducible assessment of volume changes in the clinical setting of a HIFU therapy.

## Abbreviations

CBAM: Convolutional block attention module; HIFU: High-intensity focused ultrasound; ReLU: Rectified linear unit.

## Supplementary information

The online version contains supplementary material available at <https://doi.org/10.1186/s13244-022-01342-0>.

**Additional file 1. S1.** Details on method development. **S2.** Comparison of both architectures.

## Author contributions

MT was responsible for the method development and was a major contributor in writing the manuscript. During method development and evaluation, she was intensively advised and supported by SN and AMS. TT and OS were responsible for the segmentation of all uteri, and TT was the main contributor in writing the medical section of the manuscript. WB assisted with data curation. MM was an expert in the field of HIFU therapy and helped with her medical expertise in developing the method and writing the manuscript. FR, ME, AM and UA provided additional support to the project with their medical expertise. Revision of the manuscript was mainly performed by AMS and MM. Concept of the study: AMS, MM, MT. All authors read and approved the final manuscript.

## Funding

Open Access funding enabled and organized by Projekt DEAL.

## Availability of data and materials

The datasets analyzed during the current study are not publicly available due to data protection laws.

## Declarations

### Ethics approval and consent to participate

Institutional Review Board approval was obtained by the local Ethics Committees at the Medical Faculty of the Rheinische Friedrich-Wilhelms-Universität Bonn (no. 295/19).

### Consent for publication

Written informed consent was waived by the Institutional Review Board (University of Bonn).

### Competing interests

The authors declare that they have no competing interests.

### Author details

<sup>1</sup>Department of Diagnostic and Interventional Radiology, University Hospital Bonn, Venusberg-Campus 1, 53127 Bonn, Germany. <sup>2</sup>Department of Radiotherapy and Radiation Oncology, University Hospital Bonn, Venusberg-Campus 1, 53127 Bonn, Germany. <sup>3</sup>Department of Neuroradiology, University Hospital Bonn, Venusberg-Campus 1, 53127 Bonn, Germany. <sup>4</sup>Department of Gynaecology and Gynaecological Oncology, University Hospital Bonn, Bonn, Germany. <sup>5</sup>Department of Nuclear Medicine, University Hospital Bonn, Bonn, Germany.

Received: 2 September 2022 Accepted: 2 December 2022

Published online: 05 January 2023

## References

1. Boosz AS, Reimer P, Matzko M, Römer T, Müller A (2014) The conservative and interventional treatment of fibroids. *Dtsch Arztebl Int* 111:877–183
2. Stewart EA, Cookson CL, Gandolfo RA, Schulze-Rath R (2017) Epidemiology of uterine fibroids: a systematic review. *BJOG* 124:1501–1512
3. Al-Hendy A, Myers ER, Stewart E (2017) Uterine fibroids: burden and unmet medical need. *Semin Reprod Med* 35:473–480
4. Donnez J, Dolmans MM (2016) Uterine fibroid management: from the present to the future. *Hum Reprod Update* 22:665–686
5. Mas A, Tarazona M, Dasí Carrasco J, Estaca G, Cristóbal I, Monleón J (2017) Updated approaches for management of uterine fibroids. *Int J Womens Health* 9:607–617



6. Gurusamy KS, Vaughan J, Fraser IS, Best LMJ, Richards T (2016) Medical therapies for uterine fibroids – a systematic review and network meta-analysis of randomised controlled trials. *PLoS One* 11:e0149631
7. Marinova M, Ghaei S, Recker F et al (2021) Efficacy of ultrasound-guided high-intensity focused ultrasound (USgHIFU) for uterine fibroids: an observational single-center study. *Int J Hyperthermia* 38:30–38
8. Recker F, Thudium M, Strunk H et al (2021) Multidisciplinary management to optimize outcome of ultrasound-guided high-intensity focused ultrasound (HIFU) in patients with uterine fibroids. *Sci Rep* 11:22768
9. Tonguc T, Strunk H, Gonzalez-Carmona MA et al (2021) US-guided high-intensity focused ultrasound (HIFU) of abdominal tumors: outcome, early ablation-related laboratory changes and inflammatory reaction. a single-center experience from Germany. *Int J Hyperthermia* 38:65–74
10. Kim HS, Baik JH, Pham LD, Jacobs MA (2011) MR-guided high-intensity focused ultrasound treatment for symptomatic uterine leiomyomata: long-term outcomes. *Acad Radiol* 18:970–976
11. Hindley J, Gedroyc WM, Regan L et al (2004) MRI guidance of focused ultrasound therapy of uterine fibroids: early results. *AJR Am J Roentgenol* 183:1713–1719
12. Wu F, Wang Z-B, Chen W-Z et al (2004) Extracorporeal high intensity focused ultrasound ablation in the treatment of 1038 patients with solid carcinomas in China: an overview. *Ultrason Sonochem* 11:149–154
13. Hahn M, Fugunt R, Schoenfisch B et al (2018) High intensity focused ultrasound (HIFU) for the treatment of symptomatic breast fibroadenoma. *Int J Hyperthermia* 35:463–470
14. Zhang R, Chen J-Y, Zhang L et al (2021) The safety and ablation efficacy of ultrasound-guided high-intensity focused ultrasound ablation for desmoid tumors. *Int J Hyperthermia* 38:89–95
15. Marinova M, Huxold HC, Henseler J et al (2019) Clinical effectiveness and potential survival benefit of US-guided high-intensity focused ultrasound therapy in patients with advanced-stage pancreatic cancer. *Ultraschall Med* 40:625–637
16. Marinova M, Wilhelm-Buchstab T, Strunk H (2019) Advanced pancreatic cancer: high-intensity focused ultrasound (HIFU) and other local ablative therapies. *Rofo* 191:216–227
17. Coppola F, Faggioni L, Gabelloni M et al (2021) Human, all too human? An all-around appraisal of the “artificial intelligence revolution” in medical imaging. *Front Psychol* 12:710982
18. Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts HJWL (2018) Artificial intelligence in radiology. *Nat Rev Cancer* 18:500–510
19. Yu K-H, Beam AL, Kohane IS (2018) Artificial intelligence in healthcare. *Nat Biomed Eng* 2:719–731
20. Nowak S, Mesropyan N, Faron A et al (2021) Detection of liver cirrhosis in standard T2-weighted MRI using deep transfer learning. *Eur Radiol* 31:8807–8815
21. Luetkens JA, Nowak S, Mesropyan N et al (2022) Deep learning supports the differentiation of alcoholic and other-than-alcoholic cirrhosis based on MRI. *Sci Rep* 12:8297
22. Ronneberger O, Fischer P, Brox T (2015) U-Net: convolutional networks for biomedical image segmentation. *Proc MICCAI* 2015:234–241
23. Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH (2021) nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat Methods* 18:203–211
24. Nowak S, Theis M, Wichtmann BD et al (2021) End-to-end automated body composition analyses with integrated quality control for opportunistic assessment of sarcopenia in CT. *Eur Radiol*. <https://doi.org/10.1007/s00330-021-08313-x>
25. Woo S, Park J, Lee J-Y, Kweon IS (2018) CBAM: convolutional block attention module. *Proc ECCV* 2018:3–19
26. Yang J, Xie M, Hu C et al (2021) Deep learning for detecting cerebral aneurysms with CT angiography. *Radiology* 298:155–163
27. Chen M, Zhao C, Tian X et al (2021) Placental super micro-vessels segmentation based on resnext with convolutional block attention and U-Net. *Proc IEEE EMBC* 2021:4015–4018
28. Trebing K, Stańczyk T, Mehrkanoon S (2021) SmaAt-UNet: Precipitation nowcasting using a small attention-UNet architecture. *Pattern Recognit Lett* 145:178–186
29. Shahedi M, Spong CY, Dormer JD et al (2021) Deep learning-based segmentation of the placenta and uterus on MR images. *J Med Imaging* 8:054001
30. Behboodi B, Rivaz H, Lalondrelle S, Harris E (2021) Automatic 3D ultrasound segmentation of uterus using deep learning. *Proc IEEE IUS* 2021:1–4
31. Niu Y, Zhang Y, Ying L et al (2021) Uterine magnetic resonance image segmentation based on deep learning. *J Phys Conf Ser* 1861:012067
32. Zhang C, Shu H, Yang G et al (2020) HIFUNet: multi-class segmentation of uterine regions from MR images using global convolutional networks for HIFU surgery planning. *IEEE Trans Med Imaging* 39:3309–3320
33. Fedorov A, Beichel R, Kalpathy-Cramer J et al (2012) 3D slicer as an image computing platform for the quantitative imaging network. *Magn Reson Imaging* 30:1323–1341
34. Waskom ML (2021) seaborn: statistical data visualization. *J Open Source Softw* 6:3021
35. jaketmp, & Lee Tirrell. (2021). jaketmp/pyCompare: (v1.5.2). Zenodo. <https://doi.org/10.5281/zenodo.4926654>
36. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. *Proc IEEE CVPR* 2016:770–778
37. Rundo L, Militello C, Vitabile S et al (2016) Combining split-and-merge and multi-seed region growing algorithms for uterine fibroid segmentation in MRgFUS treatments. *Med Biol Eng Comput* 54:1071–1084

### Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Submit your manuscript to a SpringerOpen<sup>®</sup> journal and benefit from:**

- Convenient online submission
- Rigorous peer review
- Open access: articles freely available online
- High visibility within the field
- Retaining the copyright to your article

---

Submit your next manuscript at ► [springeropen.com](https://www.springeropen.com)

3.2. Publication 2: End-to-end automated body composition analyses with integrated quality control for opportunistic assessment of sarcopenia in CT

Nowak S, Theis M, Wichtmann BD, Faron A, Froelich MF, Tollens F, Geißler HL, Block W, Luetkens JA, Attenberger UI, Sprinkart AM. **End-to-end automated body composition analyses with integrated quality control for opportunistic assessment of sarcopenia in CT.** Eur Radiol 2021; 32(5): 3142-3151



# End-to-end automated body composition analyses with integrated quality control for opportunistic assessment of sarcopenia in CT

Sebastian Nowak<sup>1</sup> · Maike Theis<sup>1</sup> · Barbara D. Wichtmann<sup>1</sup> · Anton Faron<sup>1</sup> · Matthias F. Froelich<sup>2</sup> · Fabian Tollens<sup>2</sup> · Helena L. Geißler<sup>1</sup> · Wolfgang Block<sup>1,3,4</sup> · Julian A. Luetkens<sup>1</sup> · Ulrike I. Attenberger<sup>1</sup> · Alois M. Sprinkart<sup>1</sup>

Received: 15 April 2021 / Revised: 6 August 2021 / Accepted: 31 August 2021 / Published online: 30 September 2021  
© The Author(s) 2021, corrected publication 2022

## Abstract

**Objectives** To develop a pipeline for automated body composition analysis and skeletal muscle assessment with integrated quality control for large-scale application in opportunistic imaging.

**Methods** First, a convolutional neural network for extraction of a single slice at the L3/L4 lumbar level was developed on CT scans of 240 patients applying the nnU-Net framework. Second, a 2D competitive dense fully convolutional U-Net for segmentation of visceral and subcutaneous adipose tissue (VAT, SAT), skeletal muscle (SM), and subsequent determination of fatty muscle fraction (FMF) was developed on single CT slices of 1143 patients. For both steps, automated quality control was integrated by a logistic regression model classifying the presence of L3/L4 and a linear regression model predicting the segmentation quality in terms of Dice score. To evaluate the performance of the entire pipeline end-to-end, body composition metrics, and FMF were compared to manual analyses including 364 patients from two centers.

**Results** Excellent results were observed for slice extraction ( $z$ -deviation =  $2.46 \pm 6.20$  mm) and segmentation (Dice score for SM =  $0.95 \pm 0.04$ , VAT =  $0.98 \pm 0.02$ , SAT =  $0.97 \pm 0.04$ ) on the dual-center test set excluding cases with artifacts due to metallic implants. No data were excluded for end-to-end performance analyses. With a restrictive setting of the integrated segmentation quality control, 39 of 364 patients were excluded containing 8 cases with metallic implants. This setting ensured a high agreement between manual and fully automated analyses with mean relative area deviations of  $\Delta$ SM =  $3.3 \pm 4.1\%$ ,  $\Delta$ VAT =  $3.0 \pm 4.7\%$ ,  $\Delta$ SAT =  $2.7 \pm 4.3\%$ , and  $\Delta$ FMF =  $4.3 \pm 4.4\%$ .

**Conclusions** This study presents an end-to-end automated deep learning pipeline for large-scale opportunistic assessment of body composition metrics and sarcopenia biomarkers in clinical routine.

## Key Points

- *Body composition metrics and skeletal muscle quality can be opportunistically determined from routine abdominal CT scans.*
- *A pipeline consisting of two convolutional neural networks allows an end-to-end automated analysis.*
- *Machine-learning-based quality control ensures high agreement between manual and automatic analysis.*

**Keywords** Body composition · Tomography, X-ray computed · Deep learning · Quality control · Sarcopenia

## Abbreviations

CDFNet	Competitive dense fully connected network
CNN	Convolutional neural network
FMF	Fatty muscle fraction
SAT	Subcutaneous adipose tissue

SM	Skeletal muscle
VAT	Visceral adipose tissue

## Introduction

Body composition analyses aim to determine the quantity of connective tissue compartments. In addition to quantifying the amount of adipose and muscle tissue, recent work proposed methods to obtain additional information about a patient's general condition by also determining the quality of skeletal muscle tissue in terms of fatty degeneration. Several studies demonstrated that these metrics

Sebastian Nowak and Maike Theis contributed equally to this study.

✉ Alois M. Sprinkart  
sprinkart@uni-bonn.de

Extended author information available on the last page of the article

determined from abdominal imaging provide prognostic implications in oncologic or cardiovascular diseases [1–8].

The amount of visceral and subcutaneous adipose tissue, as well as the amount and quality of muscle tissue, can be reliably determined from abdominal CT imaging. An opportunistic large-scale assessment in clinical routine has the potential to further enhance the understanding of the clinical value of body composition analyses in various diseases, e.g., for therapy decision and/or outcome prediction. Also, the establishment of gender-, age-, and ethnicity-specific norm values is only feasible through the widespread application of these analyses.

However, the determination of fat and muscle volume by manually annotating the region of interest by a radiologist is rather time-consuming, which currently prevents clinical routine application. Several studies have shown that area measurements of connective tissue compartments on a single slice at a certain lumbar level are highly correlated with total volume in the abdomen [9–11]. This led to greatly reduced annotation times for manual body composition analysis when applying a 2D— instead of a 3D approach. In recent years, several methods have been proposed for automating the required tissue segmentation step. It was a logical consequence that with the dominant rise of deep learning for image segmentation the previously manually segmented images were used to develop methods for automated segmentation by supervised learning [12–14]. However, manual interaction was still required for extraction of the single slice on which the automatic segmentation is performed. Only very recent work also includes deep-learning-based automated slice extraction as the next step for truly automated body composition analyses [15–17].

Moreover, to the best of our knowledge, there is currently no work that presents integrated quality control for both slice extraction and tissue segmentation. This still leaves one factor that represents an additional human effort in opportunistic analysis, namely identifying cases where the algorithm fails. Automatic determination of the predictive uncertainties can help identify cases with low-quality analyses and can additionally be used to monitor the performance of an autonomous system during deployment, as suggested for machine learning operations to manage deep learning life cycles. This can also help to detect changes in the data and to raise a warning in case of domain shifts.

Hence, the aim of this study was to develop an automated body composition analysis for abdominal CT with integrated quality checks and to evaluate the end-to-end performance of the proposed pipeline on dual-center test data.

## Material and methods

### Overview

Figure 1 shows an overview of the developed pipeline. In the first part, a single slice at the L3/L4 lumbar level is extracted

from a 3D CT scan. In the second part, the extracted 2D image is segmented into three compartment classes: visceral and subcutaneous adipose tissue (VAT, SAT) and skeletal muscle (SM). The fatty muscle fraction (FMF), a quantitative marker for fatty muscle degeneration, is determined in a subsequent post-processing step [1, 6]. For both deep-learning-based slice extraction and segmentation, classical machine learning methods were employed for integration of quality control steps that capture the predictive uncertainty during deployment.

Slice extraction and tissue segmentation were developed independently. To evaluate the end-to-end performance of the entire pipeline, automatically extracted body composition metrics and FMF were compared with manual analyses on an unselected dual-center test set. Figure 2 provides an overview of the data sets used for method development and evaluation.

### Method development for slice extraction

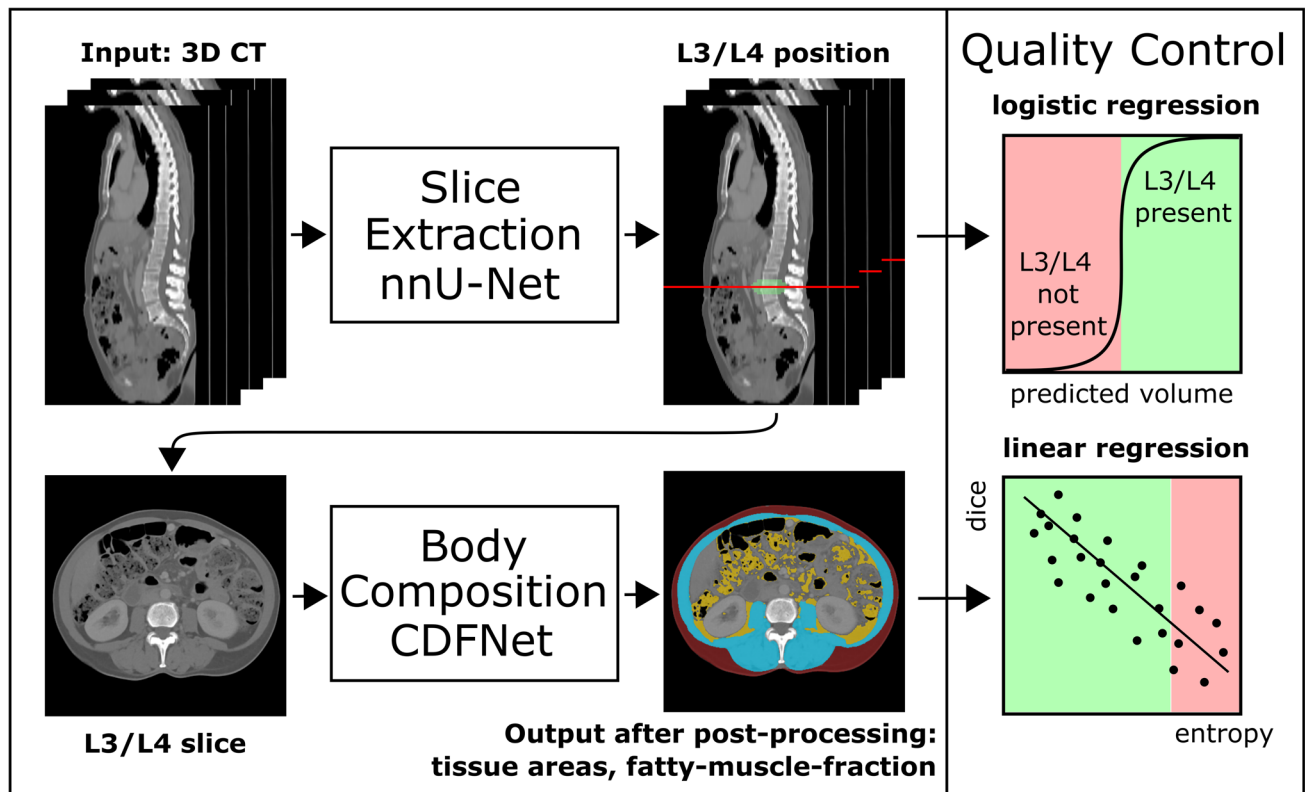
#### Dataset

With institutional review board approval, written informed patient consent was waived because of the retrospective nature of all parts of the study. Retrospectively derived 3D CT scans of 240 patients (94 female, mean age  $65 \pm 14$  years) referred for diagnostic CT including imaging of the upper abdomen acquired at eight different CT scanners were used for development of the slice extraction method. Of these patients, 43 received CT before undergoing transcatheter aortic valve implantation, 91 before transjugular intrahepatic portosystemic shunt intervention, and 106 patients received CT in the setting of immunotherapy for malignant melanoma.

The ground truth was generated by a board-certified radiologist (A.F.) by manually defining the center of the L3/L4 vertebral disk with an in-house tool (Matlab, Mathworks). Data were randomly split into a training set ( $n = 192$ , 80%) and a hold-out test ( $n = 48$ , 20%) set. The method was additionally tested on dual-center test data (described below).

#### Model

The extraction of a single slice at L3/L4 lumbar level was formulated as a segmentation task. A 3D U-Net architecture was trained using the nnU-Net framework, which has achieved high-performance values for various medical segmentation tasks and has the advantage of automatically adapting to different input sizes [18]. This is a relevant feature for the slice extraction task since the input are CT scans with a wide variety of scan lengths. The label map for



**Fig. 1** Schematic representation of the presented pipeline for autonomous body composition analysis. Input of the pipeline is a 3D CT scan. In the first part, a 3D convolutional neural network (CNN) was employed for slice extraction using nnU-Net. In the second part, a competitive dense fully connected CNN (CDFNet) is applied for segmentation of the body compartments. Classical machine learning

methods were employed for integration of quality control steps. For the slice extraction part, a logistic regression model was developed that classifies the presence of L3/L4 lumbar level in the 3D CT scan. For segmentation of the different tissues, a linear regression model was established that predicts segmentation quality in terms of the Dice score

training of the network was generated by applying a Gaussian distribution to the coordinates of the L3/L4 vertebral disk and binarizing the resulting probability map by a threshold [19]. Further details on image pre-processing, augmentation, and experimental design can be found in Supplement S1. For training, fivefold cross-validation was used and testing was performed with an ensemble of the cross-validated models.

### Quality control

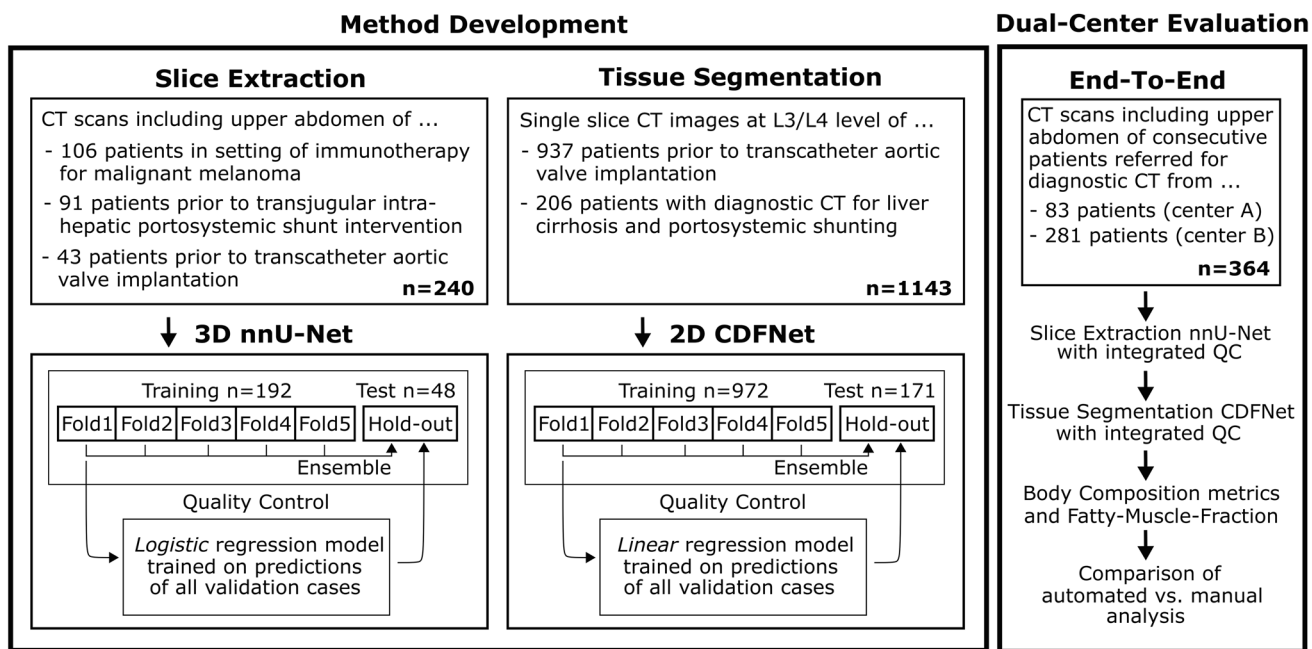
After training of the slice extraction method, a logistic regression model was built to automatically identify 3D CT scans that do not include the L3/L4 lumbar level. To obtain a balanced distribution of images with and without the L3/L4 lumbar level, for each 3D CT scan of the training, hold-out and dual-center test set, a cropped version was created. The logistic regression model was trained based on the predicted volume of all validation cases of the cross-validated slice extraction nnU-Net and applied to all test sets. Additional information about cropping and feature selection can be found in Supplement S2.

### Method development for tissue segmentation

#### Dataset

For the development of the tissue segmentation method (VAT, SAT, SM), retrospectively derived single slice images at the L3/L4 lumbar level from 1143 patients (559 female, mean age  $77 \pm 11$  years) were used. 937 patients underwent pre-interventional CT for transcatheter aortic valve implantation and 206 patients underwent diagnostic CT for liver cirrhosis with portosystemic shunting. The dataset intentionally included a high number of patients with anasarca (19.2%), ascites (9.4%), or both anasarca and ascites (6.5%). The ground truth of the segmentation was defined by manual drawing and was also used to train a different CNN in a previous work, where additional details on the dataset are reported [13].

The data for method development were randomly split into a training set ( $n=972$ , 85%) and hold-out test ( $n=171$ , 15%) set. The method was additionally tested on dual-center test data (described below).



**Fig. 2** Overview of the data sets used for method development and evaluation. The nnU-Net employed for extraction of a single slice at L3/L4 level from a 3D CT scan and the CDFNet for tissue segmentation of the 2D CT slices were developed on two different datasets. Both methods were fivefold cross-validated and an ensemble of the cross-validated models was tested on the hold-out data. The regres-

sion models for integrated quality control (QC) were developed on the validation data of the cross-validated models and were also tested on the hold-out data. Finally, the entire pipeline of slice extraction, tissue segmentation, and quality control was evaluated end-to-end on the dual-center test data and compared against manual analyses

## Model

A 2D competitive dense fully convolutional network (CDFNet), which has shown promising results for body composition analysis in magnetic resonance imaging, was used for tissue segmentation [20]. This architecture is proposed as an extension of the Dense-UNet architecture by max-out activation units. In a CDFNet, feature maps are generated by element-wise selection of the maximum values of previous feature maps, which has been shown to have a positive effect on performance and generalizability compared to unselective concatenation [20–22]. Further details on image pre-processing, augmentation, experimental design and computation of the fatty muscle fraction are provided in Supplement S3.

For training, fivefold cross-validation was used and testing was performed with an ensemble of the cross-validated models.

## Quality control

To assess the predictive uncertainty of the segmentation during employment, a linear regression model was developed that predicts the segmentation Dice score for the muscle

class based on the average entropies of the probability maps. This metric is proposed by a recent work as a feature to estimate quality of medical image segmentation and to detect out-of-distribution samples and ambiguous cases [23]. Although this method could be applied to all tissue classes, we focused on the muscle class because we consider it the most important class for the assessment of sarcopenia.

The linear regression model was trained with the predictions of all validation cases of the cross-validated tissue segmentation CDFNet and tested on all test sets.

## Dual-center test data and end-to-end evaluation

The entire pipeline was finally evaluated end-to-end, i.e., from 3D CT scan to extracted body composition metrics. The automatically determined tissue areas and the fatty muscle fraction were compared with the manually determined values. For this purpose, 3D CT scans of consecutive patients referred for diagnostic CT including imaging of the upper abdomen were retrospectively retrieved from two centers.

- Center A: 83 (41 females, mean age  $60 \pm 15$  years) patients were used as internal test data from the Department of Diagnostic and Interventional Radiology, Uni-

versity Hospital Bonn. Data were acquired at four different CT scanners.

- Center B: 281 (111 females, mean age  $63 \pm 16$  years) patients were used as external test data from the Department of Radiology and Nuclear Medicine, University Medical Centre Mannheim. Data were acquired at three different CT scanners.

In this data set, 10 patients had metallic implants. However, in the end-to-end evaluation, these cases were intentionally not excluded. For demonstration of the tissue segmentation quality control, a restrictive setting was applied excluding 10% of the cases with lowest predicted Dice score of the muscle class. End-to-end performance is reported for both included and excluded cases.

The ground truth for slice extraction and tissue segmentation was labeled by a radiology resident (B.W.) and a board-certified radiologist (A.F.). All labels of the radiology resident were validated by the board-certified radiologist.

Additional information on dual-center test data can be found in Supplement S5.

### Results

A summary of the results can be found in Fig. 3.

### Slice extraction

The mean deviations between the predictions of the ensemble of cross-validated slice extraction models and the manually defined ground truth were  $\Delta z = 2.27 \pm 7.08$  mm for the hold-out test data and  $\Delta z = 2.46 \pm 6.20$  mm for the dual-center test data. Considering an acceptable deviation of up to 10 mm, 96% of the extracted slices of the hold-out test set and 96% of the dual-center test data were extracted at the correct level. The mean deviations are listed separately for all test sets in Table 1.

### Tissue segmentation

The ensemble of fivefold cross-validated CDFNet models achieved excellent Dice scores, both on the hold-out test data (SM:  $0.96 \pm 0.02$ , VAT:  $0.98 \pm 0.02$ , SAT:  $0.98 \pm 0.01$ ) and on the dual-center test data (SM:  $0.95 \pm 0.04$ , VAT:  $0.98 \pm 0.02$ , SAT:  $0.97 \pm 0.04$ ). Table 2 lists the Dice scores separately for each test set.

### Quality control

Figure 4a shows the logistic regression model developed for identifying 3D CT scans that do not contain the L3/L4 level. High accuracy was observed for predicting the presence of

## Separate Evaluation of Slice Extraction and Tissue Segmentation

Slice Extraction			QC	Tissue Segmentation			QC	
Results	Mean $\Delta z$ [mm]	Accuracy $\Delta z \leq 10$ mm	Accuracy L3/L4 present	Results	Dice SM	Dice VAT	Dice SAT	$\Delta$ Dice SM predicted
Hold-out test data	$2.3 \pm 7.1$	0.96	1.00	Hold-out test data	0.96	0.98	0.98	0.016
Dual-center test data	$2.5 \pm 6.2$	0.96	0.98	Dual-center test data	0.95	0.98	0.97	0.016

## End-To-End Evaluation

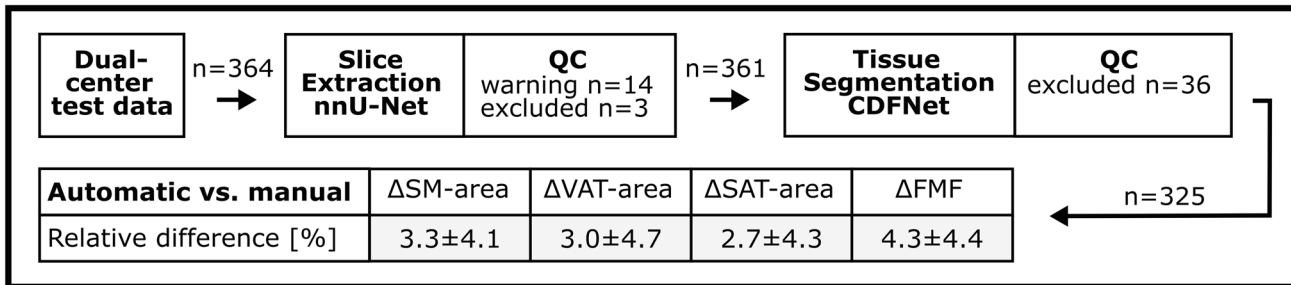


Fig. 3 Summary of results: separate analyses of slice extraction, tissue segmentation, and respective quality control (QC), as well as agreement between end-to-end automated and manual area measure-

ments of skeletal muscle (SM), visceral adipose tissue (VAT), subcutaneous adipose tissue (SAT), and the fatty muscle fraction (FMF)

**Table 1** Mean  $z$ -deviation ( $\Delta z$ ) and slice extraction accuracy for different tolerance margins obtained with the cross-validated nnU-Net ensemble for the hold-out test set and for the additional test data from center A and center B

Slice extraction	Mean, $\Delta z$ [mm]	Accuracy, $\Delta z=0$ mm	Accuracy, $\Delta z < 5$ mm	Accuracy, $\Delta z < = 10$ mm
Hold-out	$2.27 \pm 7.08$	0.79	0.96	0.96
Center A	$3.35 \pm 4.10$	0.51	0.88	0.99
Center B	$2.19 \pm 6.70$	0.85	0.96	0.96

**Table 2** Dice scores for segmentation of skeletal muscle (SM), visceral adipose tissue (VAT), and subcutaneous adipose tissue (SAT) obtained with the cross-validated CDFNet ensemble for the hold-out test set and for the additional test data from center A and center B

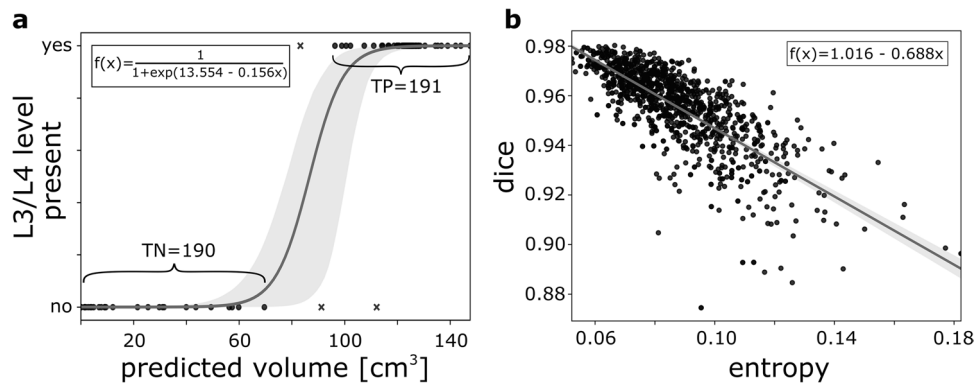
Tissue segmen- tation	Dice score, SM	Dice score, VAT	Dice score, SAT
Hold-out	$0.958 \pm 0.023$	$0.981 \pm 0.015$	$0.982 \pm 0.012$
Center A	$0.959 \pm 0.021$	$0.981 \pm 0.012$	$0.979 \pm 0.038$
Center B	$0.944 \pm 0.039$	$0.974 \pm 0.027$	$0.969 \pm 0.037$

the L3/L4 level in the original and cropped versions of the hold-out test data (100%) and also on the dual-center test data (center A: 99%, center B: 98%). Sensitivity and specificity were 97% and 99% for the dual-center test data.

The linear regression model developed for integrated quality control of the tissue segmentation is shown in Fig. 4b. Mean differences between the observed and the predicted Dice score for the hold-out test data were  $0.016 \pm 0.016$  (SM),  $0.005 \pm 0.005$  (VAT), and  $0.008 \pm 0.010$  (SAT) and for the dual-center  $0.016 \pm 0.016$  (SM),  $0.007 \pm 0.012$  (VAT), and  $0.010 \pm 0.015$  (SAT).

### End-to-end evaluation

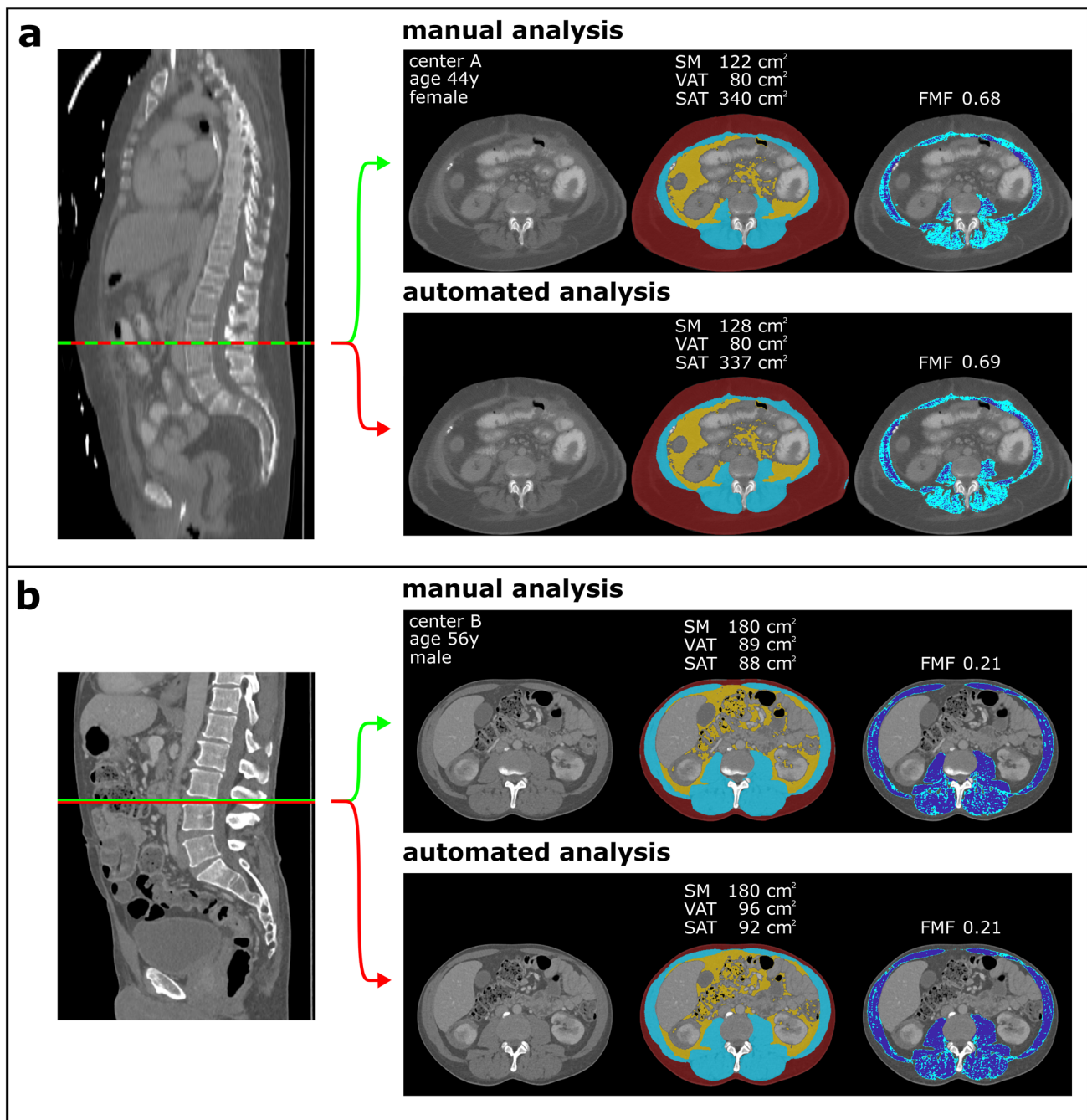
Figure 5 shows examples of the end-to-end analyses. Application of the logistic regression model to the dual-center test data, all of which contained the L3/L4 lumbar level, resulted in 14 of 364 3D CT scans with a warning that the scan may not contain the L3/L4 level. In three of these cases, the patients had implants at the L3/L4 level. For the remaining 11 cases, the difference between predicted L3/L4 level and ground truth was  $\Delta z = 6.38 \pm 10.77$  mm. Except for the three patients with implants, none of the patients were excluded from further analyses. Subsequently, the linear regression model for integrated quality control of the tissue segmentation was applied. With a restrictive setting, 36 of 361 cases were flagged as possibly having limited segmentation quality with predicted Dice scores of the muscle class ranging from 0.861 to 0.924. In 5 of these 36 cases, the patients had implants at the L3/L4 level, and 4 patients had a pronounced hernia. In the remaining cases, there were various reasons for limited segmentation quality, such as parts of the arms included in the tissue segmentation or parts of the kidney classified as muscle. In total, 8 of 10 cases with metallic implants on the L3/L4 level were excluded by the two quality control steps. For the two cases not excluded by quality



**Fig. 4** Models trained for quality control: **a** Based on the predicted volume of the nnU-Net employed for slice extraction, a logistic regression model was trained to predict the presence of the slice at L3/L4 lumbar level in the 3D CT scan. **b** For prediction of the tissue segmentation quality in terms of the Dice score, a linear regression

model was trained based on the entropy of the probability map of the CDFNet for the muscle class. Both regression models were built on features derived from cross-validation data of slice extraction and tissue segmentation, respectively. Gray areas represent the 95% confidence intervals





**Fig. 5** Compartmental areas of visceral adipose tissue, subcutaneous adipose tissue (VAT, SAT), skeletal muscle (SM), and fatty muscle fraction (FMF) derived for patients from center A (**a**) and center B

(**b**). Manual analysis is marked in green, while results from the proposed pipeline are marked with a red line

control, only minor hardening artifacts were observed, as shown in Supplement 4S.

Results of the entire end-to-end evaluation are summarized in Table 3. A high agreement was observed for the 325 cases of the dual-center data that passed the quality control. Body composition metrics and FMF derived from automated and manual analysis showed

absolute differences in area of  $\Delta\text{SM} = 5.0 \pm 6.0 \text{ cm}^2$ ,  $\Delta\text{VAT} = 3.7 \pm 5.8 \text{ cm}^2$ , and  $\Delta\text{SAT} = 5.7 \pm 10.4 \text{ cm}^2$ , corresponding to low relative differences of  $\Delta\text{SM} = 3.3 \pm 4.1\%$ ,  $\Delta\text{VAT} = 3.0 \pm 4.7\%$ , and  $\Delta\text{SAT} = 2.7 \pm 4.3\%$ . Also for FMF, low absolute deviations of  $\Delta\text{FMF} = 0.014 \pm 0.012$  and relative deviations of  $\Delta\text{FMF} = 4.3 \pm 4.4\%$  were observed.

**Table 3** Evaluation of the end-to-end performance of the body composition analyses

Center	Quality control	Fatty muscle fraction	Muscle area (cm <sup>2</sup> )	Visceral fat area (cm <sup>2</sup> )	Subcutaneous fat area (cm <sup>2</sup> )
A	Passed, <i>n</i> = 82	0.009 ± 0.008 (3.1% ± 3.5%)	3.7 ± 4.1 (2.7% ± 4.4%)	3.6 ± 4.3 (2.7% ± 3.6%)	5.4 ± 5.3 (2.7% ± 3.0%)
B	Passed, <i>n</i> = 243	0.016 ± 0.013 (4.8% ± 4.6%)	5.4 ± 6.4 (3.5% ± 4.0%)	3.8 ± 6.2 (3.1% ± 5.0%)	5.8 ± 11.7 (2.8% ± 4.6%)
A	Excluded, <i>n</i> = 1	<b>0.046 (9.3%)</b>	<b>16.0 (16.6%)</b>	2.0 (2.3%)	14.9 (10.8%)
B	Excluded, <i>n</i> = 35	<b>0.033 ± 0.036 (6.1% ± 6.6%)</b>	<b>18.6 ± 21.6 (14.1% ± 15.6%)</b>	7.2 ± 10.4 (7.0% ± 8.6%)	18.4 ± 29.5 (7.8% ± 9.5%)

Absolute and relative differences (in parentheses) between the values obtained with the proposed pipeline and the manually determined values are listed separately for center A and center B and for all 3D CT scans that were included and excluded by restrictive setting of the tissue segmentation quality control. The excluded cases show markedly lower agreement of muscle area, while FMF agreement is still reasonably good (marked in bold)

## Discussion

This paper presents a method that allows the application of body composition analysis without human interaction, thus permitting opportunistic determination of body compartment areas and FMF as a marker for sarcopenia in routine clinical practice. For both CNNs applied in the pipeline, the trained networks are available on reasonable request (<https://qilab.de>).

In recent years, a variety of deep learning methods have been presented that address the topic of automated body composition analysis. Most of these studies focus on the segmentation of the tissue compartments in a single slice at a certain lumbar level, as it has been demonstrated that 2D and 3D measurements for quantification of VAT, SAT, and SM show a high correlation [9–14]. Although very recent works have also addressed automation of slice extraction, routine clinical application additionally requires the integration of quality control methods for both slice extraction and tissue segmentation [15, 16]. For this purpose, two classic machine learning models have been developed in this study. The developed pipeline therefore provides full automation of body composition analysis in abdominal CT, including deep-learning-based slice extraction and tissue segmentation and integrated application of quality control models.

Compared to previous research in the field of automated body composition analyses, we observed similar or superior performance values for slice extraction task and tissue segmentation in our study [12–17]. In previous work, the slice extraction task was formulated either as a regression problem, a classification task, or, similar to our approach, a segmentation problem [15–17]. While the methods proposed so far for slice extraction are based on 2D images or require the generation of a maximum intensity projection in a pre-processing step, the use of the nnU-Net framework allows the direct input of 3D CT datasets of different sizes. For tissue segmentation, different variants of a 2D U-Net architecture have been used [12, 15–17]. The CDFNet architecture applied in the current study is an extension of a DenseUNet architecture with max-out activation units, which has recently also been successfully used for body composition

analyses in magnetic resonance imaging [20]. A detailed comparison to previous work can be found in Supplement S6.

For the development of the tissue segmentation CNN, patient collectives were included that also represent tissue alterations, as ascites and anasarca, which are challenging for body composition analysis [14]. In addition, segmentation results from other studies show the disadvantages of using only threshold-based pre-processing steps to define segmentation ground truth, resulting in misclassification of intermuscular fat to one of the abdominal adipose tissue classes (VAT, SAT) [15]. To overcome this limitation, intermuscular fat was manually assigned to the muscle class in this study, allowing additional analyses of muscle [13].

Several aspects of body composition, such as skeletal muscle fat infiltration as an indicator of skeletal muscle quality were shown to provide prognostic information in patients with cardiovascular and oncologic diseases [1–3]. Thereby, FMF was recently proposed as an easy-accessible body composition metric which may be considered particularly promising as it additionally integrates information on skeletal muscle quality [1, 5]. Previous studies have demonstrated its prognostic value both as an indicator of frailty in patients with planned endovascular aortic valve replacement as well as an powerful predictor of outcome in critically ill patients receiving extracorporeal membrane oxygenation therapy [1, 6].

A recent work on 3D tissue segmentation points out that for a truly automated application of body compartment analysis, the development of quality assurance procedures is warranted to identify patients with metal artifacts [24]. The dual-center end-to-end analysis presented in the current work demonstrates that the proposed quality control ensures a high agreement between manual and automated analyses by identifying cases that are unsuitable for body composition analyses not only due to hardening artifacts but also due to other reasons limiting the segmentation quality. Interestingly, end-to-end performance analysis of cases flagged by quality control as having limited segmentation quality shows that FMF is quite robust to segmentation errors.

As a limitation of this study, only the areas of VAT, SAT, and SM are determined in a single slice instead of determining the respective tissue volumes in the entire abdomen. However, we are not aware of studies demonstrating that a 3D approach has significant advantages over the established 2D measurement for assessment of sarcopenia. Also, reference values for body compartments have so far only been determined in large studies for 2D measurements [15].

## Conclusion

This study presents an end-to-end automated deep-learning pipeline for large-scale opportunistic assessment of body composition metrics and sarcopenia biomarker in clinical routine.

**Supplementary Information** The online version contains supplementary material available at <https://doi.org/10.1007/s00330-021-08313-x>.

**Funding** Open Access funding enabled and organized by Projekt DEAL. The study was supported by a grant from the BONFOR research program of the University of Bonn (application number 2020-2A-04). The funders had no influence on conceptualization and design of the study, data analysis, and data collection, preparation of the manuscript as well as the decision to publish.

## Compliance with ethical standards

**Guarantor** The scientific guarantor of this publication is PD Dr. Alois Martin Sprinkart.

**Conflict of interest** The authors declare no competing interests.

**Statistics and biometry** No complex statistical methods were necessary for this paper.

**Informed consent** Written informed consent was waived by the institutional review board (the University of Bonn and the University of Heidelberg).

**Ethical approval** This retrospective study was approved by the institutional review board with waiver of written informed consent.

## Methodology

retrospective  
diagnostic study  
performed at two institutions

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will

need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

- Luetkens JA, Faron A, Geissler HL et al (2020) Opportunistic computed tomography imaging for the assessment of fatty muscle fraction predicts outcome in patients undergoing transcatheter aortic valve replacement. *Circulation* 141:234–236
- Faron A, Pieper CC, Schmeel FC et al (2019) Fat-free muscle area measured by magnetic resonance imaging predicts overall survival of patients undergoing radioembolization of colorectal cancer liver metastases. *Eur Radiol* 29:4709–4717
- Faron A, Sprinkart AM, Pieper CC, et al (2020) Yttrium-90 radioembolization for hepatocellular carcinoma: outcome prediction with MRI derived fat-free muscle area. *Eur J Radiol* 125:108889.
- Faron A, Sprinkart AM, Kuetting DLR et al (2020) Body composition analysis using CT and MRI: intra-individual intermodal comparison of muscle mass and myosteatosis. *Sci Rep* 10:11765
- Cruz-Jentoft AJ, Bahat G, Bauer J et al (2019) Sarcopenia: revised European consensus on definition and diagnosis. *Age Ageing* 48:16–31
- Faron A, Kreyer S, Sprinkart AM et al (2020) CT fatty muscle fraction as a new parameter for muscle quality assessment predicts outcome in venovenous extracorporeal membrane oxygenation. *Sci Rep* 10:22391
- Lenchik L, Boutin RD (2018) Sarcopenia: beyond muscle atrophy and into the new frontiers of opportunistic imaging, precision medicine, and machine learning. *Semin Musculoskelet Radiol* 22:307–322
- Prado CMM, Lieffers JR, McCargar LJ et al (2008) Prevalence and clinical implications of sarcopenic obesity in patients with solid tumours of the respiratory and gastrointestinal tracts: a population-based study. *Lancet Oncol* 9:629–635
- Shen W, Punyanitya M, Wang Z et al (2004) Total body skeletal muscle and adipose tissue volumes: estimation from a single abdominal cross-sectional image. *J Appl Physiol* 97:2333–2338
- Faron A, Luetkens JA, Schmeel FC et al (2019) Quantification of fat and skeletal muscle tissue at abdominal computed tomography: associations between single-slice measurements and total compartment volumes. *Abdom Radiol* 44:1907–1916
- Irlbeck T, Massaro JM, Bamberg F, O'Donnell CJ, Hoffmann U, Fox CS (2010) Association between single-slice measurements of visceral and abdominal subcutaneous adipose tissue with volumetric measurements: the Framingham Heart Study. *Int J Obes (Lond)* 34:781–787
- Weston AD, Korfiatis P, Kline TL et al (2018) Automated abdominal segmentation of CT scans for body composition analysis using deep learning. *Radiology* 290:669–679
- Nowak S, Faron A, Luetkens JA et al (2020) Fully automated segmentation of connective tissue compartments for CT-based body composition analysis: a deep learning approach. *Invest Radiol* 55:357–366
- Park HJ, Shin Y, Park J et al (2020) Development and validation of a deep learning system for segmentation of abdominal muscle and fat on computed tomography. *Korean J Radiol* 21:88–100
- Magudia K, Bridge CP, Bay CP et al (2020) Population-scale CT-based body composition analysis of a large outpatient population using deep learning to derive age-, sex-, and race-specific reference curves. *Radiology* 298:319–329

16. Dabiri S, Popuri K, Ma C, et al (2020) Deep learning method for localization and segmentation of abdominal CT. *Comput Med Imaging Graph* 85:101776.
17. Castiglione J, Somasundaram E, Gilligan LA, Trout AT, Brady S (2021) Automated segmentation of abdominal skeletal muscle on pediatric ct scans using deep learning. *Radiol Artif Intell* 3:e200130.
18. Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH (2021) nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat Methods* 18:203–211
19. Yang D, Xiong T, Xu D et al (2017) Deep Image-to-Image Recurrent Network with Shape Basis Learning for Automatic Vertebra Labeling in Large-Scale 3D CT Volumes. *Proceedings of MIC-CAI 2017*:498–506
20. Estrada S, Lu R, Conjeti S et al (2020) FatSegNet: A fully automated deep learning pipeline for adipose tissue segmentation on abdominal dixon MRI. *Magn Reson Med* 83:1471–1483
21. Estrada S, Conjeti S, Ahmad M, Navab N, Reuter M (2018) Competition vs. concatenation in skip connections of fully convolutional networks. *Proceedings of international workshop on machine Learning in Medical Imaging*, pp 214–222.
22. Goodfellow IJ, Warde-Farley D, Mirza M, Courville A, Bengio Y (2013) Maxout networks. *Proceedings of International Conference on Machine Learning*, pp 1319–1327.
23. Mehrtash A, Wells WM, Tempany CM, Abolmaesumi P, Kapur T (2020) Confidence calibration and predictive uncertainty estimation for deep medical image segmentation. *IEEE Trans Med Imag* 39:3868–3878
24. Koitka S, Kroll L, Malamutmann E, Oezcelik A, Nensa F (2021) Fully automated body composition analysis in routine CT imaging using 3D semantic segmentation convolutional neural networks. *Eur Radiol* 31:1795–1804

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

## Authors and Affiliations

Sebastian Nowak<sup>1</sup> · Maike Theis<sup>1</sup> · Barbara D. Wichtmann<sup>1</sup> · Anton Faron<sup>1</sup> · Matthias F. Froelich<sup>2</sup> · Fabian Tollens<sup>2</sup> · Helena L. Geißler<sup>1</sup> · Wolfgang Block<sup>1,3,4</sup> · Julian A. Luetkens<sup>1</sup> · Ulrike I. Attenberger<sup>1</sup> · Alois M. Sprinkart<sup>1</sup>

<sup>1</sup> Department of Diagnostic and Interventional Radiology, Quantitative Imaging Lab Bonn (QILaB), University Hospital Bonn, Venusberg-Campus 1, 53127 Bonn, Germany

<sup>2</sup> Department of Radiology and Nuclear Medicine, University Medical Centre Mannheim, Theodor-Kutzer-Ufer 1-3, 68167 Mannheim, Germany

<sup>3</sup> Department of Radiotherapy and Radiation Oncology, University Hospital Bonn, Venusberg-Campus 1, 53127 Bonn, Germany

<sup>4</sup> Department of Neuroradiology, University Hospital Bonn, Venusberg-Campus 1, 53127 Bonn, Germany

3.3. Publication 3: Direct deep learning-based survival prediction from pre-interventional CT prior to transcatheter aortic valve replacement

Theis M, Block W, Luetkens JA, Attenberger UI, Nowak S, Sprinkart AM. **Direct deep learning-based survival prediction from pre-interventional CT prior to transcatheter aortic valve replacement.** Eur J Radiol 2023; 168: 111150

Contents lists available at [ScienceDirect](https://www.sciencedirect.com)

## European Journal of Radiology

journal homepage: [www.elsevier.com/locate/ejrad](http://www.elsevier.com/locate/ejrad)

# Direct deep learning-based survival prediction from pre-interventional CT prior to transcatheter aortic valve replacement

Maike Theis<sup>a,\*</sup>, Wolfgang Block<sup>a,b,c</sup>, Julian A. Luetkens<sup>a</sup>, Ulrike I. Attenberger<sup>a</sup>, Sebastian Nowak<sup>a,1</sup>, Alois M. Sprinkart<sup>a,1</sup>

<sup>a</sup> Department of Diagnostic and Interventional Radiology, Quantitative Imaging Lab Bonn (QILaB), University Hospital Bonn, Venusberg-Campus 1, 53127 Bonn, Germany

<sup>b</sup> Department of Radiotherapy and Radiation Oncology, University Hospital Bonn, Venusberg-Campus 1, 53127 Bonn, Germany

<sup>c</sup> Department of Neuroradiology, University Hospital Bonn, Venusberg-Campus 1, 53127 Bonn, Germany

## ARTICLE INFO

## Keywords:

Deep learning

Survival

Transcatheter aortic valve replacement

Proportional hazards models

Tomography, X-ray computed

## ABSTRACT

**Purpose:** To investigate survival prediction in patients undergoing transcatheter aortic valve replacement (TAVR) using deep learning (DL) methods applied directly to pre-interventional CT images and to compare performance with survival models based on scalar markers of body composition.

**Method:** This retrospective single-center study included 760 patients undergoing TAVR (mean age  $81 \pm 6$  years; 389 female). As a baseline, a Cox proportional hazards model (CPHM) was trained to predict survival on sex, age, and the CT body composition markers fatty muscle fraction (FMF), skeletal muscle radiodensity (SMRD), and skeletal muscle area (SMA) derived from paraspinal muscle segmentation of a single slice at L3/L4 level. The convolutional neural network (CNN) encoder of the DL model for survival prediction was pre-trained in an autoencoder setting with and without a focus on paraspinal muscles. Finally, a combination of DL and CPHM was evaluated. Performance was assessed by C-index and area under the receiver operating curve (AUC) for 1-year and 2-year survival. All methods were trained with five-fold cross-validation and were evaluated on 152 hold-out test cases.

**Results:** The CNN for direct image-based survival prediction, pre-trained in a focussed autoencoder scenario, outperformed the baseline CPHM (CPHM: C-index = 0.608, 1Y-AUC = 0.606, 2Y-AUC = 0.594 vs. DL: C-index = 0.645, 1Y-AUC = 0.687, 2Y-AUC = 0.692). Combining DL and CPHM led to further improvement (C-index = 0.668, 1Y-AUC = 0.713, 2Y-AUC = 0.696).

**Conclusions:** Direct DL-based survival prediction shows potential to improve image feature extraction compared to segmentation-based scalar markers of body composition for risk assessment in TAVR patients.

## 1. Introduction

Transcatheter aortic valve replacement (TAVR) is frequently employed in patients with severe aortic valve stenosis and high surgical risk. Patients with untreated severe aortic valve stenosis have an increased mortality risk, and aortic valve replacement can increase their life expectancy [1]. However, surgical aortic valve replacement (SAVR) is not an option for every patient because of various conditions such as

advanced age or left ventricular dysfunction [2]. In addition to the assessment of surgical risk factors, overall life expectancy plays an important role in the selection of therapy for the treatment of severe aortic valve stenosis. For instance, TAVR is preferable to SAVR in patients with a shorter life expectancy, but it is not recommended in patients with a life expectancy of less than one year [3]. To evaluate the mortality risk of TAVR patients, various clinical parameters or surgical risk scores such as the European System for Cardiac Operative Risk

**Abbreviations:** (TAVR), transcatheter aortic valve replacement; (DL), deep learning; (CPHM), Cox proportional hazards model; (FMF), fatty muscle fraction; (SMRD), skeletal muscle radiodensity; (SMA), skeletal muscle area; (CNN), convolutional neural network; (AUC), area under the curve; (SAVR), surgical aortic valve replacement; (EuroSCORE), European System for Cardiac Operative Risk Evaluation; (HR), hazard ratio, (CI), confidence interval; (AI), artificial intelligence.

\* Corresponding author.

**E-mail addresses:** [Maike.Theis@ukbonn.de](mailto:Maike.Theis@ukbonn.de) (M. Theis), [Wolfgang.Block@ukbonn.de](mailto:Wolfgang.Block@ukbonn.de) (W. Block), [Julian.Luetkens@ukbonn.de](mailto:Julian.Luetkens@ukbonn.de) (J.A. Luetkens), [Ulrike.Attenberger@ukbonn.de](mailto:Ulrike.Attenberger@ukbonn.de) (U.I. Attenberger), [Sebastian.Nowak@ukbonn.de](mailto:Sebastian.Nowak@ukbonn.de) (S. Nowak), [sprinkart@uni-bonn.de](mailto:sprinkart@uni-bonn.de) (A.M. Sprinkart).

<sup>1</sup> Contributed equally to this study.

<https://doi.org/10.1016/j.ejrad.2023.111150>

Received 12 July 2023; Received in revised form 27 September 2023; Accepted 10 October 2023

Available online 11 October 2023

0720-048X/© 2023 Elsevier B.V. All rights reserved.

Evaluation (EuroSCORE) II have been applied [4–6]. In addition, previous studies have shown that patient frailty status is an important risk factor for outcome in TAVR patients and a variety of frailty scores have been investigated, for example, based on questionnaires and/or physical performance tests [6,7]. Recently, human-defined scalar markers of body composition have been introduced to assess frailty, sarcopenia or myosteatosis. The corresponding measurements are usually performed on individual CT slices at L3/L4 lumbar level. The parameters determined in this way can also be taken into account when modelling the mortality risk of TAVR patients [4,8–10].

These scalar markers are derived from tissue segmentations and summarize an image feature, such as skeletal muscle area (SMA) or alterations in tissue density, into a scalar value. To automate the extraction of scalar markers derived from tissue segmentations, deep learning (DL) is typically employed [11,12]. DL methods such as convolutional neural networks (CNN) can autonomously identify and extract relevant image features and feature hierarchies. It is therefore a logical step to use DL not only for automated extraction of human-defined scalar markers through segmentation, but also to explore direct application on unprocessed images for survival prediction.

Several studies have already demonstrated an advantage of direct DL-based prediction of patient survival over classical methods such as Cox proportional hazards models (CPHM) [13–17]. In a CPHM, the patient's log-risk function is represented as a linear combination of several predictor variables [18]. To be able to also model non-linear relationships, Katzman et al. employed a DL method to estimate the patient's log-risk function [13]. Such DL-based analysis has already been successfully applied for survival prediction in patients with oral cancer based only on clinical parameters [14]. Also, in the field of medical imaging, CNN-based time-to-event analyses have been successfully applied to 2D or 3D data and in combination with other relevant information like gene expression data [15–17]. A direct image-based prediction of survival time has not been investigated so far.

Therefore, the aim of our study was to investigate the feasibility of applying a direct image-based DL model for prediction of survival time using a TAVR cohort as an example. The results were compared to established CPHMs based on scalar human-defined body composition markers derived from image segmentation.

## 2. Material and methods

### 2.1. Dataset

Due to the retrospective nature of this single-center study, written informed consent was waived by the institutional review board of the Medical Faculty of the University Bonn. The study was conducted in accordance with the ethical standards of the 1964 Declaration of Helsinki and its subsequent amendments. The patient cohort consists of 811 patients who underwent TAVR at the University Hospital Bonn between 2011 and 2017, with available follow-up data and pre-interventional

thoracic abdominal CT scans. 34 patients were excluded due to insufficient image quality caused e.g., by metallic implants. A further 17 patients were censored before the end of the first year and were therefore also excluded from our analyses. Therefore, the final cohort consists of 760 patients with a mean age of  $81 \pm 6$  years and 389 (51%) female patients. Inclusion and exclusion criteria are presented in a flow chart in Fig. 1. 54% of the included patients died during follow-up with a median survival time of 687 days. For patients with no observed event, median follow-up time was 1548 days. Detailed patient characteristics are shown in Table 1.

For each patient, the scalar body composition markers FMF, mean skeletal muscle radiodensity (SMRD) and skeletal muscle area (SMA) were derived from manual segmentations of the paraspinal musculature at L3/L4 lumbar level previously performed by a radiology resident with three years of experience in abdominal imaging. Detailed descriptions of the extraction of the scalar markers can be found in Appendix A.

For method development, the datasets were randomly divided into 80% ( $n = 608$ ) training and 20% ( $n = 152$ ) test cases, ensuring a similar distribution of deaths, survival times and observation periods in both datasets. Training was performed with five-fold cross-validation. A detailed description of the procedure for splitting the data can be found in Appendix B.

### 2.2. Models

The image pre-processing prior to method development is described in Appendix C.

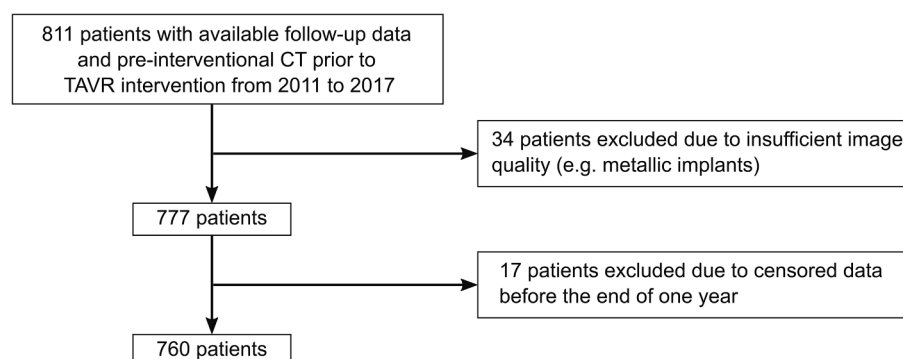
#### 2.2.1. Cox proportional hazards model

A traditional approach for survival prediction was applied to obtain a

**Table 1**

Overview of the patient characteristics of the total dataset ( $n = 760$ ), including sex, event (death), follow-up time and survival time, age, fatty muscle fraction (FMF), mean skeletal muscle radiodensity (SMRD) and skeletal muscle area (SMA).  $Q_1$  refers to the 25%,  $Q_2$  to the 50%, and  $Q_3$  to the 75% quantile.

Patient characteristics		
Variable	Absolute number	Relative number (%)
Sex male / female	371 / 389	48.82% / 51.18%
Event 0 / 1	348 / 412	45.79% / 54.21%
	$Q_1$   $Q_2$   $Q_3$	Range
Follow-up time; event = 0 (days)	1271   1548   2129.5	[365, 3603]
Survival time; event = 1 (days)	204   687   1367.5	[0, 3459]
	Mean $\pm$ Std	Range
Age (years)	81.21 $\pm$ 6.05	[57.00, 96.00]
FMF (%)	62.51 $\pm$ 20.10	[9.97, 97.22]
Mean SMRD (HU)	18.98 $\pm$ 10.80	[-11.27, 49.42]
SMA (cm <sup>2</sup> )	55.89 $\pm$ 10.91	[29.37, 107.03]



**Fig. 1.** Flow chart to illustrate the inclusion and exclusion criteria.

baseline for model comparison. Therefore, the following CPHMs were trained using the Lifelines package in Python (Python 3.9.12, Lifelines 0.27.0) [19]: First, the prognostic value of each scalar body composition marker derived from muscle segmentation (FMF, SMRD and SMA) for predicting survival in TAVR patients was assessed by a univariable analysis. In addition, patient sex and age were also examined as univariable predictors. Therefore, the categorical variable sex was binarized, where male patients were encoded with a value of 1 and female patients with a value of 0. Then, a multivariable CPHM was built using only predictors that showed a significant hazard ratio (HR) in the univariable analysis ( $p$ -value  $< 0.05$ ) ( $\text{CPHM}_{\text{Multivar,sign}}$ ). Lastly, a second multivariable CPHM including all predictors (FMF, SMRD, SMA, sex, age) was investigated, independent from its significance in univariable analysis ( $\text{CPHM}_{\text{Multivar,all}}$ ).

The univariable CPHMs were developed on the first training set from cross-validation, which included 486 cases. Both multivariable CPHMs were trained with five-fold cross-validation, and an ensemble of all five models was applied to the hold-out test set. A general description of a CPHM can be found in Appendix D.

### 2.2.2. Deep learning based survival prediction

As a new approach, a DL model was trained for direct image-based survival time prediction. Fig. 2 shows the CNN architecture developed for predicting patient survival directly on the unsegmented CT slices at L3/L4 lumbar level. In the first part of the network (encoder), relevant image features are extracted by using multiple convolutional layers. In the second part of the network, the mortality risk is predicted based on the encoded image features using fully connected linear layers, which finally output the logistic hazard rate as a single scalar value. Lifelines 0.27.0 was used to assess the probability of survival at a given time point based on the predicted log-hazard [19]. The loss function for training the CNN-based survival prediction is the negative log Cox partial likelihood divided by the number of observed events, which is similar to the loss used for training of the CPHMs [13,18].

We investigated autoencoder based pre-training of the convolutional layers of the CNN encoder to mitigate overfitting. Autoencoder pre-training involves connecting a CNN encoder to a CNN decoder via a bottleneck. This forces the encoder to learn to compress characteristic image features so that the decoder can reconstruct the original image. The CNN's encoder weights for survival prediction are then initialized with the corresponding pre-trained autoencoder weights.

Two different versions of L1-loss for autoencoder-based pre-training were examined: First, a standard L1-loss was used that considers all image areas equally. Second, a masked L1-loss was used with a focus on paraspinal musculature, which forces the encoder to preserve more image detail in this specific region containing prognostic information for survival prediction [4,8,9,20].

Details on the autoencoder pre-training can be found in Appendix E. To investigate the benefit of these two autoencoder-based pre-training strategies, a further DL model was trained from scratch, i.e., without pre-training of the encoder ( $\text{DL}_{\text{Scratch}}$ ). We refer to the survival prediction CNN with and without focus on the paraspinal musculature in pre-training as  $\text{DL}_{\text{Masked}}$  and  $\text{DL}_{\text{Unmasked}}$ . For training of  $\text{DL}_{\text{Masked}}$  and  $\text{DL}_{\text{Unmasked}}$ , the weights of the pre-trained encoder are kept frozen ( $\text{DL}_{\text{Masked,frozen}}$  and  $\text{DL}_{\text{Unmasked,frozen}}$ ) [21–23]. The best approach of  $\text{DL}_{\text{Masked,frozen}}$ ,  $\text{DL}_{\text{Unmasked,frozen}}$  and  $\text{DL}_{\text{Scratch}}$  was selected by training and evaluating on the first validation split and then trained with full five-fold cross-validation and evaluated on the hold-out test set. To investigate the benefits of altering the pre-trained parameters for survival prediction, this best frozen model was further trained with unfrozen weights ( $\text{DL}_{\text{Unfrozen}}$ ).

Finally, a combination of the baseline CPHM and the direct image-based DL approach was evaluated by implementing a further CPHM using the parameters sex, age, and the log-hazard rate of each patient predicted by the best DL model as predictor variables ( $\text{CPHM}_{\text{DL+Sex+Age}}$ ) [16].

For all DL methods, a grid search for hyperparameters such as

learning rate, weight decay, and dropout rate was conducted. For more details on the experimental design and grid searches, see Appendix F.

### 2.3. Comparison to EuroSCORE

To evaluate the clinical utility of the DL model also in comparison with the surgical risk scores EuroSCORE and EuroSCORE II [24–26], two further CPHMs based on age and sex and EuroSCORE ( $\text{CPHM}_{\text{EuroSCORE-E+Sex+Age}}$ ) and EuroSCORE II ( $\text{CPHM}_{\text{EuroSCOREII+Sex+Age}}$ ) were evaluated, respectively. In 90 of 760 patients, only the original EuroSCORE was available, as EuroSCORE II was first introduced in 2012.

### 2.4. Statistical evaluation

As a standard metric for evaluating time-to-event analysis, the C-index was calculated for comparison of model performance on the validation and hold-out test data [27,28]. The area under the receiver operating curve (AUC) for the prediction of 1-year and 2-year survival was additionally assessed on the hold-out test set, as this is a more intuitive metric for evaluating survival time prediction. All included patients had at least 1-year follow-up available. For the calculation of 2-year survival AUC, patients without 2-year follow-up data had to be excluded ( $n = 6$ ). To assess significant differences in performance, 95% confidence intervals (CI) were calculated for all metrics by bootstrapping the test set with 1000 resamples.

Lastly, Kaplan-Meier analyses with log-rank tests for 1-year and 2-year survival were conducted on the test data based on the predicted log-hazard rate of the best-performing DL model. To stratify patients into low- and high-risk groups, the median of all predicted log-hazard rates in the five validation cohorts was set as a cut-off value. A  $p$ -value  $< 0.05$  or non-overlapping 95% CIs were considered statistically significant [29].

## 3. Results

### 3.1. Cox proportional hazards model

The results of the univariable and multivariable CPHM analyses are shown in Table 2. In univariable analysis, the scalar markers FMF, SMRD, and SMA were observed to be significant predictors. Only SMA remained significant in the multivariable CPHM analysis employing solely these significant predictors ( $\text{CPHM}_{\text{Multivar,sign}}$ ). SMA and sex showed significant hazard ratios in the CPHM including all investigated variables ( $\text{CPHM}_{\text{Multivar,all}}$ ). Poor performance with a C-index of 0.508, an AUC for 1- and 2-year survival with 0.496 and 0.457 was observed applying an ensemble of all five cross-validated  $\text{CPHM}_{\text{Multivar,sign}}$  to the hold-out test set. For  $\text{CPHM}_{\text{Multivar,all}}$  a C-index of 0.608 and AUC values for 1- and 2-year survival of 0.606 and 0.594 were observed (see Table 4).

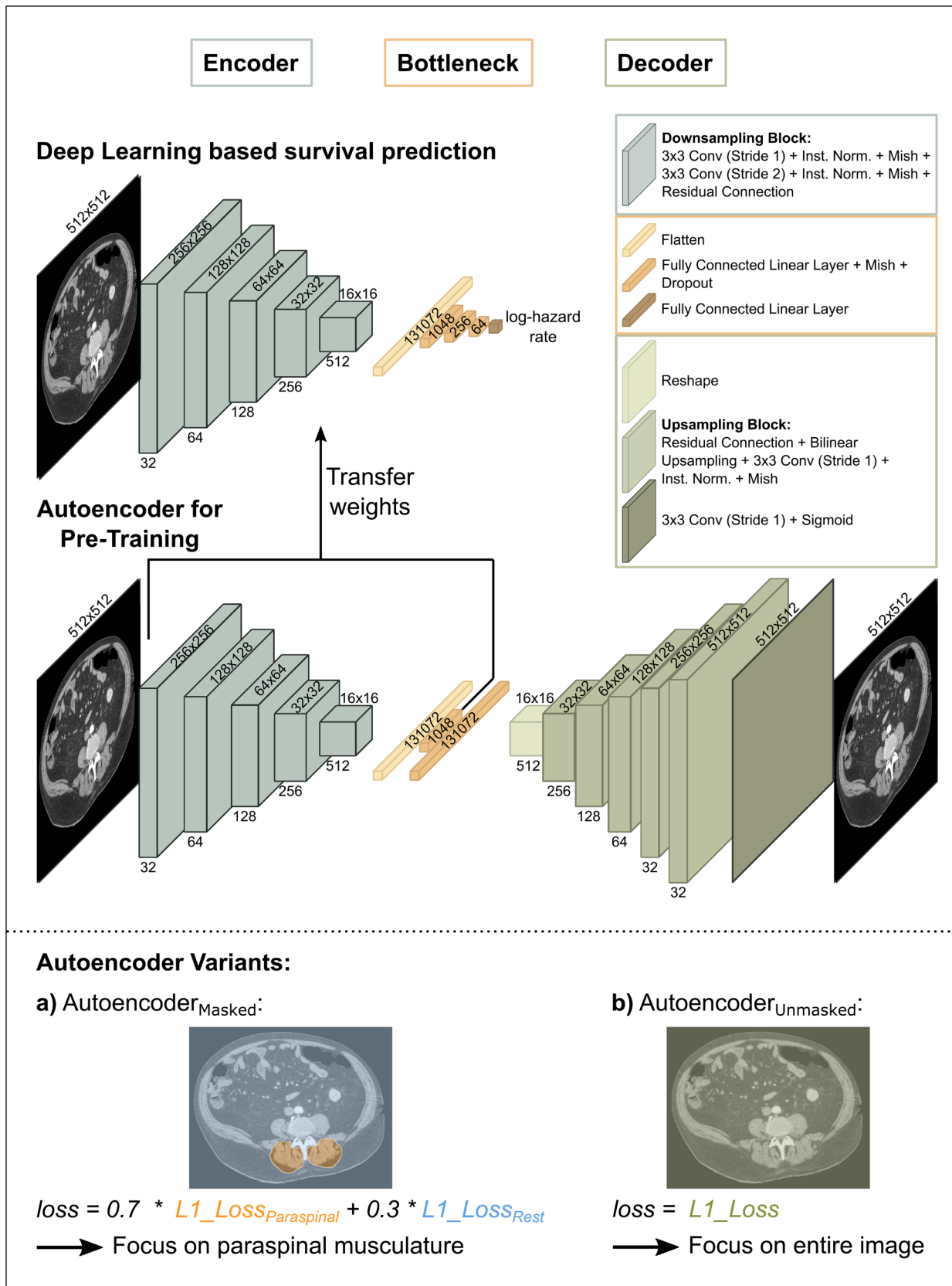
### 3.2. Deep learning based survival prediction

The performance values for the three different DL variants ( $\text{DL}_{\text{Scratch}}$ ,  $\text{DL}_{\text{Masked,frozen}}$ ,  $\text{DL}_{\text{Unmasked,frozen}}$ ) are listed in Table 3. The  $\text{DL}_{\text{Masked,frozen}}$  model showed the highest performance with a C-index of 0.636. Results of the corresponding hyperparameter optimization are listed in Appendix G.

Training of the  $\text{DL}_{\text{Masked,frozen}}$  model on all five folds and testing the ensemble on the hold-out test data resulted in a C-index of 0.637 and AUC values for 1- and 2-year survival of 0.687 and 0.683 respectively. The C-index increased slightly to 0.645 and the 2-year AUC increased to 0.692 after subsequent training with unfrozen weights of the encoder ( $\text{DL}_{\text{Unfrozen}}$ ) (see Table 4). Results of the corresponding grid search for hyperparameter optimization can be found in Appendix H.

A significantly higher C-index was achieved for the  $\text{DL}_{\text{Unfrozen}}$  model compared to the  $\text{CPHM}_{\text{Multivar,sign}}$ , which only includes the three scalar





**Fig. 2.** Overview of the investigated pre-training strategies for the development of an image-based survival prediction. Two autoencoders were trained, one with and one without focusing on paraspinal musculature. Pre-trained weights were afterwards transferred to the deep learning model that predicts patient survival.

**Table 2**

Results for univariable and multivariable analysis for a Cox proportional hazards model (CPHM) trained on the first of five training sets from cross-validation. The following predictors were considered: fatty muscle fraction (FMF), mean skeletal muscle radiodensity (SMRD), skeletal muscle area (SMA), sex, and age. Hazard ratios (HR) are given with 95% confidence intervals and p-values indicating significance of the predictors (\*: p-value < 0.05). Two multivariable CPHMs were investigated, one with only predictors that were significant in univariable analysis (CPHM<sub>Multivar,sign</sub>) and another including all variables (CPHM<sub>Multivar,all</sub>).

Variables	Univariable analysis		Multivariable analysis			
	HR	p-value	CPHM <sub>Multivar,sign</sub>		CPHM <sub>Multivar,all</sub>	
			HR	p-value	HR	p-value
FMF (%)	1.01 [1.004, 1.016]	<0.01*	1.00 [0.971, 1.029]	0.98	1.00 [0.968, 1.026]	0.82
Mean SMRD (HU)	0.98 [0.971, 0.992]	<0.01*	0.98 [0.931, 1.033]	0.46	0.97 [0.916, 1.019]	0.21
SMA (cm <sup>2</sup> )	0.99 [0.976, 0.999]	0.03*	0.99 [0.975, 0.998]	0.02*	0.98 [0.963, 0.989]	<0.01*
Sex	1.16 [0.909, 1.474]	0.23	–	–	1.79 [1.341, 2.380]	<0.01*
Age (years)	1.02 [0.998, 1.042]	0.08	–	–	1.00 [0.979, 1.027]	0.84

**Table 3**

Comparison of the DL models trained from scratch (DL<sub>scratch</sub>), pre-trained on the masked autoencoder (DL<sub>Masked,frozen</sub>) and pre-trained on the standard autoencoder (DL<sub>Unmasked,frozen</sub>). The results presented correspond to the best performance values for each model after an individual performed parameter tuning. The epoch column indicates the number of the epoch in which the lowest validation loss was observed. The model with the highest performance is marked in bold.

Model	Loss	Epoch	C-index
DL <sub>scratch</sub>	4.09	31	0.609
<b>DL<sub>Masked,frozen</sub></b>	<b>4.05</b>	<b>43</b>	<b>0.636</b>
DL <sub>Unmasked,frozen</sub>	4.09	29	0.632

**Table 4**

Performance values of all examined methods on the hold-out test set (n = 152) together with 95%-confidence intervals in brackets. The model with the highest performance is marked in bold.

Performance on hold-out test			
Model	C-index	AUC 1Y	AUC 2Y
DL <sub>Masked,frozen</sub>	0.637 [0.570, 0.701]	0.687 [0.567, 0.792]	0.683 [0.583, 0.773]
<b>DL<sub>Unfrozen</sub></b>	<b>0.645</b> <b>[0.580, 0.706]</b>	<b>0.687</b> <b>[0.564, 0.792]</b>	<b>0.692</b> <b>[0.594, 0.777]</b>
CPHM <sub>Multivar,sign</sub>	0.508 [0.439, 0.578]	0.496 [0.389, 0.614]	0.457 [0.349, 0.567]
CPHM <sub>Multivar,all</sub>	0.608 [0.543, 0.676]	0.606 [0.493, 0.720]	0.594 [0.488, 0.700]

markers FMF, SMRD and SMA. The performance of the DL approach was also higher compared to CPHM<sub>Multivar,all</sub>, which additionally included sex and age (see Table 4). Fig. 3 shows the Kaplan-Meier analysis of the log-hazard rate predicted by the DL<sub>Unfrozen</sub> model. Here, a significant difference was found between patients with a high log hazard rate ( $\geq 0.91$ ) to patients with low hazard rates predicted by the DL model for 1-year (p = 0.04) and 2-year survival (p < 0.01). When combining the log-hazard rate predicted by DL<sub>Unfrozen</sub> in a CPHM together with age and sex (CPHM<sub>DL+Sex+Age</sub>) the C-index increased to 0.668 and AUC values for 1- and 2-year survival increased to 0.713 and 0.696 (see Table 5).

### 3.3. Comparison to EuroSCORE

Performance values for the two models developed on the basis of EuroSCORE and EuroSCORE II are presented in Table 6. For both EuroSCORE models, C-index as well as AUC of 1- and 2-year survival was lower compared to the direct image-based DL approach.

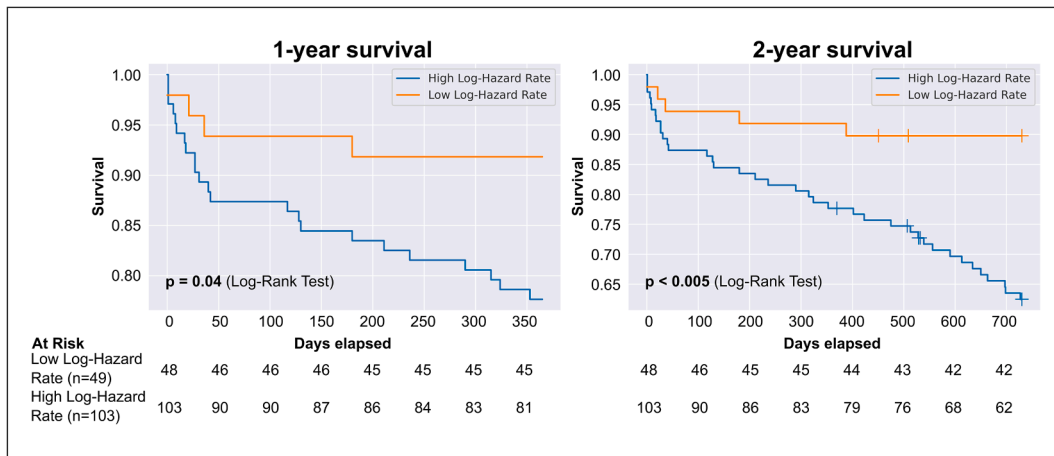
## 4. Discussion

In this study, we investigated the feasibility of DL for direct image-based survival prediction on pre-interventional CT of patients undergoing TAVR. The results were compared to CPHMs based on established scalar markers of body composition. The study shows that direct application of a thoroughly optimized image-based DL model has the potential to improve survival prediction compared to the application of scalar body composition markers.

Until now there are only a few studies that have investigated DL-based survival prediction directly on imaging data. In a previous study, a similar CNN was developed to predict loco-regional tumour control from 2D and 3D CT data [15]. In that study, an improvement was observed for the DL model based solely on CT image data in comparison to the clinical model developed using CPHM. In another study, DL-based prediction of survival time based on CT and PET image data has been examined in combination with clinical parameters for predicting survival time and other time-to-event outcomes in patients with oral cavity cancer [16]. Based on the promising results presented in these papers, our main concern was to investigate whether important information for predicting survival time can be obtained from abdominal CT examination alone using DL approaches. Furthermore, we investigated how such a DL model can be trained most efficiently. This was performed using a TAVR cohort as an example.

The machine learning-based analysis of user-selected scalar features or the autonomous selection of relevant image features by DL are two different approaches for the development of artificial intelligence (AI) models in radiology. However, several studies have reported that the use of DL over or in combination with the analysis of hand-crafted features can provide improved performance in various tasks [15–17,30]. As an example, combining CNN-based information extraction from chest CT scans with established quantitative features extracted from lesions of patients with lung adenocarcinoma has been shown to improve risk assessment [31]. However, the images examined in the present study do not show any pathology of primary interest, such as lesions. Instead, an abdominal slice from a pre-interventional CT is analysed, for which it was shown that scalar body composition markers derived from the paraspinal musculature carry prognostic information for various conditions [4,8–10,20].

The fact that the DL model directly applied to an abdominal image improved the risk assessment in the studied cohort can be attributed to the ability of a CNN to identify relevant features and feature hierarchies. For a given task, a CNN optimizes its convolution kernels autonomously and is therefore not limited to the analysis of human-defined image features. This is also an advantage over traditional methods such as CPHMs, where fixed and user-defined predictor variables, such as SMA, must be defined for method development. An extensive analysis of all variables is therefore required to ensure that only relevant predictors are considered. Unlike CPHM, the DL approach is also able to model more complex non-linear relationships between the hazard rate and the predictor variables. On the downside, the unconstrained feature exploration also makes the DL method more prone to overfitting to irrelevant features of the training data [30]. To address this issue, we investigated an autoencoder-based pre-training of the CNN encoder. Interestingly, we found that it is useful to incorporate prior knowledge from body composition analyses when training the autoencoder model. The use of a masked loss that forces focusing on the paraspinal muscles in the pre-training step led to a higher performance of the final DL model for the prediction of patient risk.



**Fig. 3.** Kaplan-Meier curves for 1- and 2-year survival. The figure illustrates Kaplan-Meier curves for patients in the hold-out test group ( $n = 152$ ) stratified by low (orange) and high (blue) predicted log-hazard rate from the  $DL_{\text{Unfrozen}}$  model, whose weights were unfrozen after previous training with frozen pre-trained autoencoder weights focusing on the paraspinal muscles. The cut-off value for stratification into low and high log-hazard rates was determined as the median of the predicted log-hazard rates from all five validation sets. Censored cases were indicated by a plus sign (+). The log-rank test shows that the probability of survival for patients with high predicted log-hazard rates is significantly lower than for patients with low predicted risk for both one-year ( $p = 0.04$ ) and two-year survival ( $p < 0.01$ ). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

**Table 5**

Performance values on the hold-out test set ( $n = 152$ ) for the CPHM model trained on sex, age, and the predicted log-hazard score from the  $DL_{\text{Unfrozen}}$  model together with 95%-confidence intervals in brackets.

Combination of CPHM and $DL_{\text{Unfrozen}}$	Performance		
	C-index	AUC 1Y	AUC 2Y
$CPHM_{DL+Sex+Age}$	0.668 [0.600, 0.726]	0.713 [0.600, 0.815]	0.696 [0.611, 0.780]

**Table 6**

Performance values together with 95% confidence intervals for both CPHM models trained on sex, age and EuroSCORE or EuroSCORE II, which were evaluated on the hold-out test cases.

Performance on hold-out test				
Model	n	C-index	AUC 1Y	AUC 2Y
$CPHM_{EuroSCORE+Sex+Age}$	152	0.615 [0.546, 0.681]	0.647 [0.529, 0.765]	0.601 [0.493, 0.701]
$CPHM_{EuroSCOREII+Sex+Age}$	139	0.609 [0.542, 0.676]	0.647 [0.514, 0.767]	0.599 [0.485, 0.702]

Two multivariable CPHMs were developed and evaluated to compare the direct DL-based evaluation of images with the established analysis of human-defined scalar markers for outcome assessment of TAVR patients. Very limited predictive power was found for the first multivariable CPHM including FMF, SMRD, and SMA on the hold-out test data. By adding sex and age into the multivariable CPHM the performance increased, although no significant hazard ratios were observed for these two predictors in univariable analysis. The outcome indicates the potential value of including further clinical parameters along with scalar markers derived from image segmentation for survival prediction. Other studies also included functional and clinical parameters in combination with established body composition markers in a CPHM for survival prediction in TAVR patients [4]. Apart from age and sex, no other clinical information was included in these CPHM models, as the main aim of this proof-of-concept study was to examine potential benefits of DL for direct image-based survival prediction. Nevertheless, we also evaluated two additional CPHMs based on sex, age, and the surgical

risk scores EuroSCORE and EuroSCORE II respectively. Although these scores are not primarily developed to estimate the life expectancy of a TAVR patient but aim to assess the surgical risk, both models showed also predictive value for patient survival. However, the performance was lower than the direct image-based DL approach regarding all evaluated metrics. Future studies are warranted to investigate the benefit of considering more comprehensive clinical information and combining this data with multimodal DL architectures to further improve patient outcome assessment.

In this context, the utilization of robust survival prediction models for patients undergoing TAVR offers an additional dimension to aid cardiologists in making informed therapy decisions. While e.g., the 1-year survival prediction has the potential to serve as a valuable adjunct, it is imperative to underscore that therapy decisions must be made through a comprehensive assessment of various clinical factors. The integration of survival prediction models into clinical practice represents an evolving area, and its true impact on decision-making should be the subject of further scientific investigation.

A limitation of our work is that the investigated methods were only applied to 2D data and thus the extraction of relevant image information is limited to this specific slice. However, body composition analysis is usually performed on 2D slices at a certain lumbar level, as a high correlation to 3D measurements has been demonstrated [32–34]. The slice extraction can also be performed automatically so that no manual input is required, and the application of the developed DL method could be completely automated end-to-end [11]. Nevertheless, it may be worthwhile to investigate a 3D application of the method and to develop a direct image-based DL model for survival prediction on 3D CT data. Again, it may be investigated whether a focus on the paraspinal musculature is beneficial and automated methods such as the Total-Segmentator could be used for the 3D segmentation [35]. It should be noted, however, that a 3D approach will be much more susceptible to overfitting. A further limitation of the DL-based survival prediction is that the interpretation of the rationale behind the decision of the CNN is not straightforward for humans. However, the aspect of interpretability is crucial for gaining confidence in DL prediction and also to identify potential new image-based biomarkers that could be specifically targeted. So far, methods of explainable AI are still very limited when it comes to bringing more transparency to individual decisions, e.g., by providing only rough and unspecific saliency maps [36]. Another limitation of the study is the use of single institution data. Multi-center studies with heterogeneous datasets are warranted to demonstrate

general applicability, which is also considered a preferred approach to validate predictive DL models over the use of explanatory AI methods by some researchers [36].

## 5. Conclusions

This study demonstrates the potential of direct image-based outcome assessment by DL on pre-interventional abdominal CT in patients undergoing TAVR, offering improved image feature extraction compared to the assessment of human-defined scalar body composition metrics.

## CRedit authorship contribution statement

**Maïke Theis:** Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing – original draft, Visualization. **Wolfgang Block:** Writing – review & editing, Resources, Data curation. **Julian A. Luetkens:** Writing – review & editing, Data curation, Conceptualization. **Ulrike I. Attenberger:** Writing – review & editing, Project administration, Funding acquisition. **Sebastian Nowak:** Writing – review & editing, Validation, Supervision, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Alois M. Sprinkart:** Conceptualization, Methodology, Validation, Formal analysis, Investigation, Data curation, Writing – review & editing, Supervision, Project administration.

## Declaration of Competing Interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: S. N. was funded over a part of the study duration by RACOON (NUM), which is supported by the Federal Ministry of Education and Research of Germany under BMBF grant number 01KX2121. M.T. was funded over a part of the study duration by a grant from the BONFOR research program of the University of Bonn (application number 2020-2A-04). The funders had no influence on the conception and design of the study, the data analysis, the data collection, the preparation of the manuscript, and the decision to publish.

## Appendix A–G. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.ejrad.2023.111150>.

## References

- [1] P. Varadarajan, N. Kapoor, R.C. Bansal, R.G. Pai, Survival in elderly patients with severe aortic stenosis is dramatically improved by aortic valve replacement: results from a cohort of 277 patients aged  $\geq 80$  years, *Eur. J. Cardiothorac. Surg.* 30 (2006) 722–727, <https://doi.org/10.1016/j.ejcts.2006.07.028>.
- [2] B. Iung, A. Cachier, G. Baron, D. Messika-Zeitoun, F. Delahaye, P. Tornos, C. Gohlke-Bärwolf, E. Boersma, P. Ravaud, A. Vahanian, Decision-making in elderly patients with severe aortic stenosis: why are so many denied surgery? *Eur. Heart J.* 26 (2005) 2714–2720, <https://doi.org/10.1093/eurheartj/ehi471>.
- [3] C.M. Otto, R.A. Nishimura, R.O. Bonow, B.A. Carabello, J.P. Erwin, F. Gentile, H. Jneid, E.V. Krieger, M. Mack, C. McLeod, P.T. O’Gara, V.H. Rigolin, T.M. Sundt, A. Thompson, C. Toly, ACC/AHA guideline for the management of patients with valvular heart disease: a report of the American college of cardiology/American heart association joint committee on clinical practice guidelines, *Circulation* 143 (2021) (2020) e72–e227, <https://doi.org/10.1161/CIR.0000000000000923>.
- [4] J.A. Luetkens, A. Faron, H.L. Geissler, B. Al-Kassou, J. Shamekhi, A. Stundl, A. M. Sprinkart, C. Meyer, R. Fimmers, H. Treede, E. Grube, G. Nickenig, J.-M. Sinning, D. Thomas, Opportunistic computed tomography imaging for the assessment of fatty muscle fraction predicts outcome in patients undergoing transcatheter aortic valve replacement, *Circulation* 141 (2020) 234–236, <https://doi.org/10.1161/CIRCULATIONAHA.119.042927>.
- [5] K. Maeda, T. Kuratani, K. Pak, K. Shimamura, I. Mizote, S. Miyagawa, K. Toda, Y. Sakata, Y. Sawa, Development of a new risk model for a prognostic prediction after transcatheter aortic valve replacement, *Gen. Thorac. Cardiovasc. Surg.* 69 (2021) 44–50, <https://doi.org/10.1007/s11748-020-01436-w>.
- [6] T. Shimura, M. Yamamoto, S. Kano, A. Kagase, A. Kodama, Y. Koyama, E. Tsuchikane, T. Suzuki, T. Otsuka, S. Kohsaka, N. Tada, F. Yamanaka, T. Naganuma, M. Araki, S. Shirai, Y. Watanabe, K. Hayashida, F. Yashima, T. Inohara, Y. Kakefuda, T. Arai, R. Yanagisawa, M. Tanaka, T. Kawakami, Y. Maekawa, K. Takashi, A. Yoshitake, Y. Iida, M. Yamazaki, H. Shimizu, Y. Yamada, M. Jinzaki, H. Tsuruta, Y. Itabashi, M. Murata, M. Kawakami, S. Fukui, M. Sano, K. Fukuda, S. Hosoba, H. Sato, T. Teramoto, M. Kimura, M. Sago, T. Tsunaki, S. Watarai, M. Tsuzuki, K. Irokawa, K. Shimizu, T. Kobayashi, Y. Okawa, M. Miyasaka, Y. Enta, K. Shishido, T. Ochiai, T. Yamabe, K. Noguchi, S. Saito, H. Kawamoto, H. Onishi, H. Yabushita, S. Mitomo, S. Nakamura, M. Yamawaki, Y. Akatsu, Y. Honda, T. Takama, A. Isotani, M. Hayashi, N. Kamioka, M. Miura, T. Morinaga, T. Kawaguchi, M. Yano, M. Hanyu, Y. Arai, H. Tsubota, M. Kudo, Y. Kuroda, A. Kataoka, H. Hioki, Y. Nara, H. Kawashima, F. Nagura, M. Nakashima, K. Sasaki, J. Nishikawa, T. Shimokawa, T. Harada, K. Kozuma, Impact of the clinical frailty scale on outcomes after transcatheter aortic valve replacement, *Circulation* 135 (2017) 2013–2024, <https://doi.org/10.1161/CIRCULATIONAHA.116.025630>.
- [7] J. Afilalo, S. Lauck, D.H. Kim, T. Lefevre, N. Piazza, K. Lachapelle, G. Martucci, A. Lamy, M. Labina, M.D. Peterson, R.C. Arora, N. Noisieux, A. Rassi, I.F. Palacios, P. G n reux, B.R. Lindman, A.W. Asgar, C.A. Kim, A. Trnkus, J.A. Morais, Y. Langlois, L.G. Rudski, J.-F. Morin, J.J. Popma, J.G. Webb, L.P. Perrault, Frailty in older adults undergoing aortic valve replacement: the FRAILTY-AVR study, *J. Am. Coll. Cardiol.* 70 (2017) 689–700, <https://doi.org/10.1016/j.jacc.2017.06.024>.
- [8] M. Soud, F. Alahdab, G. Ho, K.O. Kuku, M. Cejudo-Tejeda, A. Hideo-Kajita, P. de Araujo Goncalves, R.C. Teles, R. Waksman, H.M. Garcia-Garcia, Usefulness of skeletal muscle area detected by computed tomography to predict mortality in patients undergoing transcatheter aortic valve replacement: a meta-analysis study, *Int. J. Cardiovasc. Imaging* 35 (2019) 1141–1147, <https://doi.org/10.1007/s10554-019-01582-0>.
- [9] A. Faron, S. Kreyer, A.M. Sprinkart, T. Muders, S.F. Ehrentraut, A. Isaak, R. Fimmers, C.C. Pieper, D. Kuetting, J.-C. Schewe, U. Attenberger, C. Putensen, J. A. Luetkens, CT fatty muscle fraction as a new parameter for muscle quality assessment predicts outcome in venovenous extracorporeal membrane oxygenation, *Sci. Rep.* 10 (2020) 22391, <https://doi.org/10.1038/s41598-020-79495-5>.
- [10] A. Faron, A.M. Sprinkart, D.L.R. Kuetting, A. Feisst, A. Isaak, C. Enderl, J. Chang, S. Nowak, W. Block, D. Thomas, U. Attenberger, J.A. Luetkens, Body composition analysis using CT and MRI: intra-individual intermodal comparison of muscle mass and myosteatosis, *Sci. Rep.* 10 (2020) 11765, <https://doi.org/10.1038/s41598-020-68797-3>.
- [11] S. Nowak, M. Theis, B.D. Wichtmann, A. Faron, M.F. Froelich, F. Tollens, H. L. Gei ler, W. Block, J.A. Luetkens, U.I. Attenberger, A.M. Sprinkart, End-to-end automated body composition analyses with integrated quality control for opportunistic assessment of sarcopenia in CT, *Eur. Radiol.* (2021), <https://doi.org/10.1007/s00330-021-08313-x>.
- [12] S. Nowak, A. Faron, J.A. Luetkens, H.L. Gei ler, M. Praktiknjo, W. Block, D. Thomas, A.M. Sprinkart, Fully automated segmentation of connective tissue compartments for CT-based body composition analysis: a deep learning approach, *Invest. Radiol.* 55 (2020) 357, <https://doi.org/10.1097/RLI.0000000000000647>.
- [13] J.L. Katzman, U. Shaham, A. Cloninger, J. Bates, T. Jiang, Y. Kluger, DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network, *BMC Med. Res. Method.* 18 (2018) 24, <https://doi.org/10.1186/s12874-018-0482-1>.
- [14] D.W. Kim, S. Lee, S. Kwon, W. Nam, I.-H. Cha, H.J. Kim, Deep learning-based survival prediction of oral cancer patients, *Sci. Rep.* 9 (2019) 6994, <https://doi.org/10.1038/s41598-019-43372-7>.
- [15] S. Starke, S. Leger, A. Zwanenburg, K. Leger, F. Lohaus, A. Linge, A. Schreiber, G. Kalinauskaitė, I. Tinhofer, N. Guberina, M. Guberina, P. Balermipas, J. von der Gr n, U. Ganswindt, C. Belka, J.C. Peeken, S.E. Combs, S. Boeke, D. Zips, C. Richter, E.G.C. Troost, M. Krause, M. Baumann, S. L ck, 2D and 3D convolutional neural networks for outcome modelling of locally advanced head and neck squamous cell carcinoma, *Sci. Rep.* 10 (2020) 15625, <https://doi.org/10.1038/s41598-020-70542-9>.
- [16] P. Afshar, A. Mohammadi, P.N. Tyrrell, P. Cheung, A. Sigiuk, K.N. Plataniotis, E. T. Nguyen, A. Oikonomou, DRTOP: deep learning-based radiomics for the time-to-event outcome prediction in lung cancer, *Sci. Rep.* 10 (2020) 12366, <https://doi.org/10.1038/s41598-020-69106-8>.
- [17] L.A. Vale-Silva, K. Rohr, Long-term cancer survival prediction using multimodal deep learning, *Sci. Rep.* 11 (2021) 13505, <https://doi.org/10.1038/s41598-021-92799-4>.
- [18] D.R. Cox, Regression models and life-tables, *J. R. Stat. Soc. B. Methodol.* 34 (1972) 187–202, <https://doi.org/10.1111/j.2517-6161.1972.tb00899.x>.
- [19] C. Davidson-Pilon, lifelines: survival analysis in Python, *J. Open Source Softw.* 4 (2019) 1317, <https://doi.org/10.21105/joss.01317>.
- [20] A. Faron, N.S. Opheys, S. Nowak, A.M. Sprinkart, A. Isaak, M. Theis, N. Mesropyan, C. Enderl, J. Sirokay, C.C. Pieper, D. Kuetting, U. Attenberger, J. Landsberg, J. A. Luetkens, Deep learning-based body composition analysis predicts outcome in melanoma patients treated with immune checkpoint inhibitors, *Diagnostics* 11 (2021) 2314, <https://doi.org/10.3390/diagnostics11122314>.
- [21] S. Nowak, N. Mesropyan, A. Faron, W. Block, M. Reuter, U.I. Attenberger, J. A. Luetkens, A.M. Sprinkart, Detection of liver cirrhosis in standard T2-weighted MRI using deep transfer learning, *Eur. Radiol.* 31 (2021) 8807–8815, <https://doi.org/10.1007/s00330-021-07858-1>.
- [22] R. Mormont, P. Geurts, R. Maree, Comparison of Deep Transfer Learning Strategies for Digital Pathology, in: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, 2018, pp. 2262–2271.
- [23] J.A. Luetkens, S. Nowak, N. Mesropyan, W. Block, M. Praktiknjo, J. Chang, C. Bauchhage, R. Sifa, A.M. Sprinkart, A. Faron, U. Attenberger, Deep learning

- supports the differentiation of alcoholic and other-than-alcoholic cirrhosis based on MRI, *Sci. Rep.* 12 (2022) 8297, <https://doi.org/10.1038/s41598-022-12410-2>.
- [24] F. Roques, S.A. Nashef, P. Michel, E. Gauducheau, C. de Vincentiis, E. Baudet, J. Cortina, M. David, A. Faichney, F. Gabrielle, E. Gams, A. Harjula, M.T. Jones, P.P. Pintor, R. Salamon, L. Thulin, Risk factors and outcome in European cardiac surgery: analysis of the EuroSCORE multinational database of 19030 patients, *Eur. J. Cardio-Thorac. Surg. Off. J. Eur. Assoc. Cardio-Thorac. Surg.* 15 (1999) 816–822; discussion 822–823. Doi: 10.1016/s1010-7940(99)00106-2.
- [25] S.A.M. Nashef, F. Roques, P. Michel, E. Gauducheau, S. Lemeshow, R. Salamon, the EuroSCORE study group, European system for cardiac operative risk evaluation (EuroSCORE), *Eur. J. Cardiothorac. Surg.* 16 (1999) 9–13, [https://doi.org/10.1016/S1010-7940\(99\)00134-7](https://doi.org/10.1016/S1010-7940(99)00134-7).
- [26] S.A.M. Nashef, F. Roques, L.D. Sharples, J. Nilsson, C. Smith, A.R. Goldstone, U. Lockowandt, EuroSCORE II†, *Eur. J. Cardiothorac. Surg.* 41 (2012) 734–745, <https://doi.org/10.1093/ejcts/ezs043>.
- [27] F.E. Harrell Jr., R.M. Califf, D.B. Pryor, K.L. Lee, R.A. Rosati, Evaluating the yield of medical tests, *J. Am. Med. Assoc.* 247 (1982) 2543–2546, <https://doi.org/10.1001/jama.1982.03320430047030>.
- [28] F.E. Harrell Jr., K.L. Lee, R.M. Califf, D.B. Pryor, R.A. Rosati, Regression modelling strategies for improved prognostic prediction, *Stat. Med.* 3 (1984) 143–152, <https://doi.org/10.1002/sim.4780030207>.
- [29] G. Cumming, Inference by eye: Reading the overlap of independent confidence intervals, *Stat. Med.* 28 (2009) 205–220, <https://doi.org/10.1002/sim.3471>.
- [30] A. Hosny, C. Parmar, J. Quackenbush, L.H. Schwartz, H.J.W.L. Aerts, Artificial intelligence in radiology, *Nat. Rev. Cancer* 18 (2018) 500–510, <https://doi.org/10.1038/s41568-018-0016-5>.
- [31] R. Paul, S.H. Hawkins, Y. Balagurunathan, M. Schabath, R.J. Gillies, L.O. Hall, D. B. Goldgof, Deep feature transfer learning in combination with traditional features predicts survival among patients with lung adenocarcinoma, *Tomography*. 2 (2016) 388–395, <https://doi.org/10.18383/j.tom.2016.00211>.
- [32] W. Shen, M. Punyanitya, Z. Wang, D. Gallagher, M.-P. St-Onge, J. Albu, S.B. Heymsfield, S. Heshka., Total body skeletal muscle and adipose tissue volumes: estimation from a single abdominal cross-sectional image, *J. Appl. Physiol.* 97 (2004) 2333–2338, <https://doi.org/10.1152/jappphysiol.00744.2004>.
- [33] A. Faron, J.A. Luetkens, F.C. Schmeel, D.L.R. Kuetting, D. Thomas, A.M. Sprinkart, Quantification of fat and skeletal muscle tissue at abdominal computed tomography: associations between single-slice measurements and total compartment volumes, *Abdom. Radiol.* 44 (2019) 1907–1916, <https://doi.org/10.1007/s00261-019-01912-9>.
- [34] T. Irlbeck, J. Massaro, F. Bamberg, C. O'Donnell, U. Hoffmann, C. Fox, Association between single-slice measurements of visceral and abdominal subcutaneous adipose tissue with volumetric measurements: the Framingham Heart Study, *Int. J. Obes.* 2005 (34) (2010) 781–787, <https://doi.org/10.1038/ijo.2009.279>.
- [35] J. Wasserthal, H.-C. Breit, M.T. Meyer, M. Pradella, D. Hinck, A.W. Sauter, T. Heye, D.T. Boll, J. Cyriac, S. Yang, M. Bach, M. Segeroth, TotalSegmentator: Robust Segmentation of 104 Anatomic Structures in CT Images, *Radiol. Artif. Intell.* 5 (2023) e230024.
- [36] M. Ghassemi, L. Oakden-Rayner, A.L. Beam, The false hope of current approaches to explainable artificial intelligence in health care, *Lancet Digit. Health* 3 (2021) e745–e750, [https://doi.org/10.1016/S2589-7500\(21\)00208-9](https://doi.org/10.1016/S2589-7500(21)00208-9).

## 4. Discussion with references

### 4.1 Discussion and conclusion

This thesis shows three examples of automated quantitative image analysis using DL algorithms. The first addresses a targeted quantitative analysis, in which the uterine volume is assessed from T2-weighted MRI. To automatically measure uterine volume, a CNN based on the nnU-Net framework from Isensee et al. (2021) is developed to segment uterine tissue before and after HIFU therapy. This approach enables fast and reliable detection of volumetric changes in patients with uterine fibroids after HIFU intervention, allowing for an objective assessment of the therapy response. In clinical routine, a determination of the uterine volume has so far been either estimated based on diameter measurements or determined exactly by laborious manual segmentations. This study is thus an example of how a DL approach can enable efficient quantitative analysis in routine clinical practice.

The second work presents an end-to-end approach for an automated BCA. The demonstrated pipeline consists of two main steps: The slice extraction at the L3/L4 lumbar level and the segmentation of skeletal muscle, visceral adipose tissue, and subcutaneous adipose tissue, from which the body composition markers are derived. Due to the integrated quality control, this automatic approach enables the extraction of body composition metrics directly from CT scans in clinical routine. The use of these opportunistically collected quantitative markers is diverse, and the association of body composition parameters with various diseases and therapies is still under investigation. To date, a significant correlation between body composition and survival time has been demonstrated in patients with cardiogenic shock, advanced pancreatic cancer, and in patients with severe aortic stenosis undergoing TAVR (Luetkens et al., 2020; Nowak et al., 2024; Salam et al., 2023).

The fact that relevant information for the survival of TAVR patients can be obtained from an abdominal CT slice was also addressed in the last publication of this thesis. Here, a DL model directly predicts the patient's survival from pre-interventional abdominal CT. The model outperforms the baseline method, which uses different predefined body composition markers extracted from the abdominal CT slice. Thus, the study could demonstrate that further predictive features can be obtained from radiological image data.

The two presented papers on automated targeted and opportunistic quantitative analysis are mainly concerned with semantic segmentation tasks, where each voxel in the image is assigned to a certain label. In contrast, the last paper presents a direct predictive model that outputs a single numerical value representing a patient's mortality risk based on a high-dimensional radiological image. Such a challenge is therefore much more susceptible to overfitting problems. In addition to standard approaches, such as data augmentation and regularization of the loss function, the third paper therefore investigates the use of autoencoder pre-training as an additional technique to counteract overfitting. In fact, the transfer of pre-trained autoencoder weights, trained to learn appropriate data compression, improves the generalizability of the survival prediction.

As with the automated BCA described in this thesis, the direct image-based survival prediction is also based on 2D abdominal slice images. Unlike the DL model, which determines uterine volume from 3D MRI, the extracted quantitative image information for outcome analysis is restricted. However, especially in the case of BCA, several studies showed the prognostic value of these features extracted from a certain abdominal slice (Faron et al., 2021; Luetkens et al., 2020; Nowak et al., 2024; Prado et al., 2008, Salam et al., 2023). A recent study has also shown that important information for the prediction of survival time in TAVR patients can be obtained from the analysis of cardiac adipose tissue on single CT slices (Salam et al., 2024). This study measured the area and density of pericardial and epicardial adipose tissue (EAT) at the level of the aortic valve and found a significant association between EAT density and 2-year mortality. Only single-slice EAT measurements were performed, as a high correlation with the whole cardiac fat compartment has been previously demonstrated (Oyama et al., 2011; Vach et al., 2023). Regarding the presented DL approach for survival prediction in TAVR patients, where feature extraction is restricted to a single abdominal slice, an additional consideration of cardiac CT slices may be beneficial for this task. In general, the study by Salam et al. (2024) motivates a further extension of the automated opportunistic BCA described in this work to the additional assessment of cardiac adipose tissue. This opportunistic assessment would allow further investigation of the impact of the cardiac biomarker in future studies. Instead of restricting the BCA to certain CT slices, it is also possible to determine body composition markers in the entire 3D dataset. By now, there are also DL

pipelines that enable automated BCA based on 3D CT data (Haubold et al., 2024; Koitka et al., 2021). However, the impact of 2D vs 3D markers on clinical endpoints is still an open question. In a very specific cohort of cancer patients treated with anti-angiogenic therapy, 3D body composition features provided significant prognostic value for mortality prediction, while 2D features extracted at the L3 lumbar level did not (Decazes et al., 2023). On the other hand, certain tissues, such as EAT, are not robustly measurable over the entire volume, as a very high inter-reader variability has been reported (Commandeur et al., 2019; Greif et al. 2009). A 2D evaluation of EAT at the level of the aortic valve is much more reliable, as this specific slice allows for better delineation from the pericardium (Salam et al., 2024). However, a generalized comparison in different cohorts for the prognostic value of 2D and 3D quantitative image features remains a future task. Additionally, it must be considered that using 3D data for direct DL-based outcome prediction would further amplify the problem of overfitting.

In order to develop precise and valid prediction models, another important point must be taken into account, namely the integration of additional non-image-based parameters into the DL model. The third study in this thesis shows that direct DL-based survival prediction leads to an improvement compared to the sole use of image-based numerical markers. However, the study also reveals that image features alone are not sufficient to get an accurate prediction of survival. Considering age and sex as additional, non-image-based parameters already improves the performance, but the inclusion of further important clinical parameters and laboratory values has not yet been investigated in more detail. This would require multimodal DL models that combine the information extracted from the CT images with the individual parameters. In general, the handling and application of multimodal input data is one major challenge in ML (Castiglioni et al., 2021). When combining image data with clinical input parameters, different dimensionalities of both modalities must be taken into account and the information extracted from the low-dimensional clinical parameters has to be sufficiently considered. In the presented DL approach for survival prediction, age and sex are used as additional input for a CPHM, as suggested by Afshar et al. (2020). For the combination with image features, the image-based log-hazard risk predicted by the direct DL approach is used as an additional low-dimensional predictor variable. The high-dimensional input image is therefore first mapped



by the DL model to a numerical value representing the patient's risk. In future studies, so-called "Plugin Networks" may be investigated to directly integrate the numerical parameters into the DL model (Koperski et al., 2020). In this approach, the base CNN is first trained only on the image data. Then the additional knowledge, e.g. certain laboratory values, forms the input to a set of fully connected neural networks (plugin layers) attached to the backbone of the pre-trained base CNN. In a second training run, only the weights of these plugin layers are adapted. This effectively combines different modalities without a large increase in computational effort.

A general difficulty when using DL techniques is the lack of interpretability (Castiglioni et al., 2021, Hosny et al., 2018). In the case of the presented direct image-based survival prediction model, the extracted image features improve the mortality prediction, but it is not straightforward to determine exactly which part of the image is of particular importance. Interpretability can be significantly easier when using classic ML algorithms, such as decision trees (Castiglioni et al., 2021). However, efforts are also being made to increase the explainability of DL algorithms. A well-known method is the use of saliency maps aiming to localize the relevant image regions (Selvaraju et al., 2017). However, this method has some weaknesses, e.g. a rough localization is not sufficient to fully understand the network's decision-making process (Ghassemi et al., 2021). Rombach et al. (2020) also argued that interpreting DL models requires considering the model's learned invariances. Therefore, they have presented an approach to recover both the model's semantic concept and its learned invariances. Especially for multimodal models, interpretability is crucial to understand which parameters contributed to the algorithm's decision and to draw conclusions for clinical application.

In summary, this work has shown how DL models can be used for fast and robust automated quantitative image analysis. This applies to targeted quantitative analyses such as uterine volume measurement and opportunistic assessments such as end-to-end automated BCA. In addition, a method improving the survival prediction of TAVR patients through the development of a direct image-based DL model was presented. However, there are still open issues that need to be further investigated, e.g. the development of reliable multimodal DL models and the interpretability and explainability of the methods.

## 4.2 References

- Afshar P, Mohammadi A, Tyrrell PN, Cheung P, Sigiuk A, Plataniotis, KN, Nguyen ET, Oikonomou A. DRTOP: deep learning-based radiomics for the time-to-event outcome prediction in lung cancer. *Sci Rep* 2020; 10(1): 12366
- Castiglioni I, Rundo L, Codari M, Di Leo G, Salvatore C, Interlenghi M, Gallivanone F, Cozzi A, D'Amico NC, Sardanelli F. AI applications to medical images: From machine learning to deep learning. *Phys Med* 2021; 83: 9-24
- Commandeur F, Goeller M, Razipour A, Cadet S, Hell MM, Kwiecinski J, Chen X, Chang HJ, Marwan M, Achenbach S, Berman DS, Slomka PJ, Tamarappoo BK, Dey D. Fully automated CT quantification of epicardial adipose tissue by deep learning: a multicenter study. *Radiol Artif Intel* 2019; 1(6): e190045
- Decazes P, Ammari S, De Prévia A, Mottay L, Lawrance L, Belkouchi Y, Benatsou B, Albiges L, Balleyguier C, Vera P, Lassau N. Body composition to define prognosis of cancers treated by anti-angiogenic drugs. *Diagnostics* 2023; 13(2): 205
- Faron A, Opheys NS, Nowak S, Sprinkart AM, Isaak A, Theis M, Mesropyan N, Endler C, Sirokay J, Pieper CC, Kuetting D, Attenberger U, Landsberg J, Luetkens JA. Deep learning-based body composition analysis predicts outcome in melanoma patients treated with immune checkpoint inhibitors. *Diagnostics* 2021; 11(12): 2314
- Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. *Lancet Digit Health* 2021; 3(11): e745-e750
- Greif M, Becker A, von Ziegler F, Lebherz C, Lehrke M, Broedl UC, Tittus J, Parhofer K, Becker C, Reiser M, Knez A, Leber AW. Pericardial adipose tissue determined by dual source CT is a risk factor for coronary atherosclerosis. *Arterioscler Thromb Vasc Biol* 2009; 29(5): 781-786
- Haubold J, Baldini G, Parmar V, Schaarschmidt BM, Koitka S, Kroll L, van Landeghem N, Umutlu L, Forsting M, Nensa F, Hosch R. BOA: A CT-based body and organ analysis for radiologists at the point of care. *Invest Radiol* 2024; 59(6): 433-441

Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts HJWL. Artificial intelligence in radiology. *Nat Rev Cancer* 2018; 18(8): 500-510

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: 203–211

Koitka S, Kroll L, Malamutmann E, Oezcelik A, Nensa F. Fully automated body composition analysis in routine CT imaging using 3D semantic segmentation convolutional neural networks. *Eur Radiol* 2021; 31: 1795-1804

Koperski M, Konopczynski T, Nowak R, Semberecki P, Trzcinski T. Plugin networks for inference under partial evidence. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) 2020*; 2883-2891

Luetkens JA, Faron A, Geissler HL, Al-Kassou B, Shamekhi, J, Stundl A, Sprinkart AM, Meyer C, Fimmers R, Treede H, Grube E, Nickenig G, Sinnig JM, Thomas D. Opportunistic computed tomography imaging for the assessment of fatty muscle fraction predicts outcome in patients undergoing transcatheter aortic valve replacement. *Circulation* 2020; 141(3): 234-236

Nowak S, Kloth C, Theis M, Marinova M, Attenberger UI, Sprinkart AM, Luetkens JA. Deep learning-based assessment of CT markers of sarcopenia and myosteotosis for outcome assessment in patients with advanced pancreatic cancer after high-intensity focused ultrasound treatment. *Eur Radiol* 2024; 34(1): 279-286

Oyama N, Goto D, Ito YM, Ishimori N, Mimura R, Furumoto T, Kato F, Tsutsui H, Tamaki N, Terae S, Shirato H. Single-slice epicardial fat area measurement: do we need to measure the total epicardial fat volume?. *Jpn J Radiol* 2011; 29: 104-109

Prado CMM, Lieffers JR, McCargar LJ, Reiman T, Sawyer MB, Martin L, Baracos VE. Prevalence and clinical implications of sarcopenic obesity in patients with solid tumours of the respiratory and gastrointestinal tracts: a population-based study. *Lancet Oncol* 2008; 9(7): 629-635

Rombach R, Esser P, Ommer B. Making sense of CNNs: Interpreting deep representations and their invariances with INNs. In: Vedaldi A, Bischof H, Brox T, Frahm JM, Eds. *Computer Vision—ECCV 2020 Cham: Springer, 2020: 647-664*

Salam B, Al Zaidi M, Sprinkart AM, Nowak S, Theis M, Kuetting D, Aksoy A, Nickenig G, Attenberger U, Zimmer S, Luetkens JA. Opportunistic CT-derived analysis of fat and muscle tissue composition predicts mortality in patients with cardiogenic shock. *Sci Rep* 2023; 13(1): 22293

Salam B, Al-Kassou B, Weinhold L, Sprinkart AM, Nowak S, Theis M, Schmid M, Al Zaidi M, Weber M, Pieper CC, Kuetting D, Shamekhi J, Nickenig G, Attenberger U, Zimmer S, Luetkens JA. CT-derived Epicardial Adipose Tissue Inflammation Predicts Outcome in Patients Undergoing Transcatheter Aortic Valve Replacement. *J Thorac Imaging* 2024; 39(4): 224-231

Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In: *Proceedings of the IEEE international conference on computer vision (ICCV) 2017; 618-626*

Vach M, Luetkens JA, Faron A, Isaak A, Salam B, Thomas D, Attenberger UI, Sprinkart AM. Association between single-slice and whole heart measurements of epicardial and pericardial fat in cardiac MRI. *Acta Radiol* 2023; 64(7): 2229-2237

## 5. Acknowledgement

I want to take this opportunity to thank everyone who has supported me during my doctorate. Special thanks go to my doctoral committee: Above all, my supervisor Univ.-Prof. Dr. Ulrike Attenberger, who enabled this doctorate and has supported me over the last few years. Many thanks to Prof. Dr. rer. nat. Jürgen Hesser, who always helped me with his expertise during our discussions and thus steered this dissertation in the right direction. I would also like to thank Univ.-Prof. Dr. med. Ass. jur. Alexander Radbruch, and finally Priv.-Doz. Dr.-Ing. Alois Martin Sprinkart: Martin, you certainly played the biggest part in enabling me to submit this thesis. Thank you very much for your support over the last few years, your honest and constructive feedback and for always giving me helpful advice, regardless of the issue.

I would also like to thank my colleagues Sebastian Nowak and Laura Garajová for their assistance during this time. Sebastian, thank you for always advising me on all my projects and enriching me with your expertise! I have always enjoyed working with both of you and I hope that we will carry out many more projects together. Many thanks also to Priv.-Doz. Dr. rer. nat. Wolfgang Block, who was always available as a contact person and supported me in many of my projects, as well as all other colleagues and co-authors who contributed to this dissertation.

Finally, I would like to thank my family: my mom for her love, her open ear in every situation, and her constant belief in me; my dad, who has supported me throughout my life and motivated my scientific career; and my grandparents, who have always been there for me. Special thanks go to my siblings: my brother, who was at least as excited about my published papers as I was, and my sister, who has always been my biggest support and refuge. My deepest gratitude goes to my boyfriend Marvin: Thank you so much for your love, your trust, your motivation and your advice!