Aus der
Klinik für Diagnostische und Interventionelle Radiologie
Universitätsklinikum Bonn
Direktor: Univ.-Prof. Dr. med. Julian Alexander Luetkens

Advancing radiological workflows through AI: Deep learning for automated tissue quantification, disease classification, generating synthetic contrast imaging and free-text report content extraction

Habilitationsschrift
zur Erlangung der Venia Legendi
der Hohen Medizinischen Fakultät
der Rheinischen Friedrich-Wilhelms-Universität Bonn
für das Fachgebiet
„Experimentelle Radiologie“
mit Prüfungsdatum
04.11.2025

Vorgelegt von
Dr. Sebastian Nowak
Wissenschaftlicher Mitarbeiter der
Rheinischen Friedrich-Wilhelms-Universität Bonn
Bonn 2026

Meiner Familie

gewidmet

The following publications are included in this cumulative habilitation thesis

Advances in artificial intelligence (AI) algorithms have raised expectations for the transformation of knowledge-based workflows, also in radiology. In this thesis, the applicability of AI to automate, optimize or support various image- or report-based analysis was investigated. It was the aim to provide insights into the potential of AI for advancing radiological workflows and thereby improving patient care. The scope of the eight original works included in this cumulative thesis can be categorized into three main topics:

Processing of free-text radiological reports

1. **Privacy-ensuring, open-weights large language models are competitive with closed GPT-4o in extracting chest X-ray findings from free-text reports.** Nowak S, Wulff B, Layer YC, Theis M, Isaak A, Salam B, Block W, Kuetting D, Pieper CCC, Luetkens JA, Attenberger UI, Sprinkart AM. Radiology. Radiology 2025; 314(1):e240895.
2. **Transformer-based structuring of free-text radiology report databases.** Nowak S*, Biesner D*, Layer Y, Theis M, Schneider H, Block W, Wulff B, Attenberger UI*, Sifa R*, Sprinkart AM*. European Radiology. 2023;33(6):4228–4236.
3. **Development of image-based decision support systems utilizing information extracted from radiological free-text report databases with text-based transformers**. Nowak S*, Schneider H*, Layer YC, Theis M, Biesner D, Block W, Wulff B, Attenberger UI, Sifa R*, Sprinkart AM*. European Radiology. 2024;34(5):2895-2904.

Generating synthetic radiological images

4. **Deep learning virtual contrast-enhanced T1 mapping for contrast-free myocardial extracellular volume assessment. Journal of the American Heart Association**. Nowak S*, Bischoff LM*, Pennig L, Kaya K, Isaak A, Theis M, Block W, Pieper CC, Kuetting D, Zimmer S, Nickenig G, Attenberger UI, Sprinkart AM*, Luetkens JA*. Journal of the American Heart Association. 2024;13(19):e035599.

Supporting diagnostic and treatment decisions based on radiological imaging

5. **Deep learning supports the differentiation of alcoholic and other-than-alcoholic cirrhosis based on MRI.** Luetkens JA*, Nowak S*, Mesropyan N, Block W, Praktiknjo M, Chang J, Bauckhage C, Sifa R, Sprinkart AM*, Faron A*, Attenberger UI*. Scientific reports. 2022;12(1):8297.
6. **Deep learning–based assessment of CT markers of sarcopenia and myosteatosis for outcome assessment in patients with advanced pancreatic cancer after high-intensity focused ultrasound treatment.** Nowak S*, Kloth C*, Theis M, Marinova M, Attenberger UI, Sprinkart AM*, Luetkens JA*. European Radiology. 2024;34(1):279–286.
7. **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*. European Journal of Radiology. 2023;168:111150.
8. **Computer tomography-based assessment of perivascular adipose tissue in patients with abdominal aortic aneurysms.** Ginzburg D*, Nowak S*, Attenberger U, Luetkens J, Sprinkart AM*, Kuetting D*. *Scientific Reports* 2024;*14*(1):20512.

* contributed equally

1 Table of contents

List of abbreviations

AI	Artificial intelligence
ANN	Artificial neural network
CNN	Convolutional neural networks
ChatGPT	Chatting generative pre-trained transformer
AGI	Artificial general intelligence
UTF-8	8-Bit UCS transformation format
Swin	Hierarchical vision transformers using shifted windows
BERT	Bidirectional encoder representations from transformers
GPT	Generative pre-trained transformer
LLM	Large language model
RLHF	Reinforcement learning from human feedback
GPU	Graphics processing unit
GAN	Generative adversarial network
CLIP	Contrastive language-image pre-training
CE	Contrast-enhanced
MRI	Magnetic resonance imaging
ECV	Extracellular volume
CT	Computed tomography
CPH	Cox proportional hazard
RAG	Retrieval augmented generation
GBCA	Gadolinium-based contrast agent
LGE	Late gadolinium enhancement
BMI	Body-mass-index
ECOG	Eastern cooperative oncology group score

2 Introduction

2.1 Introduction into artificial intelligence

In this first part of the thesis, a brief introduction into the milestones of artificial intelligence (AI) is given, which form the basis for the set of methods used in the works included in this thesis.

2.1.1 Perceptron and the winter of artificial intelligence

In 1958 the American psychologist Frank Rosenblatt introduced the perceptron, an electric device that aimed to model the biological functioning of a neuron, i.e. an artificial neuron (van der Malsburg 1986). This invention of a learning electrical system is considered as one of the earliest ideas on AI and that mark the beginning of the developments leading to the modern AI of today (LeCun et al. 2015; Tappert 2019). Figure 1 illustrates the function of a perceptron.

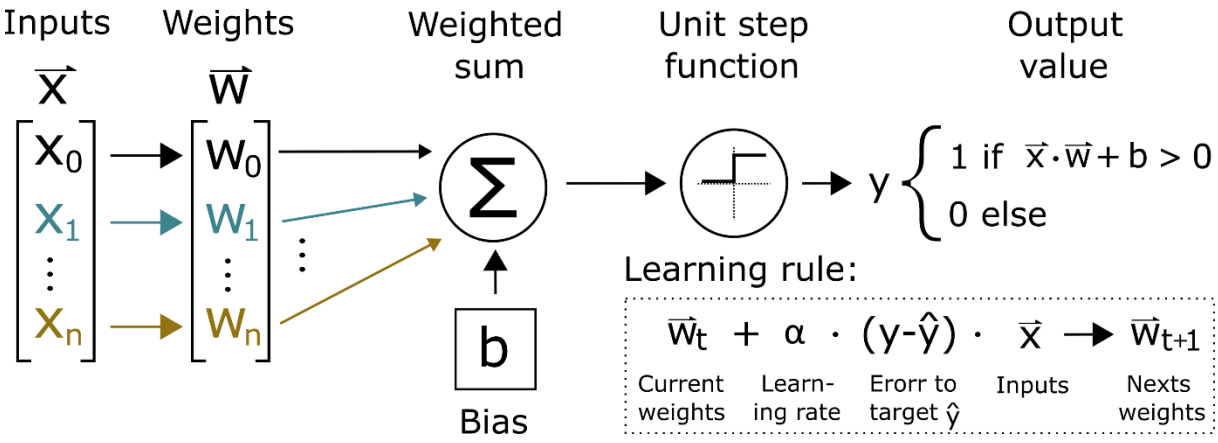


Figure 1: Schematic representation of a perceptron. The perceptron calculates a weighted sum of its input values, which can also be represented as a dot product between an input vector ($\vec{x}$) and a weight vector ($\vec{w}$). In addition, an estimated threshold (bias) is also added. The output value of the perceptron is binarized by a unit step function. The weights defining the summation of inputs are the learnable parts of the perceptron and are usually initialized randomly at the beginning. The perceptron can learn to perform a linearly separable classification by comparing the error of its outputs with the desired output (target) and changing the weights based on this error.

Rosenblatt has shown that a single perceptron with two input values and a single output value could learn linear separable operations such as the logical AND, OR, and NOT operations. However, following work that mathematically investigated the perceptron in 1969 reported limited abilities and showed that even basic functions such as the XOR operation could not be achieved by a single layered perceptron (Minsky and Papert 1988). Due to such negative reporting on perceptrons and due to missing understanding on how to efficiently train these systems, research on neural networks as learning algorithm declined, which is commonly referred to as the winter of AI.

2.1.2 Artificial neural networks and deep learning: Breakthrough in training multi-layered perceptrons

Two key aspects led to the revival of research into neural networks in the 1980s. First, it was shown that the limited capability of perceptrons to represent non-linear separable functions can be overcome by building them into multiple sequential layers, where the input of a perceptron of a given layer is the output of all perceptrons of the previous layer. In addition, different activation functions other than the unit step function were used in their intermediate "hidden" layers. Such a multilayered system, which is illustrated in Figure 2, is also referred to as fully connected artificial neural network (ANN) today. However, the definition of the learning rule, i.e. the algorithm by which the weights of the network are adapted to minimize the error of model prediction to target, is not as straight forward as the learning rule of the single layered perceptron shown in Figure 1.

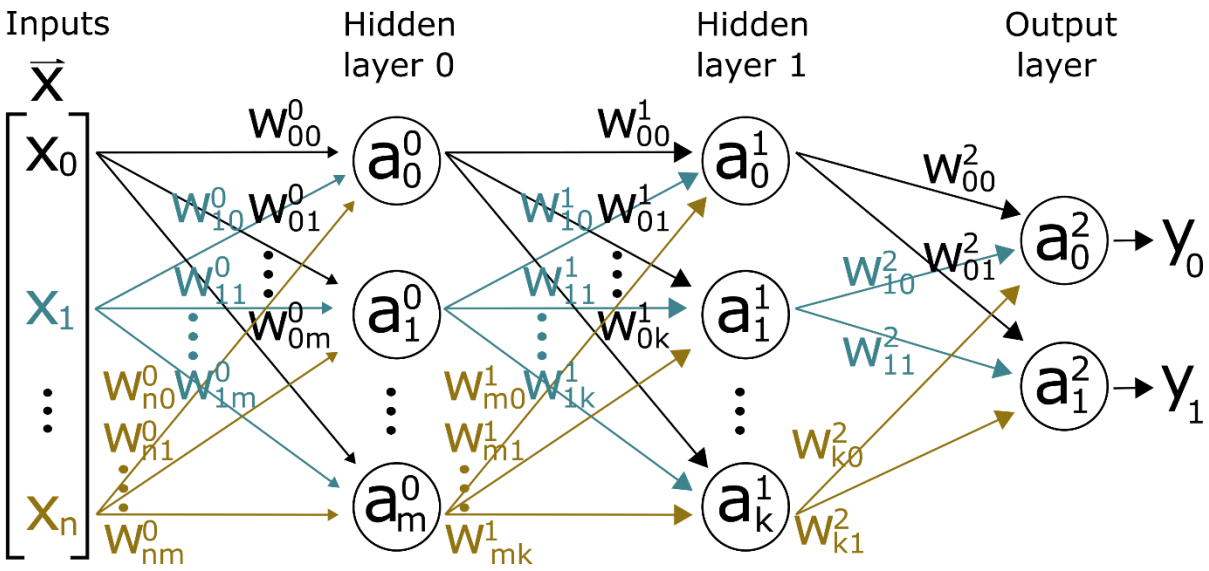


Figure 2: Schematic representation of a fully connected artificial neural network with three layers: two hidden layers, and one output layer. Each input value (x_0 to x_n) is connected to the m+1 nodes (a^0_0 to a^0_m) of the first hidden layer. Each output value of the first hidden layer is input of the k+1 nodes (a^1_0 to a^1_k) of the second hidden layer, which outputs are then finally the inputs for the two nodes (a^2_0 and a^2_1) of the output layer. A node represents an artificial neuron that does weighted summation of its inputs with subsequent application of an activation function, as shown in Figure 1. Connections between nodes are represented by arrows, with weights denoted as w^i_{jg}, where i is the layer number, j is the index of the neuron a^i_j within layer i and g is the index of the neuron in the next layer a^{i+1}_g to which a^i_j is connected by the weight w^i_{jg}. For simplicity, biases are not illustrated. The illustrated network could be used for classification of two binary classes or for regression of two continuous values.

The second aspect that led to increased interest in neural networks is the backpropagation algorithm that was popularized by David E. Rumelhart, Ronald J. Williams, and Geoffrey Hinton that received the physics Nobel prize for his achievements in AI research in 2024. This new learning algorithm enabled the calculation of how the error between model prediction and target backpropagates into the numerous nodes of the multilayered networks using principles of differential calculus (Rumelhart et al. 1986). This enabled efficient training of ANNs by calculation of the gradient, i.e. the direction in which the weight vector must be updated for model improvement. This new approach for machine learning by backpropagating errors deeply into multilayered neural networks was basis for the term “deep learning” (Dechter 1986).

2.1.3 Convolutional neural networks: Specializing on images and other grid-based data

Computer vision, i.e. the application of computer algorithms for analyzing images, including the task of recognizing objects within images, was an important field of research in the 1980s and is still an important field today. Classical approaches used human-defined filters or heuristic rules to extract local features from the image, such as the presence of edges in different directions, and further human defined rules to combine them into a higher form of features (LeCun 1989). Inspired by breakthroughs of neural networks training using backpropagation, LeCun et al. investigated the detection of handwritten digits in small 16x16 pixel sized images by flattening the images into 256-element vectors. These were then input for fully-connected ANN trained with backpropagation directly on the pixel data (LeCun et al. 1989). This work represents an early example of the application of ANNs that shows promising results for object recognition.

However, the task of recognizing objects in images is inherently invariant to scaling, shifting, rotating, or slight distortions of the object within the image (LeCun 1989). For instance, scaling or shifting a digit does not alter its identity. With the described approach, each pixel in the 16x16 grid is assigned to a fixed input position of an ANN, and consequently to fixed learnable weights. Reflecting this spatial invariance in the model design would be preferable as, for example, shifting a digit into a different image section does not change its underlying features, such as the orientation and relation of the digit’s edges. Therefore, weights should be shared or re-used across different image sections. Another fact that should be utilized for efficiency is that images represent a form of grid-based data, where each pixel occupies a specific location within the grid and maintains spatial relationships with neighboring pixels. Therefore, pixels with lower distance to each other are more important for a combined processing compared to pixels with high distance. This inherent aspect of processing visual information was also found in the functioning of the visual cortex in neuroscience. Here, it could be shown that single neurons respond to stimuli in specific areas of the retina, called receptive fields (Hubel and Wiesel 1962). Furthermore, visual neurons have a hierarchical organization, where simple cells respond to basic features, like edges, and higher level neurons combine inputs to detect more complex patterns (D’Souza et al. 2022).

Consequently, later works on symbol and digit recognition in documents popularized convolutional neural networks (CNN), a architecture type that mimics the functioning of biological visual systems more closely (LeCun et al. 1989; LeCun et al. 1998). In the following years, researches could achieve impressive results with CNNs on the large-scale, real-world ImageNet classification challenge including over 1 million 224x244 sized images and a thousand different classes, such as different animals or vehicles (Krizhevsky et al. 2012). This achievement marked the beginning of CNN's status as a state-of-the-art approach for image processing.

The architectural design of CNNs is strongly adapted to grid-based data, as images, and achieves spatially invariant image processing, efficient reuse of weights, and independence of the number of its model parameters from the image size. The following sections briefly describe the central concepts and applications of CNNs.

Applying weights as convolutional operation: In a convolutional operation a small matrix (termed kernel, for two-dimensions commonly sized 3x3 or 5x5) slides across the image matrix. For each image coordinate an element-wise multiplication of the kernel with the overlapping image region is performed, followed by summation to produce a single output value. This value is written into the corresponding position of a new matrix. In convolutional layers the kernel values represent the feature extracting, learnable weights of the neural network. The new matrix is therefore termed feature map. A convolutional layer commonly consists of numerous kernels that are applied in parallel producing multiple feature maps. CNNs consists of an input layer that processes the values of the input image and multiple subsequent hidden convolutional layers that process the intermediate feature representations encoded in the feature maps.

Receptive field enlargement and spatial subsampling: A single convolutional layer using 3x3 kernels combines input image information from a 3x3 pixel area, i.e. has a receptive field of 3x3. After applying another convolutional layer with 3x3 kernels to the first feature maps, the values of the next feature maps have a receptive field size of 5x5 with respect to the input image. However, objects that are content of the image will mostly be part of a large section or even the entire image. To achieve a receptive field close to the size of an ImageNet image, i.e. 224x224 pixels, would require the cumulative effect of over 100 convolutional layers. Therefore, the second central concept of CNNs is the application of down-sampling operations to feature maps, also termed pooling. In the most common pooling method, the so-called max pooling, a 2x2 kernel slides over the feature maps while skipping every second matrix position, which is termed stride of two. The maximum value in the overlapping region is written into a new down-sampled feature map that is half the size, i.e. the max pooling kernel has no learnable weights. By this convolutional operation the next feature maps and convolutional layers have drastically increased receptive field size with respect to the input image. Some more recent CNN implementations use convolutions with stride two and 2x2 kernels with learnable weights for spatial subsampling instead of max pooling kernels. Figure 3 illustrates these convolution and pooling operations.

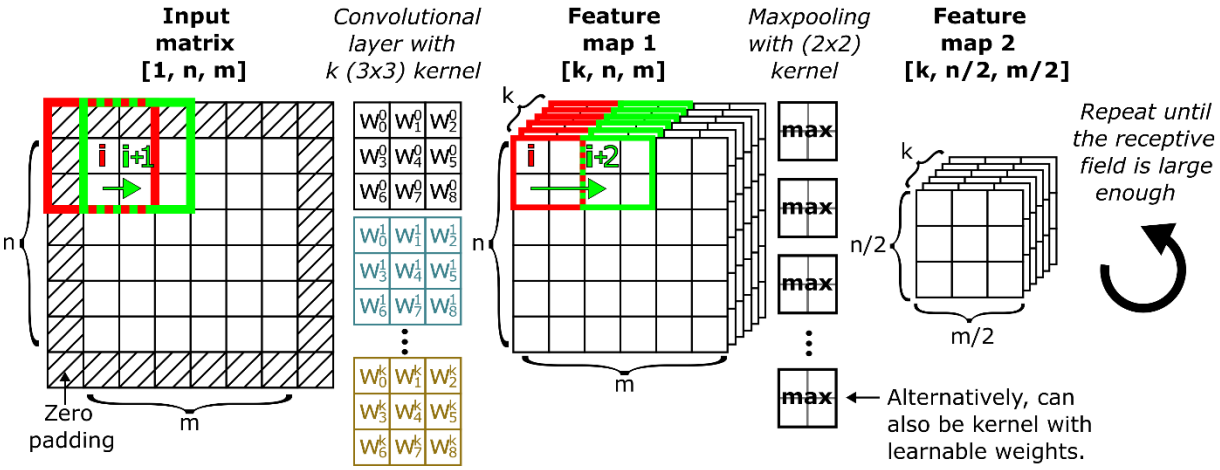


Figure 3: Schematic illustration of a convolutional and a maxpooling layer. In the convolutional layer, element-wise multiplication of k 3x3 sized convolutional kernels is performed for each coordinate i of the input matrix (stride of one). The kernel values represent the learnable weights. For simplicity bias and activation functions are not illustrated. To remain matrix size after convolution, a small border is added to the input matrix prior to convolution, which is termed padding. Therefore, the resulting feature map is of same spatial dimension [n, m], but now has k channels created by convolution with k kernels. Spatial resampling by maxpooling can be applied to increase the receptive field of subsequent feature maps. Here, a 2x2 sized kernel extracts the maximum value in the overlapping region for every second coordinate i of the feature map (stride of two), resulting in a new feature map with half spatial dimension [n/2, m/2]. This process of repeating multiple convolutional layers with subsequent spatial resampling can be repeated until the receptive field has an appropriate size for detecting objects within the image.

Classification based on low-dimensional feature maps: A CNN for image classification commonly flattens the last low dimensional feature map that encodes combinations of feature representations into a single dimensional vector. This vector can then be input for hidden, fully connected ANN layers that end in an output layer with a number of neurons representing the number of classification classes. The application of a softmax activation function results in the output values representing classification probabilities. A system that encodes higher into lower level feature representations is commonly referred to as “encoder”.
Image segmentation by feature map up-sampling: Segmentation of structures within an image, i.e. classifying each pixel of the input image into different object classes, can be achieved by a CNN that also features a so called “decoder”. In a decoder convolutional layers and up-sampling operations are alternated to resize low dimensional feature map of the encoder until the original image size is reached. Then an output convolutional layer with number of kernels equal to the number of segmentation classes and with softmax activation allows for pixel-wise image classification. A prominent architecture for image segmentation is the U-net, that is named based on the typical shape when illustrating a down-sampling encoder with up-sampling decoder (Ronneberger et al. 2015). To enable stable training, the U-net also introduces so called skip-connections, which are illustrated in Figure 4.

Image regression: In principle, the activation function of the output layer of a CNN is crucial for the use case the model is applied on. While in classification and segmentation the softmax function is used to create discrete outputs, representing classification classes, a sigmoid activation function can map the output to continues values and thereby achieve a regression task. Examples are the prediction of an image coordinate or the patient's hazard (Nowak et al. 2023; Kim et al. 2020).

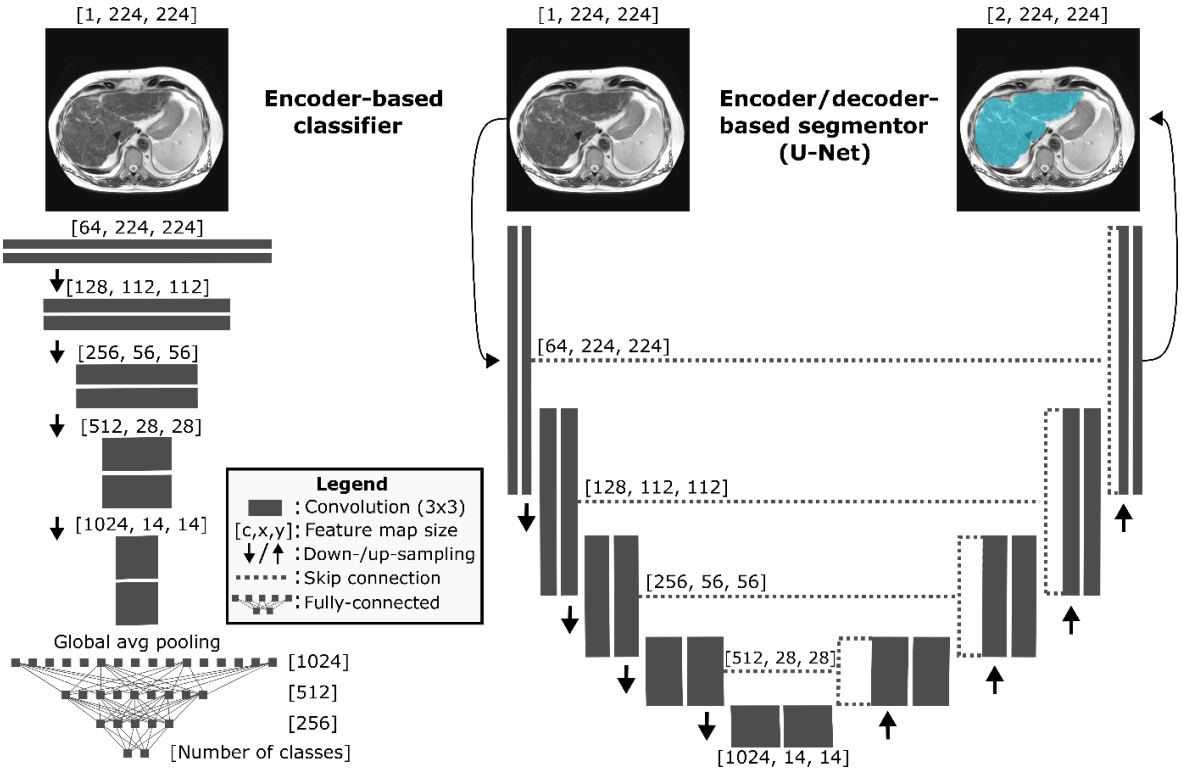


Figure 4: Schematic representation of an encoder-based CNN for classification (left) and an encoder-decoder based U-Net architecture for segmentation (right). Both share a common encoder structure designed to encode conclusive image feature representations into low-dimensional feature maps by optimizing convolutional kernel weights. The encoder employs spatial subsampling to expand the receptive field, i.e. the area within the input image from which information is combined into a single value in the final feature maps. Fully connected layers are frequently utilized as classification heads atop the final, lowest-dimensional feature maps to generate binary classification output. The image segmentation task requires pixel-wise classification of the input image. This is achieved by combining an encoder with a subsequent decoder that up-samples the feature maps to original image dimension. To maintain stable gradients after backpropagation the U-Net introduces skip-connections, where feature maps of the encoder are attached to the feature maps of the decoder.

Leveraging the convolutional operation for definition of the structure of weight connections within the network leads to efficient processing. This includes re-using of the learnable weights for different image sections and enables translation invariance i.e. the processing of features regardless of their position. Furthermore, by primarily combining neighboring pixels

for processing and by hierarchical ordering of neurons and pooling operations to increase complexity of features and receptive field size, a CNN follows more closely the functioning of a biological visual cortex compared to a ANN (Zeiler and Fergus 2014). Figure 4 shows a schematic representation of an encoder-based CNN for classification and an encoder-decoder-based U-net for segmentation.

2.1.4 Transformers: The self-attention mechanism revolutionized artificial intelligence

The release of the chatting generative pre-trained transformer (ChatGPT) by OpenAI in late 2023 brought the topic of artificial general intelligence (AGI) to the forefront of public discourse. The demonstration of a learnable systems that is capable projecting its capabilities on numerous tasks represents a significant advancement compared to the task-specific and tool-based AI methods also described in previous sections and, thereby, resulted in a massive increase in interest and investments in AI technologies. These recent developments are ultimately triggered by the "Attention is all you need" publication and the subsequent rise of the transformer architecture (Vaswani et al. 2017). The following section and the visual explanations of Figure 5 aim to briefly introduce basics of natural language processing, and the transformer architecture as published in 2017.

Tokenization

This represents an efficient formatting of the input text into numerical format. Transformers, and other natural language processing algorithms, analyze the input text by mathematical operations, such as matrix multiplications. In general, as computers fundamentally function with binary coded data, i.e. bits and bytes, the characters of a text are already stored as numerical values. For example, the 8-Bit UCS transformation format (UTF-8) describes how characters and symbols can be mapped to numerical values that can be stored and processed by a computer. However, using a character-level tokenization of text would create numerical sequences of maximum length, which is inefficient for the subsequent mathematical processing. Therefore, a tokenizer aims to generate shorter numerical sequences by introducing a fixed vocabulary with 30,000 or more entries. The vocabulary includes mappings of common words, sub words and lastly also single characters and symbols to individual numerical values, called tokens. The tokenizer used by the base model of ChatGPT would create the following text subdivisions indicated by “|”: |This| text| is| token|ized| into| |10| numerical| values| and convert them into following numerical values: 2028, 1495, 374, 4037, 1534, 1139, 220, 605, 35876, 2819.

Token embeddings

After tokenization, the first learnable layer of a transformer network (named embedding layer) maps the individual numerical tokens into a vector within a higher dimensional space of e.g. 768 dimensions. The mapping tokens to embedding vectors enables a trained transformer to encode the meaning of individual words or tokens, which should result in the embeddings of similar words being closer to each other in this higher dimensional space (Gong et al. 2018). Additionally, directions within the embedding space of a trained transformer should also encode the relation of words to each other. For example, the distance between the words “Berlin” and “Washington” should be lower than between “Berlin” and “dog”, or

adding the distance between the embeddings between “king” and “man” to “woman” should result in a point close to “queen” in the embedding space.

Positional encoding of embeddings

The learnable mapping of numerical tokens to embedding vectors and the later described self-attention mechanism of a transformer are not inherently able to represent ordering and positional distance of tokens. However, the position of a token or a word within the sequence can be important for its meaning and should therefore be represented. Therefore, in positional encoding the embedded tokens are infused with the relative and absolute positions of the tokens in the sequence by adding values of sine and cosine functions to the embedding vectors, which are dependent of the tokens position in the sequence.

The self-attention mechanism

This is the core innovation of the transformer architecture, enabling the model to weight the importance of different tokens or intermediate token representation of a sequence relative to each other. In self-attention layers, the processing and forwarding of values of the input sequence is dynamically influenced by the values itself. This contrasts with other traditional neural network architectures, such as ANNs or CNNs, where the process of multiplying the inputs of a fully connected or a convolutional layer with fixed numerical weights does not depend on the value of the input that is multiplied. Also, in contrast to CNNs, a classical transformer equally considers the entire input sequence simultaneously when generating an output sequence of token representations. This, however, results in a quadratic relationship of model complexity to input sequence lengths of a classical transformer. This makes self-attention computational intensive compared to e.g. a CNN that has linear relationship of model complexity to input size, which is highly beneficial in processing large images (Vaswani et al. 2017). Of note, in later developments it was shown that so called vision transformers can process images by converting 16x16 pixel patches into embedding space vectors, comparable to text-based tokens. Also, with similar ideas of the local windowed processing of CNNs, hierarchical vision transformers using shifted windows (Swin) have also linear relationship of model complexity to input length (Dosovitskiy et al. 2020; Liu et al. 2021).

Encoders, decoders and the effectiveness of scaling training data and compute

The transformer presented by Vaswani et al. in 2017 is designed for the task of translation between two languages. Essentially, this task consists of two sub tasks. First, comprehension of the meaning of the input text to be translated and generating the tokens or words of the translated sentence as output. Thereby, the presented transformers consisted of an encoder that encodes the meaning of the input text into a sequence of intermediate token representations. These representations (keys and values, see Figure 5) are fed into an intermediate state of the decoder, for which the sequence of already translated tokens is input. The decoder outputs a probability for the next token in the translated sequence and can therefore be used to iteratively generate a full translated text. From an implementation perspective, a decoder differs from an encoder only by the masking of its attention matrix, resulting in token representations only being able to attend to previous token representations. In a later work, researches from Google proposed using a solely encoder-based transformer named bidirectional encoder representations from transformers (BERT) for

the task of classifying the content of text into different binary classes, without generation of text (Devlin et al. 2018). The authors showed that transformers can be pre-trained on large amounts of unannotated text from the internet without prior human processing by requiring the model to fill out randomly blacked out tokens (named masked language modeling). The insight that transformers can learn general knowledge from unannotated texts was essential for further investigation of its potential to replace task-specific AI architectures. Later work from researchers of OpenAI proposed generative pre-trained transformers (GPT). This model essentially consists of a decoder-based transformer that is pre-trained to predict each token of a given sequence of unannotated text based on all prior tokens (termed causal language modeling). The authors could show that their third generation model (GPT-3) demonstrates great potential for text generation when using almost a trillion words of unlabeled internet text as training data and scaling of trainable parameters to 175 billion with massive computational hardware (Brown et al. 2020). Such large language models (LLMs) showed great abilities to compress extensive general knowledge and language understanding, LLMs were expected as basis for overcoming the task-specific nature of previous AI and thereby create AGI. However, OpenAI's researchers also reported that giving text-based definitions of user-defined tasks in a chatbot scenario (also termed prompting) to models solely pre-trained by causal language modelling, can result in invention of false facts (often termed hallucinations), the generation of biased texts, or simply not following user instructions. Therefore the authors proposed a multi-step post-training named reinforcement learning from human feedback (RLHF) (Ouyang et al. 2022). Here, the model is first trained on human-generated examples of instruction-response pairs to infuse general "willingness" to answer given questions. In a next step, a reward model is trained with user ratings, aiming to distinguish good from bad answers. Lastly, the answer quality is then improved by adapting the text-generating model to generate answers that the reward model considers good. This was the basis for the release of ChatGPT by OpenAI at the end of 2023.

Open- versus closed weights LLMs: OpenAI and other prominent closed-model providers, such as Anthropic, Google or X, do not publish their frontier models for free use under an open-weights license (Nowak and Sprinkart 2024). This is due to the extensive financial investments required to obtain a sufficient amount of compute by graphics processing units (GPU) for training LLMs and fulfillment of commercial interests. Also, the implementation of these closed LLMs on local hardware is not possible to date. To use these closed LLMs, data must be transferred to external servers via dedicated application user interfaces. In many countries, such as countries of the European Union, processing protected health information on servers outside the secure clinical infrastructure is highly regulated by strict data protection laws. To democratize the application of LLMs and to avoid privacy issues of closed LLMs there are initiatives by communities, but also companies, that release their models under open licenses. This allows model implementation and even further training on local hardware. This is complemented by algorithmic contributions that lower hardware requirements for adapting pre-trained LLMs, such as low rank-adaptation of frozen model weights or full training by low-rank projection of the gradients (Hu et al. 2021; Zhao et al. 2024a). Figure 6 shows a list of the top 50 LLMs according to a public leaderboard.

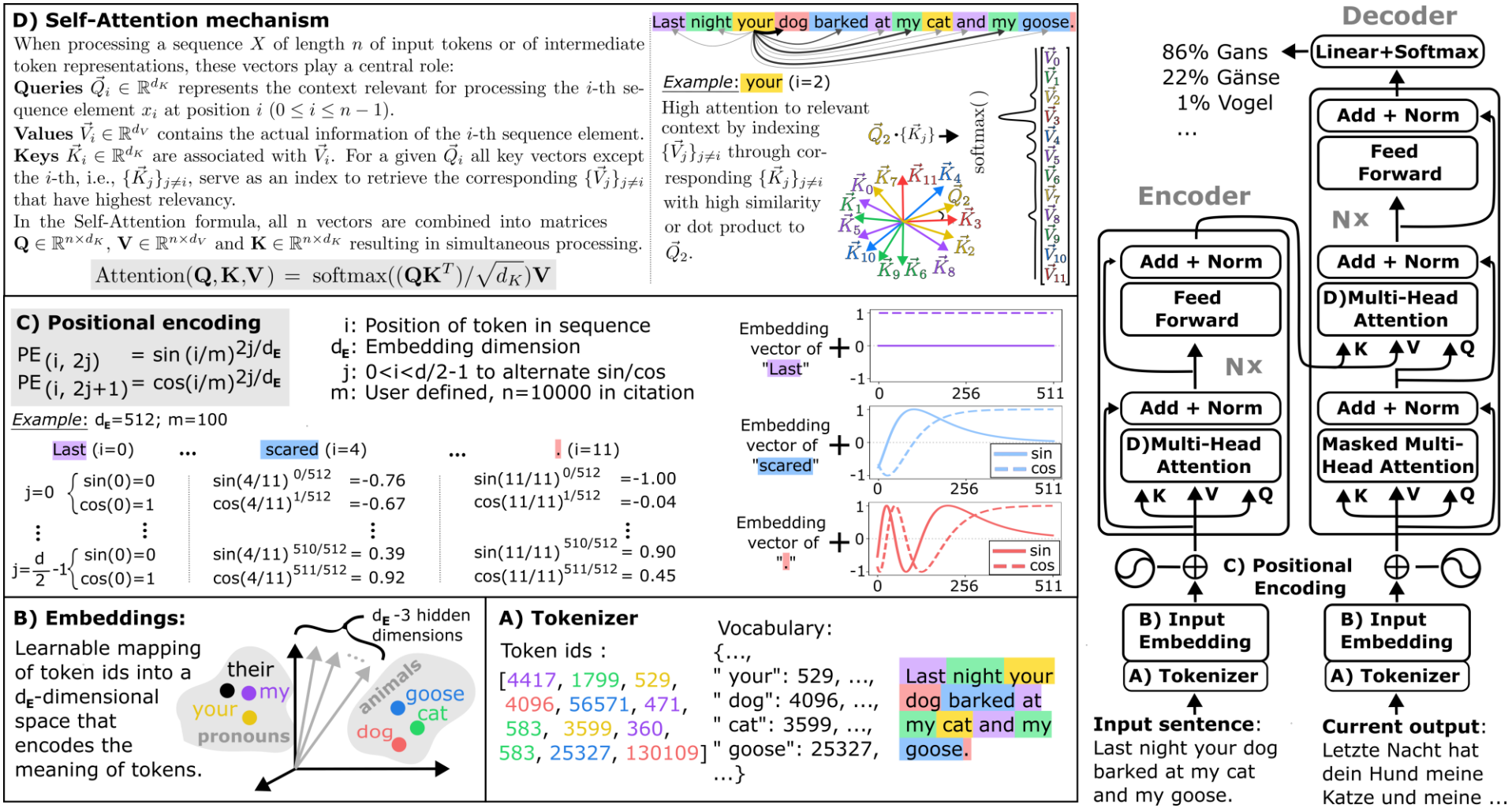


Figure 5: (left) Visualization of the core mechanisms of a transformer: A) tokenization, B) token embeddings, C) positional encoding and D) self-attention mechanism. (Right) An encoder-decoder based transformer, as proposed in the "Attention is all you need" publication, consisting of multiple (N) encoder and decoder layers that utilize multi-head attention blocks for machine translation (Vaswani et al. 2017).

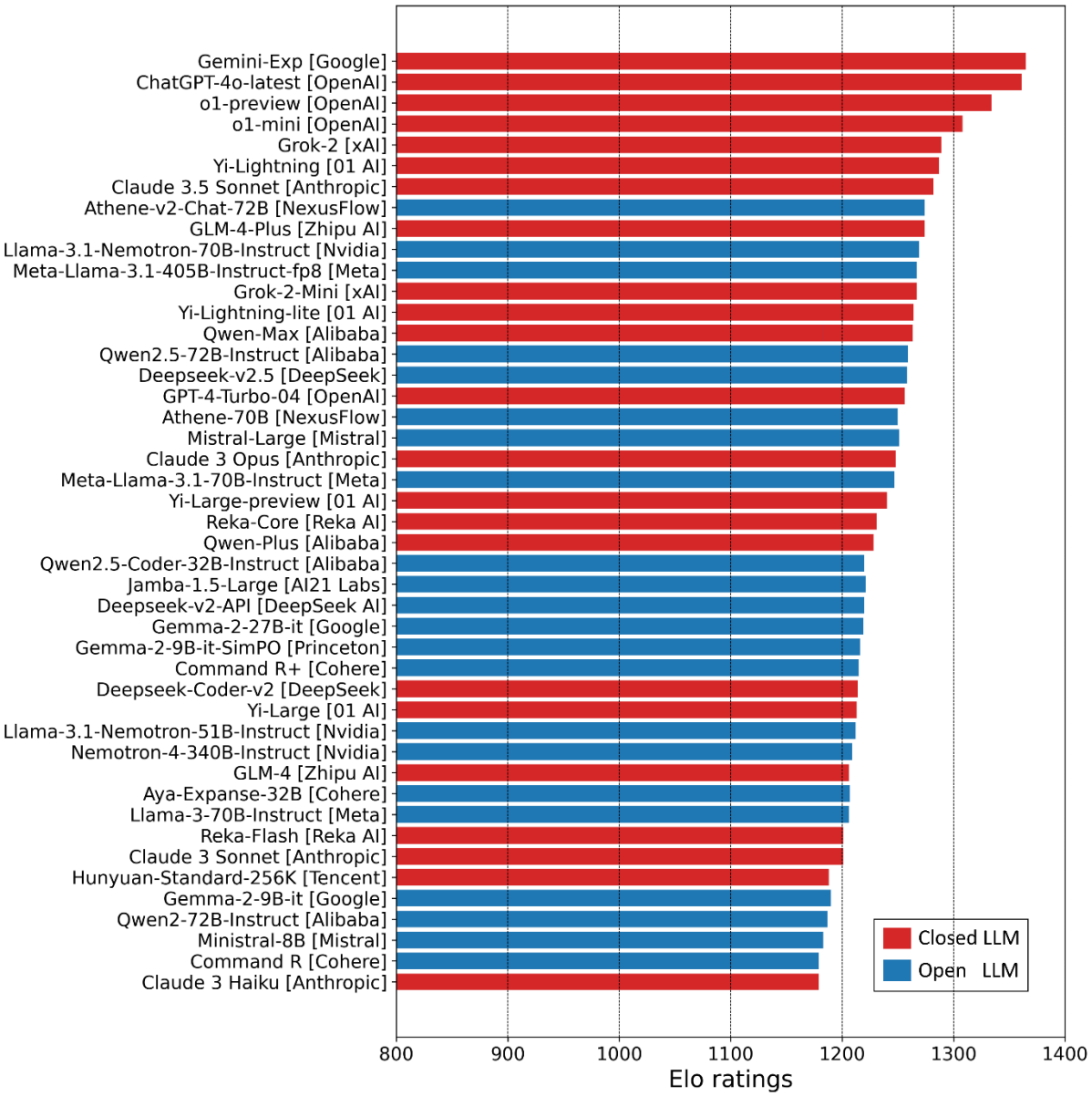


Figure 6: Elo rating of the top 50 closed and open LLMs according to a public leaderboard created by the University Berkeley, California (Chiang et al. 2024). For a given user question, two models compete for the best answer based on user preference. Similar to professional chess games, the models can gain or lose rating during these matches based on over a million community votes. Data were derived from https://lmarena.ai/ and https://github.com/fboulnois/llm-leaderboard-csv and accessed at 11/28/2024.

2.1.5 The different learning paradigms of deep learning

The previous illustrations show use-cases, where deep learning can be trained based on training data with the desired output being already defined, i.e. based on data with ground truth. However, this learning paradigm (termed supervised learning) is only one of many, each tailored to specific use cases and to the inherent properties of the data on which the model is trained. In the following section the diverse learning paradigms of deep learning are explained.

- Supervised learning describes the learning paradigm, in which a set of paired inputs with desired outputs are used as examples for the AI model. Given a specific input, the learning algorithm can calculate a target-based loss, e.g. the mean-squared error between the model's prediction and the desired output. Examples for supervised learning are the image classification or segmentation use-cases described in the previous sections.
- Unsupervised learning operates in scenarios where no human-defined outputs or other sources of ground truth are available. This means that the model cannot be optimized based on a prediction to the target error, as done in supervised learning. Therefore, in unsupervised learning the error function has to be designed in such a way that the model can learn useful skills based on the inherent structure of the unlabeled data itself. A prominent example of an unsupervised learning paradigm can be found in generative adversarial networks (GAN), which can be trained on unlabeled sets of images to generate artificial images. This is achieved by a user-defined model adversarial set up, in which an image generating model and a fake image detecting model compete (Goodfellow et al. 2014).
- Reinforcement learning describes the setting, in which an AI model (commonly referred to agent) interacts with environment states by taking actions (Dominic et al. 1991). Subsequently, the model receives rewards for positive actions on the environment. The probability of an action at a given state is described by the agent's policy. The goal is to tackle sequential decision-making by learning a policy that maximizes cumulative rewards over time. A very prominent example is the AlphaGo algorithm created by Google Deepmind that mastered the ancient game of Go, which was previously considered to have to many board states to be mastered by AI (Silver et al. 2016).
- Self-supervised learning algorithms have become increasingly important in recent years. This learning paradigm can be seen as the intersection between supervised and unsupervised learning. Both unsupervised and self-supervised learning share the absence of human labeled data on the ultimate learning task to be achieved. However, self-supervised learning often utilizes other inherent structures of the data in a pre-training stage that are often human-defined such as text-based image descriptions. These human-defined structures are leveraged as labels for supervised training to gain general insights about the data. Then, in a post training stage, the skills of the pre-trained models are adapted to achieve the main goal. Prominent examples for self-supervised learning algorithms are the contrastive language-image pre-training (CLIP) that utilizes images with paired text descriptions scraped from the html files of web pages. The CLIP pre-training aims to align the image feature representation of an image encoder with the text feature

representation of a text encoder model (Radford et al. 2021). The text encoder´s feature representations extracted are then used as a basis for a diffusion model to generate images based on text instructions (Ramesh et al. 2022). Another high impact example of self-supervised pre-training and targeted post-training is the previously described causal language modeling and RLHF approach used for creating LLM-based chatbots, like ChatGPTs (Brown et al. 2020; Ouyang et al. 2022).

2.2 Aims and scope of the presented thesis

In this thesis, the applicability of AI for automated radiological image and free-text report analyses was investigated. The aim was to provide insights into the potential of AI for advancing radiological workflows and thereby improving patient care. The scope of the included eight original works consists of the following workflows:

- Processing of free-text radiological reports: The potential of encoder-based transformers and decoder-based LLMs to structure radiological databases by extracting information from free-text reports was investigated using public English and nonpublic German chest X-ray examinations collected from the radiological information system of the University Hospital Bonn. The application and further training of transformers were compared to traditional, simple rule-based annotation systems and to conventional machine-learning approaches. Finally, it was investigated whether transformer-based report text annotations can be used to unlock clinical databases for supervised training of image-based AI to support the detection of pathological findings described in chest X-ray reports.

- Generating synthetic radiological images: The applicability of image generating AI for creating virtual contrast-enhanced (CE), quantitative T1 maps from native T1 maps was investigated for cardiac magnetic resonance imaging (MRI), with the goal to achieve contrast-free estimation of the extracellular volume (ECV) fraction.

- Supporting diagnostic and treatment decisions based on radiological imaging: It was investigated, if body composition metrics extracted by AI from computed tomography (CT) are valuable for the prognosis of patients with pancreatic cancer treated with high-intensity focused ultrasound. Furthermore, a direct application of AI for solely image-based hazard prediction of patients undergoing transcatheter aortic valve implantation was compared to the evaluation of AI-based scalar body composition metrics with Cox proportional hazard (CPH) models that are commonly applied for creation of prognostic models in clinical research. Furthermore, the potential of AI for characterization of the etiology of a liver cirrhosis into alcoholic and non-alcoholic disease based on T2-weighted MRI was investigated. Lastly, it was examined, if AI for precise segmentation of aneurysms of the abdominal aorta can be applied to enable evaluation of the perivascular adipose tissue.

3 Results

3.1 **Nowak S**, Wulff B, Layer YC, Theis M, Isaak A, Salam B, Block W, Kuetting D, Pieper CCC, Luetkens JA, Attenberger UI, Sprinkart AM. Privacy-ensuring, open-weights large language models are competitive with closed GPT-4o in extracting chest X-ray findings from free-text reports. Radiology 2025; 314(1):e240895, DOI: 10.1148/radiol.2408

Background

Large-scale secondary use of clinical databases requires automated tools for retrospective extraction from free-text radiology reports.

Purpose

Sharing data and insights on the application of privacy-preserving open-weights large language models (LLM) for report content extraction with comparison to standard rule-based systems and the closed-weights LLMs from OpenAI.

Methods

In this retrospective exploratory study, zero-shot prompting of 17 open-weights LLMs was compared to rule-based annotation and to GPT-4o, GPT-4o-mini, GPT-4-turbo, and GPT-3.5-turbo on a manually annotated public English chest X-ray dataset (University Indiana, 3,927 patients and reports). An annotated nonpublic German chest X-ray dataset (18,500 reports, 16,844 patients, 10,340 male, mean age: 62.6 ± 21.5 years) was used to compared local fine-tuning of all open-weights LLMs via low-rank adaptation (LoRA) and 4-bit quantization to BERT with different subsets of reports (N: from 10 to 14,580). Nonoverlapping 95% confidence intervals of macroaveraged F1 values (MAF1) were defined as relevant differences.

Results

For the English reports, the highest zero-shot MAF1 was observed for GPT-4o (92.4 [87.9-95.9]); GPT-4o outperformed the rule-based CheXpert (73.1 [65.1-79.7]) but was comparable in performance to several open-weights LLMs (top three: Mistral-Large: 92.6 [88.2-96.0], Llama-3.1-70b: 92.2 [87.1-95.8], Llama-3.1-405b: 90.3 [84.6-94.5]). For the German reports, Mixtral-Large (91.6 [90.5-92.7]) had the highest zero-shot MAF1 among the six other open-weights LLMs and outperformed the rule-based annotation (74.8 [73.3-76.1]). Using 1,000 reports for fine-tuning, all LLMs (top three: Mistral-Large: 94.3 [93.5-95.2], OpenBioLM-70b: 93.9 [92.9-94.8], and Mixtral-8x22b: 93.8 [92.8-94.7]) achieved significantly higher MAF1 values than did BERT (86.7 [85.0-88.3]); however, the differences were not relevant when 2,000 or more reports were used for fine-tuning.

Conclusions

LLMs have the potential to outperform rule-based systems for zero-shot 'out-of-the-box' structuring of report databases, with privacy-ensuring open-weights LLMs being competitive with closed-weights GPT-4o. Additionally, the open-weights LLM outperformed BERT when moderate numbers of reports were used for fine-tuning.

Radiology

ORIGINAL RESEARCH • COMPUTER APPLICATIONS

Privacy-ensuring Open-weights Large Language Models Are Competitive with Closed-weights GPT-4o in Extracting Chest Radiography Findings from Free-Text Reports

Sebastian Nowak, PhD • Benjamin Wulff, MSc • Yannik C. Layer, MD • Maike Theis, MSc • Alexander Isaak, MD • Babak Salam, MD • Wolfgang Block, PhD • Daniel Kuetting, MD • Claus C. Pieper, MD • Julian A. Luetkens, MD • Ulrike Attenberger, MD • Alois M. Sprinkart, PhD

From the Department of Diagnostic and Interventional Radiology, University Hospital Bonn, Venusberg-Campus 1, 53127 Bonn, Germany. Received March 25, 2024; revision requested May 1; final revision received November 8; accepted November 19. **Address correspondence to** S.N. (email: *Sebastian.Nowak@ukbonn.de*).

Supported by the Open Access Publication Fund of the University of Bonn and by the state of North Rhine-Westphalia (SIM-1-1).

Conflicts of interest are listed at the end of this article.

See also the editorial by Gee and Yao in this issue.

Radiology 2025; 314(1):e240895 • https://doi.org/10.1148/radiol.240895 • Content codes: CH AI

Background: Large-scale secondary use of clinical databases requires automated tools for retrospective extraction of structured content from free-text radiology reports.

Purpose: To share data and insights on the application of privacy-preserving open-weights large language models (LLMs) for reporting content extraction with comparison to standard rule-based systems and the closed-weights LLMs from OpenAI.

Materials and Methods: In this retrospective exploratory study conducted between May 2024 and September 2024, zero-shot prompting of 17 open-weights LLMs was preformed. These LLMs with model weights released under open licenses were compared with rule-based annotation and with OpenAI's GPT-4o, GPT-4o-mini, GPT-4-turbo, and GPT-3.5-turbo on a manually annotated public English chest radiography dataset (Indiana University, 3927 patients and reports). An annotated nonpublic German chest radiography dataset (18 500 reports, 16 844 patients [10 340 male; mean age, 62.6 years ± 21.5 {SD}]) was used to compare local fine-tuning of all open-weights LLMs via low-rank adaptation and 4-bit quantization to bidirectional encoder representations from transformers (BERT) with different subsets of reports (from 10 to 14 580). Nonoverlapping 95% CIs of macro-averaged F1 scores were defined as relevant differences.

Results: For the English reports, the highest zero-shot macro-averaged F1 score was observed for GPT-4o (92.4% [95% CI: 87.9, 95.9]); GPT-4o outperformed the rule-based CheXpert [Stanford University] (73.1% [95% CI: 65.1, 79.7]) but was comparable in performance to several open-weights LLMs (top three: Mistral-Large [Mistral AI], 92.6% [95% CI: 88.2, 96.0]; Llama-3.1–70b [Meta AI], 92.2% [95% CI: 87.1, 95.8]; and Llama-3.1–405b [Meta AI]: 90.3% [95% CI: 84.6, 94.5]). For the German reports, Mistral-Large (91.6% [95% CI: 90.5, 92.7]) had the highest zero-shot macro-averaged F1 score among the six other open-weights LLMs and outperformed the rule-based annotation (74.8% [95% CI: 73.3, 76.1]). Using 1000 reports for fine-tuning, all LLMs (top three: Mistral-Large, 94.3% [95% CI: 93.5, 95.2]; OpenBioLLM-70b [Saama]: 93.9% [95% CI: 92.9, 94.8]; and Mixtral-8×22b [Mistral AI]: 93.8% [95% CI: 92.8, 94.7]) achieved significantly higher macro-averaged F1 score than did BERT (86.7% [95% CI: 85.0, 88.3]); however, the differences were not relevant when 2000 or more reports were used for fine-tuning.

Conclusion: LLMs have the potential to outperform rule-based systems for zero-shot "out-of-the-box" structuring of report databases, with privacy-ensuring open-weights LLMs being competitive with closed-weights GPT-4o. Additionally, the open-weights LLM outperformed BERT when moderate numbers of reports were used for fine-tuning.

Supplemental material is available for this article.

The diagnostic workflow in a radiologic department consists of two fundamental steps: image acquisition and subsequent assessment by the radiologist of the images alongside relevant clinical information. The radiologist's assessment is documented in a report that is still typically written in a free-text format. This has substantial disadvantages, as the contents of the reports are not easily accessible retrospectively. A tool for standardized, automatic extraction of information from reports would considerably facilitate the construction of structured clinical databases for secondary use (1–5). Moreover, this approach would allow structured information of imaging findings to be linked to clinical, laboratory, and pathology data in clinical data repositories, enabling data-integrated health care. Such systems could support rapid, automatic identification of study cohorts for not only epidemiologic research but also for developing image-based or even multimodal prognostic artificial intelligence systems (4,5). Additionally, the identification of patients with similar imaging and clinical findings may support diagnosis or may be used for teaching purposes. Therefore, automated and robust natural language processing systems for the extraction of report contents that require little or no manual annotation are of great interest (1,6).

The first typical systems achieved data extraction by recognizing terms and negations with human-defined, hard-coded rules (7). The capabilities of such rule-based systems are limited, especially if the task is more complex than extracting findings that can be recognized by simple mention with or

Abbreviations

BERT = bidirectional encoder representations from transformers, LLM = large language model

Summary

Privacy-ensuring, open-weights large language models showed great potential in the extraction of structured content from free-text radiology reports, facilitating the secondary use of clinical databases for applications in data-driven medicine.

Key Results

- Zero-shot prompting of open-weights large language models (LLMs) achieved higher performance than CheXpert (rule-based system) and competitive performance to that of the closed-weight OpenAI's GPT-4o on public English reports (macro-averaged F1 score: CheXpert [Stanford University], 73.1% [95% CI: 65.1, 79.7]; GPT-4o, 92.4% [95% CI: 87.9, 95.9]; Mistral-Large [Mistral AI], 92.6% [95% CI: 88.2, 96.0]).
- Locally fine-tuned open-weights LLMs using 1000 or fewer annotated reports showed higher performance to training the previous state-of-the-art bidirectional encoder representations from transformers (BERT) on nonpublic German reports (macro-averaged F1 score: rule-based, 74.8% [95% CI: 73.3, 76.1]; BERT, 86.7% [95% CI: 85.0, 88.3]; Mistral-Large, 94.3% [95% CI: 93.5, 95.2]).

without accompanying negation. Indeed, findings characterized by variable wording and formulations might be of particular interest for secondary use in research (8). Newer natural language processing systems, such as term frequency-inverse document frequency classifiers or encoder-based bidirectional encoder representations from transformers (BERT) are based on deep learning (8,9); however, as these techniques require training of the classification head, zero-shot prompting without the use of annotated examples is not feasible.

Since then, OpenAI (10) has triggered a massive increase in interest and investment in artificial intelligence with the release of the chat generative pretrained transformer, or ChatGPT, on November 30, 2022. Such decoder-based transformers (such as the now GPT-4o–based ChatGPT) showed extraordinary capabilities when trained with massively scaled data and computational hardware. Their generative nature allows for zero-shot prompting to any text-based task, without the need for adaptation of a classification head (11). However, OpenAI and other prominent closed-model providers with capable frontier models, such as Anthropic, Google, or X, do not publish them for free use under the open-weights license nor do they allow customers to obtain access to the model weights for implementation on local hardware. Importantly, processing protected health information on servers outside the secure clinical infrastructure is highly regulated by strict data protection laws in many countries (including the European Union). Processing on external servers still constitutes a legal problem even if de-identification methods are used, as these methods do not currently guarantee 100% accuracy (5,8). Thus, there are initiatives by open-source communities and companies that seek to democratize the application of LLMs (12–16). These initiatives allow models to be implemented on local hardware in the same secure clinical network where the health information is stored, even when the network is disconnected from the public internet. Moreover, such local implementation of models with openly available weights allows clinical researchers to further optimize LLMs with their nonpublic clinical data to achieve even higher performance for their desired applications without the extensive efforts for required regulatory compliance when sharing protected information with closed model providers. Implementation within secure networks without public sharing also obviates the need for testing model resistance against adversarial attacks aimed at re-identifying protected information from training data and omits vendor dependencies with vulnerability to the changing price policies of model providers. However, the capabilities of LLMs for language understanding are typically reported on nonmedical tasks, such as the public benchmarks Massive Multitask Language Understanding (17) and Grade School Math 1000 (18).

In this exploratory study, the multilingual capabilities of different publicly available open-weights LLMs for extracting information from radiology free-text reports were compared with each other and with established approaches using rule-based systems and BERT. The aim was to share data and insights from comprehensive experiments on the efficient retrospective structuring of radiology databases with the medical community and to release the code of the experiments for local LLM prompting and fine-tuning under an open-source license *(https://github.com/ukb-rad-cfqiai/LLM_based_report_info_extraction/)*.

Materials and Methods

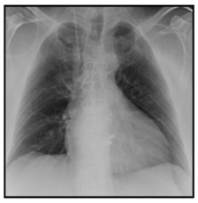
Datasets

This retrospective exploratory study uses a German nonpublic dataset and an English public dataset that were compiled and manually annotated in previous open-access works of the University of Bonn and Indiana University. Figure 1 shows an overview of the entire study, including how the datasets were divided into training and test sets.

English public dataset.—An English public chest radiography dataset with outpatients from the Indiana University (Open-i; *https://openi.nlm.nih.gov*) was analyzed; permission was obtained from the corresponding author to use their data with OpenAI services (19). Pleural effusion (160 of 3927, 4%) and pneumothorax (27 of 3927, 1%) were investigated as the union of labels included in the public English dataset, the CheXpert labeler (Stanford University), and the nonpublic test set. All reports in the dataset (n = 3927) were used for testing zero-shot prompting. The public CheXpert system was used for rule-based content extraction (7). Labels with "uncertain" values were treated as negative classifications.

German nonpublic dataset.—The requirement for written informed consent was waived by the institutional review board approval (AZ411/21). The retrospective dataset derived in a previous study included 93 368 free-text chest radiography reports written in German that contained data of 20 912 patients (mean age, 62.6 years ± 21 [SD]; 8081 female) who were treated between December 2015 and July 2021 in the intensive care units of the University Hospital Bonn; the data were extracted consecutively from the radiologic information system (8). For a subset of 18 500 reports, a radiology resident (Y.C.L., 3 years of experience reporting chest radiography)

Datasets



English: 3,927 public reports from the University Hospital Indiana (Open-i)

human annotation

A test set 3,927

German: 93,368 nonpublic reports from the University Hospital Bonn

human annotation

B train set 14,580 | **C** valid set 1,620 | **D** test set 2,300

Rule-based systems

Public English reports
Apply public rule-based system: *github.com/stanfordmlgroup/chexpert-labeler*
for the findings effusion, pneumothorax that both can be found in Open-i, chexpert and non-public dataset labels.

Nonpublic German reports
Mention extraction:
1. Find mentionings of typical terms.
2. Identify relevant part of the sentence.

Four typical findings:
Infiltrates, effusion, congestion, pneumothorax
Mention classification:
3. **If** description of negation or uncertainty.
→ class=0; else → class=1

Central venous catheter (CVC):
regular/misplaced
Mention classification:
3. **If** terms are mentioned that directly state whether CVC is regular/misplaced.
→ regular=1 or misplaced=1
else if projection on superior vena cava or cavoatrial junction is mentioned.
→ regular=1
else → misplaced=1

Large Language Models (LLM)

Decoder-based transformer, pre-trained by causal language modelling and instruction fine-tuned by reinforcement learning from human feedback.

Closed-weights LLMs
Foundation models:
OpenAI:
GPT-4o
GPT-4o-mini
GPT-4-turbo
GPT-3.5-turbo

Versus

Open-weights LLMs

Foundation models (FM):		Foundation models further fine-tuned:	
Meta AI:	**Microsoft:**	aaditya:	lmsys:
Llama-3.1-405b	Phi-3-medium	OpenBioLM-70b	Vicuna-13b
Llama-3.1-70b	Phi-3-mini	OpenBioLM-8b	BioMistral:
Llama-3.1-8b	**Mistral AI:**	epfl-llm:	Biomistral-7b
Google:	Mistral-Large	Meditron-70b	
Gemma-2-27b	Mixtral-8x7b/22b	Meditron-7b	
Gemma-2-9b	Mistral-7b		

Experiments

Public English reports

1) Zero-shot application without fine-tuning
Compare open- and closed-weights LLMs without fine-tuning to the rule-based system on test set A.

Statistical analysis

$$\text{F1 score} = \frac{2\text{xTP}}{2\text{xTP} + \text{FP} + \text{FN}}$$

TP=true positive
FP=false positive
FN=false negative

Macroaveraged F1 score:
Mean of all classes without class weighting.

Definition of relevant difference:
Non-overlapping 95% confidence intervals created by bootstrapping with 1,000 resamples.

Nonpublic German reports

2) Zero- and one-shot application without fine-tuning
Compare open-weights LLMs without fine-tuning to rule-based system on test set D.

3) Model fine-tuning with human annotated reports
Compare open-weights LLMs to BERT on test set D, after fine-tuning on train set B.

Note, application and fine-tuning of closed-weights LLMs was not possible due to German data protection law.

BERT: Encoder-based transformer, open pre-trained and further custom pre-trained on the non-public reports (excluding test D) by masked language modelling.

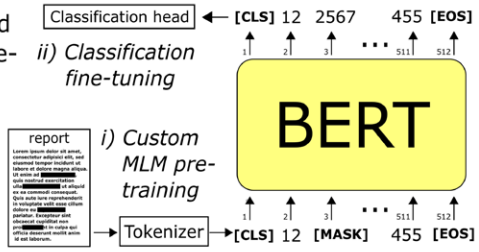


Figure 1: Overview of the current study, which included both public English (Open-i; *https://openi.nlm.nih.gov*) and nonpublic German chest radiography reports. The performance following zero- and one-shot prompting of 17 open-weight large language models (LLMs) was compared with that of the OpenAI's closed-weights LLMs and hard-coded rules for report content analysis. Additionally, the efficacy of fine-tuning of the open-weights LLMs using nonpublic German reports was compared with that of bidirectional encoder representations from transformers (BERT). b = billion, CLS = classification token, EOS = end-of-sequence token, MLM = masked language modelling.

Prompt for public English chest x-ray reports:

You are a helpful AI assistant who structures radiological chest X-ray reports in JSON format. At the end of this instruction, I will give you a report for which you summarize the radiologist's assessments and findings in the following JSON format:

```
{"pleural_effusion": 0,
 "pneumothorax": 0}
```

You always give this full JSON format with all 2 classes and replace 0 with 1 if the following is found in the report: pleural_effusion: If the report mentions effusion or pleural fluid in the lung you enter 1 for pleural_effusion in the JSON. pneumothorax: If the report mentions pneumothorax, hemopneumothorax, hydropneumothorax or collapse of the lung you enter 1 for pneumothorax in the JSON. If the radiologist describes negates the presence of a pathology (examples: "No evidence of pleural effusion", "No pneumothorax") or if he describes uncertainties (example: "Pleural effusion be excluded with certainty / no reliable evidence"), then leave 0 in the JSON for the respective pathology.

Prompt for nonpublic German chest x-ray reports (translated from German to English):

You are a helpful AI assistant who structures radiological chest X-ray reports in JSON format. At the end of this instruction, I will give you a report for which you summarize the radiologist's assessments and findings in the following JSON format:

```
{"CVC": 0,
 "CVC has incorrect projection/location": 0,
 "effusion": 0,
 "congestion": 0,
 "infiltrates": 0,
 "pneumothorax": 0}
```

You always give this full JSON format with all 6 classes and replace 0 with 1 if the following is found in the report: for "CVC" you replace 0 with 1 in the JSON format if the patient has a central venous catheter (CVC). Other foreign materials, such as Shaldon catheters or gastric tubes, are not relevant for you. For "CVC has incorrect projection/position", replace 0 with 1 in JSON format if the central venous catheter (CVC) described in the report has an incorrect position. For the classes "effusion", "congestion", "infiltrates" and "pneumothorax", replace 0 with 1 if the radiologist has noted in the report that he has recognized the respective pathology in the image, regardless it was already present in a previous examination (example: differential diagnosis (DD) pneumonic infiltrates). If the radiologist describes that he does not see the pathology in question on the image (example: "No evidence of circumscribed pneumonic infiltrates") or if he describes uncertainties (example: "Infiltrates cannot be excluded with certainty / no reliable evidence"), then leave 0 in the JSON for the respective pathology. Notes: For descriptions of a central venous catheter (CVC) with "projection onto the upper caval vein", "projection onto the superior vena cava (SVC)" or "projection onto the cavoatrial junction", the position of the CVC is correct and you leave 0 for "CVC has incorrect projection/location". Replace 0 with 1 for "CVC has incorrect projection/location" for all descriptions of projections onto other anatomies, e.g. "projection onto the right atrium" or the description of a folded CVC, because then there is an incorrect projection/location. This is an example of a report classification: Subject: Mrs. PPPPPPPPPP PPPPPPPPP, born DDDDDDDDDDDD Examination: X-ray intensive thorax from DDDDDDDDDD. Indication: Condition after right jugular CVC insertion, location, exclusion of pneumothorax. Findings: A preliminary examination from DDDDDDDDDD is used for comparison. Right jugular CVC with a folded CVC tip and projection onto the right subclavian vein, position correction recommended. No definable pneumothorax limit (as assessable in supine position). ETT in projection onto the trachea, 4 cm above the carina. Nasogastric tube, the course of which cannot be reliably traced to the subdiaphragmatic level due to overlay. Pleural effusions on both sides. Stained, basal parenchymal thickening on both sides of the lower lung subfields, pneumonic infiltrates are possible here with corresponding clinical findings, DD dystelectasis. PV congestion. Wide mediastinal shadow. This is a correctly completed JSON for the example report:

```
{"CVC": 1,
 "CVC has incorrect projection/position": 1,
 "effusion": 1,
 "congestion": 1,
 "infiltrates": 1,
 "pneumothorax": 0}
```

Color legend: System prompt, user instruction, special notes for CVC evaluation, one-shot example

Figure 2: Example of a prompt given to the large language models for reporting content classification of public English reports and nonpublic German reports. AI = artificial intelligence, DD = differential diagnosis, ETT = endotracheal tube, JSON = JavaScript Object Notation, PV = pulmonary venous.

and two medical research assistants (intensively trained by the resident) manually annotated the four frequently observed findings on chest radiographs in an intensive care unit and whether the central venous catheter had regular or misplaced positioning. These labels and their relative appearances were pulmonary infiltrates (3624 of 18 500, 20%), pleural effusion (8422 of 18 500, 46%), pulmonary congestion (6285 of 18 500, 34%), pneumothorax (669 of 18 500, 4%), regular position (8456 of 18 500, 46%), and misplaced position (1562 of 18 500, 8%) of the central venous catheter. This dataset was randomly split into a training set (n = 14 580), a validation set (n = 1620), and a test set (n = 2300). More detailed information on the dataset can be found in Appendix S1 and the publication where it was originally described (8). German

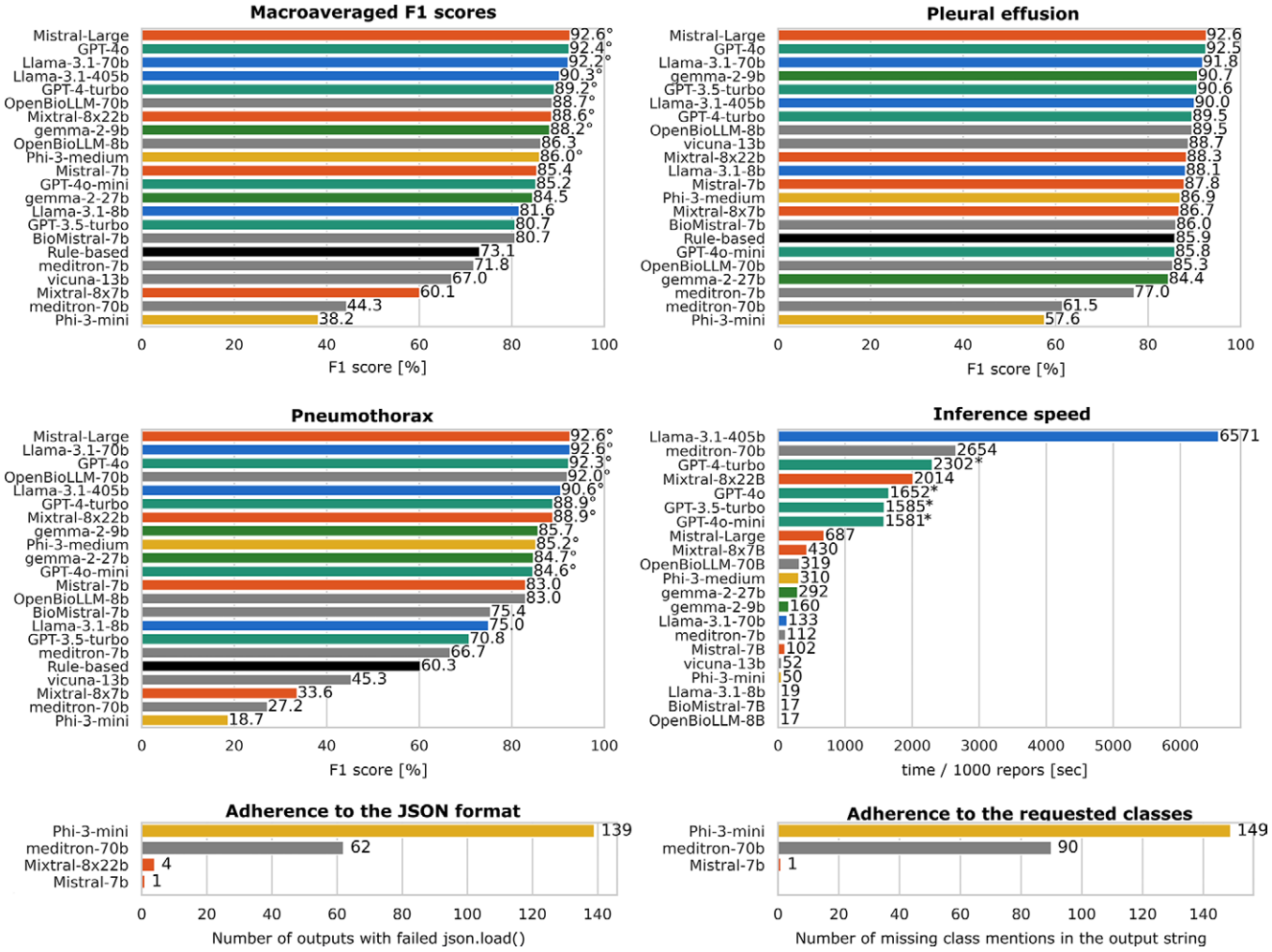


Figure 3: Bar plots of the results of zero-shot prompting of open- and closed-weights large language models without fine-tuning on public English reports. ° Indicates F1 scores with nonoverlapping 95% CIs exceeding the rule-based CheXpert system. The OpenAI's models could not be applied with batched inference, as the only access was to a tier 1 account with rate limits. For all other open-weights models, a single node with eight Nvidia A100 GPUs was used for inference. The adherence to the JavaScript Object Notation (JSON) format and to the requested classes is given for models with failures in the output string.

data protection regulations precluded sharing this dataset with OpenAI services. A rule-based annotation system for German reports was created that searches for specific terms, negations, and descriptions of uncertainty or applies further text-based rules in the "findings" section of the report. Detailed descriptions of the rule-based system can be found in Appendix S2.

Prompting of Open- and Closed-weights LLMs without Fine-tuning

The performance of 17 recently published open-weights LLMs was compared with that of OpenAI's closed-weights LLMs. These experiments were conducted between May 2024 and September 2024. Given the abundance of open- and closed-weights LLMs available, only the most recent foundation models released by the following U.S. and E.U. technology companies were included as follows: OpenAI: GPT-4o, GPT-4o-mini, GPT-4-turbo, GPT-3.5-turbo; Mistral AI: Mistral-Large, Mixtral-8×22b, Mixtral-8×7b, Mistral-7b; Meta AI: Llama-3.1–405b, Llama-3.1–70b, Llama-3.1–8b; Google: Gemma-2–27b, Gemma-2–9b; Microsoft: Phi-3-medium, Phi-3-mini) (12–16,20). Additionally, medical fine-tuned versions of foundation models (BioMistral-7b [Nantes Université] fine-tuned from Mistral-7B; Meditron-70b and 7b [Swiss Federal Institute of Technology] fine-tuned from Llama-2–70b and 7b; and OpenBioLLM-70b and 8b [Saama] fine-tuned from Llama-3–70b and 8b) and Vicuna-13b [LMSYS], a nonmedical fine-tuned version of Llama-2–13b that was previously proposed as a tool for privacy-ensuring report content extraction in radiology reports, were investigated (21–23). For each model, the used version of the Python transformers library (Huggingface) and the git commit identifier of the model weights are provided in Table S1. For all LLMs, the version adapted to follow instructions through reinforcement learning from human feedback was applied. The models were prompted to generate a JavaScript Object Notation (JSON) text that included all classes to be extracted, followed by a "0" if the finding was not found or a "1" if it was found. If the output could not be loaded with Python's "json.load ()" due to format problems (JSON error), the classes were identified programmatically directly from the output string. If a class still could not be identified, the prediction was considered incorrect in the performance evaluation (class error). For the nonpublic German reports, a one-shot prompting in which a single report with correct JSON output was provided as an example in the prompt was explored. The full prompts are shown in Figure 2.

Table 1: Results for Zero-shot Prompting of Open- and Closed-weights Large Language Models without Fine-tuning on the Public English Report Test Set

Model	MAF1 (%)	F1 Score (%)	
		Pleural Effusion	Pneumothorax
CheXpert	73.1	85.9	60.3
FM			
GPT-4o	92.4*	92.5	92.3*
GPT-4o-mini	85.2	85.8	84.6*
GPT-4-turbo	89.2*	89.5	88.9*
GPT-3.5-turbo	80.7	90.6	70.8
Mistral-Large	92.6*	92.6	92.6*
Mixtral-8×22b	88.6*	88.3	88.9*
Mixtral-8×7b	60.1	86.7	33.6
Mistral-7b	85.4	87.8	83.0
Llama-3.1–405b	90.3*	90.0	90.6*
Llama-3.1–70b	92.2*	91.8	92.6*
Llama-3.1–8b	81.6	88.1	75.0
Gemma-2–27b	84.5	84.4	84.7*
Gemma-2–9b	88.2*	90.7	85.7
Phi-3-medium	86.0*	86.9	85.2*
Phi-3-mini	38.2	57.6	18.7
Fine-tuned from FM			
Vicuna-13b	67.0	88.7	45.3
BioMistral-7b	80.7	86.0	75.4
Meditron-70b	44.3	61.5	27.2
Meditron-7b	71.8	77.0	66.7
OpenBioLLM-70b	88.7*	85.3	92.0*
OpenBioLLM-8b	86.3	89.5	83.0

Note.—b = billion, FM = foundation model, MAF1 = macro-averaged F1 score.
* igher F1 scores with nonoverlapping 95% CIs compared with the rule-based CheXpert system.

Fine-tuning of Open-weights LLMs on Nonpublic German Reports

Local distributed data parallel fine-tuning of open-weights LLMs was explored, that is, training with model replication across all graphics processing units (GPUs). Specifically, low-rank adaptation and 4-bit quantization were applied to lower computational requirements and enable LLM training on a single node with eight Nvidia A100 80-GB GPUs (24). When fine-tuning with low-rank adaptation, the parameters of the LLM were kept frozen, and adjustments to the parameters were made through adapters representing low-rank decompositions of the weight matrices (25). Notably, distributed data parallel fine-tuning of Llama-3.1–405b could not be investigated with the available GPU setup. LLM fine-tuning was compared with training of the current state-of-the-art encoder-based BERT using various numbers of reports from the training set (10, 50, 100, 250, 500, 1000, 2000, 3500, 7000, and all 14 580 reports). BERT was trained on a workstation with a single Nvidia RTX 3090 24-GB GPU. Detailed information on the LLMs and BERT and their training scheme can be found in Appendix S3.

Statistical Analysis

F1 scores were calculated (S.N.) via Python version 3.9.12 *(https://www.python.org/)* and scikit-learn version 0.24.2 *(https://scikit-learn.org/)*. Macro-averaged F1 scores represent the mean of F1 scores. The 95% CIs were calculated by bootstrapping with 1000 resamples (26). Nonoverlapping CIs were interpreted as relevant differences between models.

Results

Prompting of Open- and Closed-weights LLMs without Fine-tuning

Zero-shot prompting on public English reports.— GPT-4o achieved a high macro-averaged F1 score (92.4% [95% CI: 87.9, 95.9]) that was higher than that of the rule-based CheXpert (73.1% [95% CI: 65.1, 79.7]); however, its performance was comparable to that of the best performing open-weights LLMs (top three: Mistral-Large, 92.6% [95% CI: 88.2, 96.0]; Llama-3.1–70b, 92.2% [95% CI: 87.1, 95.8]; Llama-3.1–405b, 90.3% [95% CI: 84.6, 94.5]). Notably, for some LLMs, including GPT-3.5-turbo, 95% CIs overlapped with the rule-based CheXpert annotation. F1 scores, JSON error rates, class error rates, and inference speeds can be found in Figure 3 and Table 1.

Zero-shot prompting on the nonpublic German reports.— The top seven open-weights LLMs (Mistral-Large, 91.6% [95% CI: 90.5, 92.7]; Llama-3.1–405b, 90.4% [95% CI: 89.2, 91.4]; Llama-3.1–70b, 87.5% [95% CI: 86.1, 88.7]; Mixtral-8×22b, 87.3% [95% CI: 85.8, 88.6]; OpenBioLLM-70b, 84.6 [95% CI: 83.2, 85.8]; Llama-3.1–8b, 83.5% [95% CI: 82.1, 84.8]; Gemma-2–27b: 78.1% [95% CI: 76.7, 79.4]) differed from the rule-based annotation system in their macro-averaged F1 score (74.8% [95% CI: 73.3, 76.1]).

One-shot prompting on nonpublic German reports.— Mistral-Large (91.1% [95% CI: 90.0, 92.1]) still had the highest macro-averaged F1 score of the other seven open-weights LLMs that had showed relevant differences in performance from rule-based annotations (74.8% [95% CI: 73.3, 76.1]). Not all LLMs benefitted from the provision of an example within the prompt (eg, Mistral-7b: zero-shot macro-averaged F1 score of 77.7% [95% CI: 75.9, 79.2] vs one-shot macro-averaged F1 score of 70.8% [95% CI: 68.7, 72.7]; Llama-3.1–8b: zero-shot macro-averaged F1 score of 83.5% [95% CI: 82.1, 84.8] vs one-shot macro-averaged F1 score of 81.8% [95% CI: 80.5, 83.2]). F1 scores, JSON error rates, and class error rates can be found in Figure 4 and Table 2.

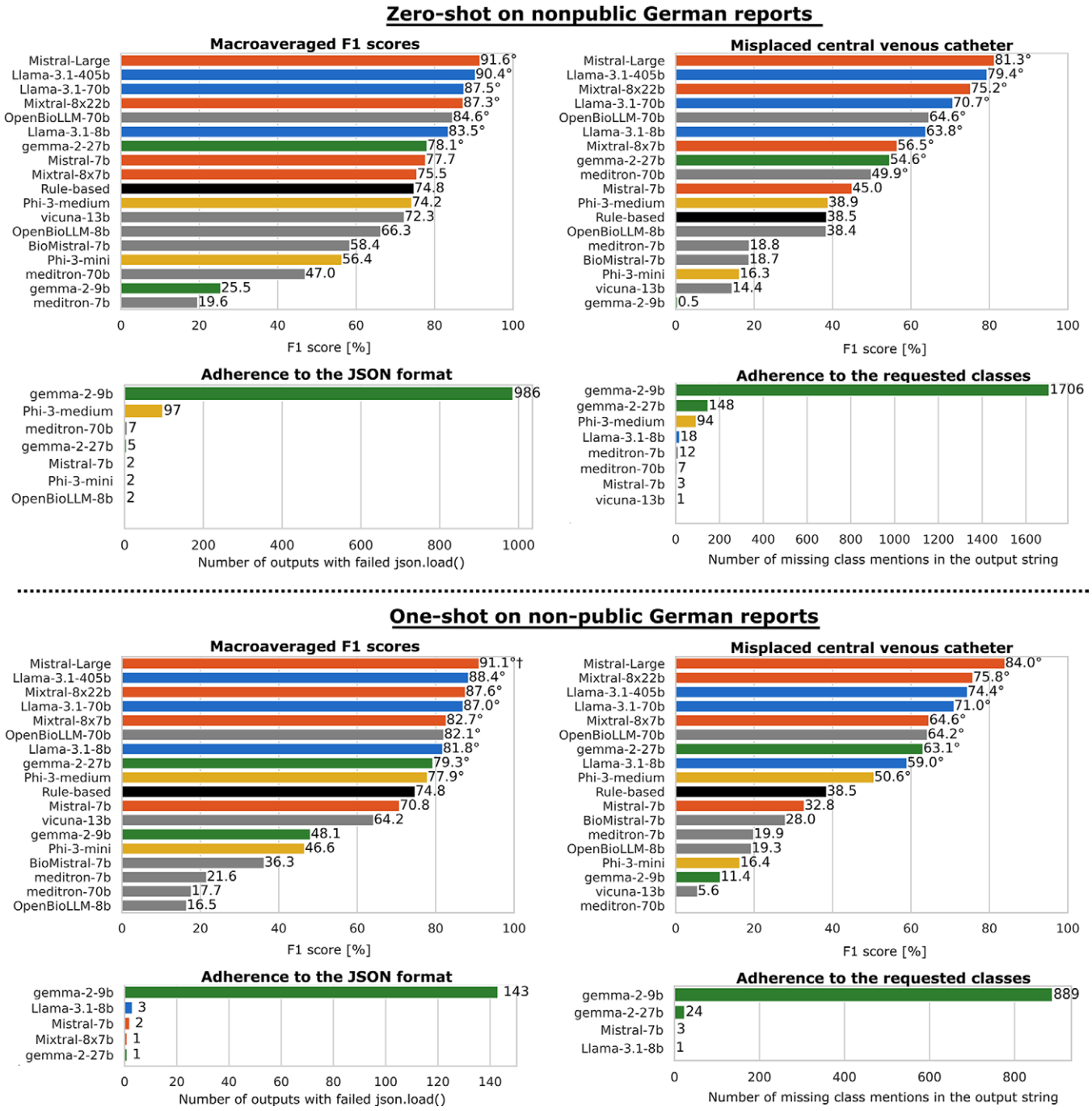


Figure 4: Bar plots of the results of zero- and one-shot prompting of open-weights large language models without fine-tuning on nonpublic German reports. ° Indicates F1 scores with nonoverlapping 95% CIs exceeding the rule-based system. † F1 scores with nonoverlapping 95% CIs exceeding all other models. Zero- or one-shot prompting of BERT was not feasible; only open-weights models were applied, as sharing reports with third parties by using closed-weights OpenAI models was not possible owing to data protection. The adherence to the JavaScript Object Notation (JSON) format and to the requested classes is given for models with failures in the output string.

Fine-tuning of Open-weights LLMs on Nonpublic German Reports

Tables 3–6 and Figures 5 and 6 show the results of fine-tuning the privacy-ensuring, open-weights LLMs. When 1000 or fewer manually annotated reports were used for fine-tuning, all the models (top three: Mistral-Large, 94.3% [95% CI: 93.5, 95.2]; OpenBioLLM-70b, 93.9% [95% CI: 92.9, 94.8]; Mixtral-8×22b, 93.8% [95% CI: 92.8, 94.7]) achieved higher macroaveraged F1 scores than did the fine-tuned BERT (86.7% [95% CI: 85.0, 88.3]). The performance of BERT was not different when 2000 or more annotated reports were used for fine-tuning. Additional experiments on the effects of low-rank adaptation settings and additional scatterplots with the number of model parameters and training examples can be found in Appendixes S4 and S5. All results reported in balanced accuracy are shown in Tables S3–S9 and Figures S3–S6.

Discussion

In this explorative study, we investigated how privacy-ensuring open-weights large language models (LLMs) for chest radiography

Table 2: Results for Zero-shot and One-shot Prompting of Open-weights Large Language Models without Fine-tuning on the Nonpublic German Report Test Set

		F1 Score (%)					
Model	MAF1 (%)	Misplaced CVC	Regular CVC	Infiltrates	Congestion	Pleural Effusion	Pneumothorax
				Zero-shot Prompting			
Rule based	74.8	38.5	68.3	68.0	93.3	95.1	85.2
Mistral-Large	91.6*	81.3*	92.5*	85.4*	98.0*	98.3*	93.9
Mixtral-8×22b	87.3*	75.2*	87.9*	79.5*	96.3*	96.9*	88.2
Mixtral-8×7b	75.5	56.5*	82.3*	70.9	95.1	87.1	61.1
Mistral-7b	77.7	45.0	84.1*	70.0	94.2	88.1	84.7
Llama-3.1–405b	90.4*	79.4*	91.4*	81.4*	97.9*	98.6*	93.4
Llama-3.1–70b	87.5*	70.7*	84.8*	82.7*	95.8*	98.1*	92.9
Llama-3.1–8b	83.5*	63.8*	84.8*	75.1*	94.8	93.6	89.0
Gemma-2–27b	78.1*	54.6*	76.4*	70.0	84.2	96.7*	86.7
Gemma-2–9b	25.5	0.5	6.6	41.8	52.5	45.7	6.1
Phi-3-medium	74.2	38.9	75.2*	63.1	91.0	95.5	81.7
Phi-3-mini	56.4	16.3	3.7	54.6	89.3	87.5	86.9
Vicuna-13b	72.3	14.4	79.7*	74.1*	88.0	90.4	87.4
BioMistral-7b	58.4	18.7	28.4	54.1	82.9	88.9	77.6
Meditron-70b	47.0	49.9*	23.5	68.1	46.6	51.7	42.0
Meditron-7b	19.6	18.8	0.0	24.4	32.2	36.2	6.2
OpenBioLLM-70b	84.6*	64.6*	85.7*	72.9	95.6	95.6	93.0
OpenBioLLM-8b	66.3	38.4	62.2	61.4	85.2	81.1	69.8
				One-shot Prompting			
Mistral-Large	91.1*†	84.0*	93.2*	79.3*	97.9*†	98.0*	94.5*
Mixtral-8×22b	87.6*	75.8*	89.0*	77.4*	96.2*	96.4	91.0
Mixtral-8×7b	82.7*	64.6*	88.4*	75.8*	92.9	92.0	82.5
Mistral-7b	70.8	32.8	76.6*	67.5	90.8	96.3	69.0
Llama-3.1–405b	88.4*	74.4*	93.0*	75.4*	96.2*	97.3*	93.9
Llama-3.1–70b	87.0*	71.0*	90.0*	75.3*	96.3*	97.9*	91.6
Llama-3.1–8b	81.8*	59.0*	82.7*	69.1	95.7*	95.7	88.9
Gemma-2–27b	79.3*	63.1*	78.3*	60.1	90.0	92.0	92.4
Gemma-2–9b	48.1	11.4	15.1	52.7	67.5	92.6	49.2
Phi-3-medium	77.9*	50.6*	81.7*	64.3	91.0	78.4	88.8
Phi-3-mini	46.6	16.4	5.6	44.7	67.6	80.3	67.2
Vicuna-13b	64.2	5.6	75.3*	44.7	80.5	80.3	77.9
BioMistral-7b	36.3	28.0	47.5	29.4	38.0	32.9	42.2
Meditron-70b	17.7	0.0	0.0	35.6	23.5	15.5	31.7
Meditron-7b	21.6	19.9	0.0	30.6	37.2	39.2	2.3
OpenBioLLM-70b	82.1*	64.2*	87.5*	64.7	91.5	92.4	92.7
OpenBioLLM-8b	16.5	19.3	12.7	16.7	20.3	18.9	11.1

Note.—b = billion, CVC = central venous catheter, MAF1 = macro-averaged F1 score. Foundation models are from Mistral AI, Meta AI, Google, and Microsoft.

* igher F1 scores with nonoverlapping 95% CIs compared with the rule-based system.

† Higher F1 scores with nonoverlapping 95% CIs compared with all other models at a given number of reports.

report content extraction from public English and nonpublic German reports compared with simple rule-based systems and with closed-weights LLMs from OpenAI. We have examined both prompting and fine-tuning of LLMs.

Zero-shot Prompting

Open-weights LLMs showed potential for efficient zero-shot "out-of-the-box" structuring of report databases compared with rule-based systems on public English reports, with some privacy-ensuring open-weights LLMs being competitive with the closed-weights GPT-4o, despite the use of 4-bit quantization to lower computational requirements. Open-weights LLMs also outperformed rule-based systems on nonpublic German reports, particularly for descriptions that differed considerably between reports; for example, the central venous catheter placement, which is frequently evaluated on radiographs in the intensive care unit, is reflected in various manifestations and, ultimately, has very variable descriptive formulations (8). Such variability complicates detection with hard-coded rules.

Table 3: Macro-averaged F1 Scores for Fine-tuning of Open-weights Large Language Models on Nonpublic German Reports

		Foundation Models										Fine-tuned from Foundation Models					
No. of Reports	BERT (0.1b)	Mistral-Large (123b)	Mixtral-8×22b	Mixtral-8×7b	Mistral-7b	Llama-3.1–70b	Llama-3.1–8b	Gemma-2–27b	Gemma-2–9b	Phi-3-medium (14b)	Phi-3-mini (3.8b)	Vicuna-13b	BioMistral-7b	Meditron-70b	Meditron-7b	OpenBio-LLM-70b	OpenBio-LLM-8b
10	15.1	92.2*†	88.2*	77.7*	76.3*	87.6*	67.4*	84.1*	20.9*	78.0*	68.5*	64.7*	54.8*	48.6*	36.4*	83.7*	64.0*
50	27.2	92.3*	90.4*	86.7*	84.8*	89.2*	83.7*	81.0*	22.8	81.1*	74.5*	66.2*	84.0*	85.1*	70.8*	87.3*	82.2*
100	38.0	93.8*	92.6*	87.2*	86.9*	89.5*	86.3*	84.3*	26.0	82.0*	71.1*	66.4*	84.0*	85.9*	68.8*	88.0*	83.1*
250	63.1	93.9*	93.0*	90.8*	90.8*	92.2*	87.0*	90.8*	31.2	89.1*	85.4*	81.2*	89.5*	90.6*	85.4*	92.2*	82.6*
500	80.7	93.8*	92.8*	91.6*	91.4*	92.8*	89.1*	90.0*	30.2	89.8*	86.2*	86.9*	90.8*	90.9*	86.1*	92.6*	91.3*
1000	86.7	94.3*	93.8*	93.2*	93.0*	93.6*	91.6*	92.0*	30.2	92.8*	90.0*	89.9*	92.9*	93.2*	91.2*	93.9*	93.1*
2000	92.9	95.0*	94.7	93.5	93.7	94.0	93.2*	92.2	31.5	93.3	90.0	92.0	93.7	94.1	92.1	94.4	94.5
3500	93.8	95.0	93.7	94.8	94.1	93.8	93.9	93.5	30.8	93.7	89.0	92.9	94.3	94.7	93.2	93.3	94.6
7000	95.0	95.1	95.1	94.7	94.7	95.5	93.7	94.4	34.0	94.4	93.3	93.6	94.7	95.1	94.8	95.0	95.2
14 580	95.1	95.3	95.5	95.3	94.9	94.6	94.7	93.9	29.3	94.8	94.7	94.1	95.2	95.3	94.7	95.3	95.3

Note.—BERT = bidirectional encoder representations from transformers. Foundation models are from Mistral AI, Meta AI, Google, and Microsoft. Labeling 500 reports required 5.5 hours of human annotation. The distributed data parallel fine-tuning of Llama-3.1–405b could not be investigated with the graphic processing unit setup available. The number of billion (b) parameters for the models is noted where this is not given by the model name.

* F1 sco es with nonoverlapping 95% CIs exceeding BERT.

† F1 scores with nonoverlapping 95% CIs exceeding all other models at a given number of reports.

Table 4: Results on Central Venous Catheter Positioning for Fine-tuning of Open-weights Large Language Models on Nonpublic German Reports

		Foundation Models										Fine-tuned from Foundation Models					
No. of Reports	BERT (0.1b)	Mistral-Large (123b)	Mixtral-8×22b	Mixtral-8×7b	Mistral-7b	Llama-3.1–70b	Llama-3.1–8b	Gemma-2–27b	Gemma-2–9b	Phi-3-medium (14b)	Phi-3-mini (3.8b)	Vicuna-13b	Bio-Mistral-7b	Meditron-70b	Meditron-7b	OpenBio-LLM-70b	OpenBio-LLM-8b
		Misplaced Central Venous Catheter															
10	0.0	84.1*	77.8*	64.5*	48.1*	71.5*	64.6*	60.6*	0.0	60.6*	30.1*	6.6*	3.9*	18.4*	0.0*	68.2*	39.9*
50	3.4	83.5*	80.6*	74.2*	59.9*	73.4*	68.4*	46.0*	0.1	46.0*	31.7*	28.5*	53.9*	63.5*	22.0*	68.0*	57.7*
100	7.1	88.5*	87.4*	70.9*	66.1*	76.3*	70.1*	54.0*	0.1	54.0*	30.3*	28.0*	59.0*	67.2*	32.4*	69.4*	57.7*
250	10.2	85.8*	83.2*	77.8*	76.8*	83.2*	73.3*	76.2*	0.3	76.2*	57.9*	38.7*	73.4*	73.7*	58.5*	85.9*	65.0*
500	32.7	86.7*	84.1*	80.0*	79.3*	84.3*	80.3*	74.9*	0.3	74.9*	58.7*	65.0*	77.5*	78.0*	56.7*	85.9*	79.6*
1000	57.6	88.5*	87.1*	84.3*	86.6*	87.5*	84.8*	79.2*	0.3	79.2*	70.8*	70.8*	84.4*	85.0*	76.6*	87.2*	84.2*
2000	84.8	90.4	90.3	84.8	86.7	87.7	89.2	83.8	0.4	83.8	73.6	81.2	86.5	89.0	80.8	90.1	89.4
3500	86.0	90.0	86.0	89.8	86.9	87.4	87.5	86.7	1.0	86.7	70.1	84.2	86.9	89.2	82.8	85.9	87.9
7000	90.4	90.8	91.0	87.9	87.5	93.0	91.2	88.8	0.4	88.8	85.8	85.7	87.8	91.2	87.2	90.6	89.4
14 580	91.0	92.0	91.6	89.5	89.4	90.1	91.0	85.8	0.5	85.8	88.4	88.6	90.6	91.3	89.5	91.0	89.7
		Regular Central Venous Catheter															
10	11.6	93.4*	92.9*	87.3*	82.5*	85.4*	85.6*	89.2*	3.2	78.9*	60.6*	68.7*	9.6	9.4	0.0	89.0*	4.1
50	64.8	93.0*	93.7*	91.7*	89.9*	90.9*	88.4*	78.6*	5.0	87.0*	80.2*	48.2	87.9*	89.0*	85.7*	90.4*	82.6*
100	72.4	94.2*	93.5*	90.8*	90.1*	91.1*	84.7*	83.2*	4.8	85.1*	60.6	44.7	86.3*	87.8*	65.7	88.9*	83.4*
250	80.9	94.1*	93.6*	92.5*	92.7*	92.3*	88.7*	92.7*	5.8	91.2*	88.3*	87.4*	91.6*	91.2*	88.7*	93.3*	91.8*
500	87.1	93.6*	91.5*	92.1*	92.8*	91.8*	91.1*	91.1*	5.4	90.5*	88.3	89.0	92.0*	88.3	88.6	91.9*	89.6
1000	90.0	93.1*	91.7	92.8*	93.5*	94.4*	93.1*	92.5*	5.0	93.0*	90.8	91.2	93.8*	92.9*	92.3	93.2*	93.4*
2000	94.0	93.5	93.9	92.8	93.7	94.3	92.4	93.7	5.9	93.5	91.0	90.8	93.1	93.4	92.4	94.6	93.7
3500	94.6	93.7	93.3	94.0	93.2	94.1	92.1	94.2	8.4	94.1	90.2	92.1	94.3	94.4	93.0	93.5	94.4
7000	94.5	93.6	95.3	94.3	94.1	94.4	93.8	94.4	6.4	93.9	92.8	91.2	94.6	95.0	94.7	94.9	94.6
14 580	94.9	94.4	94.1	94.8	93.8	94.7	93.1	94.2	7.3	94.6	94.4	92.3	94.5	94.7	94.0	94.7	95.0

Note.—BERT = bidirectional encoder representations from transformers. Foundation models are from Mistral AI, Meta AI, Google, and Microsoft. Labeling 500 reports required 5.5 hours of human annotation. The distributed data parallel fine-tuning of Llama-3.1–405b could not be investigated with the graphic processing unit setup available. The number of billion (b) parameters for the models is noted where this is not given by the model name.

* F1 sco es with nonoverlapping 95% CIs exceeding BERT.

One-shot Prompting

Interestingly, not all LLMs benefitted from inclusion of a one-shot report example in the prompt. The performance of a model given a one-shot prompt strongly depends on the chosen example (27). Here, the zero-shot prompt for the German reports consisted of broad, generalized instructions. Adding the one-shot example resulted in the last 30% of the prompt being a specific report created in the writing style of one radiologist.

Table 5: Results on Pulmonary Infiltrates and Congestion for Fine-tuning of Open-weights Large Language Models on Nonpublic German Reports

		Foundation Models										Fine-tuned from Foundation Models					
No. of Reports	BERT (0.1b)	Mistral-Large (123b)	Mixtral-8×22b	Mixtral-8×7b	Mistral-7b	Llama-3.1–70b	Llama-3.1–8b	Gemma-2–27b	Gemma-2–9b	Phi-3-medium (14b)	Phi-3-mini (3.8b)	Vicuna-13b	Bio-Mistral-7b	Meditron-70b	Meditron-7b	OpenBio-LLM-70b	OpenBio-LLM-8b
							Pulmonary Infiltrates										
10	14.8	86.4*	74.8*	64.9*	71.3*	82.5*	75.4*	76.2*	37.6*	64.5*	59.2*	70.6*	61.6*	79.6*	58.3*	72.3*	72.6*
50	21.1	84.9*	83.8*	79.7*	83.6*	83.8*	78.6*	75.6*	36.8*	69.7*	71.0*	68.5*	82.9*	81.4*	72.1*	78.1*	78.7*
100	25.0	89.5*	87.8*	78.2*	84.5*	82.0*	79.9*	81.7*	41.2*	73.8*	73.4*	70.1*	82.2*	84.0*	71.6*	82.1*	79.0*
250	59.9	89.7*	89.6*	89.7*	88.9*	89.8*	84.5*	85.9*	51.0	86.1*	84.6*	81.1*	88.0*	88.8*	83.2*	88.2*	82.9*
500	86.9	89.7	90.4	89.8	87.9	89.9	87.4	84.3	49.1	87.1	84.1	83.5	87.3	90.0	84.9	90.0	88.2
1000	89.1	90.5	91.7	89.7	88.5	88.8	89.7	88.3	50.1	89.7	89.1	89.6	89.2	91.3	89.0	90.5	89.3
2000	90.6	92.0	90.7	91.7	90.7	89.5	90.7	87.6	52.4	89.9	88.3	90.6	91.3	91.7	89.2	90.9	91.1
3500	91.0	92.1	89.4	92.1	91.3	90.5	90.1	89.6	49.1	90.4	88.3	90.3	91.0	91.6	90.0	89.7	91.6
7000	91.4	91.5	91.3	91.5	91.2	91.0	89.3	91.0	55.2	90.5	90.4	91.7	90.9	91.4	92.4	91.2	91.8
14 580	92.2	91.5	92.6	91.9	91.1	89.5	91.3	91.9	49.2	91.2	90.7	91.7	91.2	91.7	90.8	91.4	92.1
							Pulmonary Congestion										
10	6.4	97.9*	96.3*	95.9*	96.7*	95.7*	94.9*	90.0*	47.6*	90.6*	90.5*	78.7*	87.1*	54.4*	54.8*	93.3*	89.4*
50	26.1	98.2*	97.0*	96.4*	96.3*	96.0*	95.0*	94.0*	52.2*	90.1*	90.5*	87.9*	95.8*	92.2*	83.5*	95.9*	93.2*
100	57.4	98.1*	97.2*	97.0*	96.2*	97.0*	96.3*	94.9*	55.6	93.5*	90.7*	87.5*	95.8*	92.1*	83.9*	96.5*	93.6*
250	94.8	98.0*	98.4*	96.2	96.3	97.3*	96.9*	97.1*	66.5	97.3*	96.2	96.1	96.6	96.9*	96.6	96.1	91.3
500	97.3	97.7	98.0	97.9	97.4	97.9	97.9	97.1	64.0	97.7	96.9	95.7	97.8	98.2	97.0	97.4	97.0
1000	97.7	98.1	98.1	98.3	98.2	97.6	98.1	97.4	65.3	97.8	97.7	98.0	97.9	98.0	97.8	97.7	97.9
2000	98.0	97.7	98.1	98.6	98.1	98.3	98.4	97.0	65.0	98.2	97.5	98.3	98.2	97.9	98.3	98.2	98.5
3500	98.2	98.5	98.5	98.2	98.7	98.1	98.4	97.5	61.3	98.5	97.5	97.8	98.0	98.3	98.3	97.8	98.2
7000	97.9	97.9	97.7	98.2	98.3	98.2	98.1	97.6	71.2	97.6	98.1	98.2	98.6	98.1	98.1	98.1	98.3
14 580	98.0	98.2	98.5	98.4	98.2	98.1	98.0	97.7	62.2	98.2	98.1	98.0	98.5	98.5	98.6	97.8	98.3

Note.—BERT = bidirectional encoder representations from transformers. Foundation models are from Mistral AI, Meta AI, Google, and Microsoft. Labeling 500 reports required 5.5 hours of human annotation. The distributed data parallel fine-tuning of Llama-3.1–405b could not be investigated with the graphic processing unit setup available. The number of billion (b) parameters for the models is noted where this is not given by the model name.

* F1 sco es with nonoverlapping 95% CIs exceeding BERT.

It may be speculated that this led some LLMs to allocate less attention to the general instructions, as previous work has shown that for longer prompts, LLMs tend to focus primarily on information at the beginning and end of the instructions (28). To address this issue, advanced prompt techniques that select suitable short k-shot examples on the basis of the LLM's own generated outputs could be applied (27).

Adaptation by Fine-tuning

To increase the robustness of classifications, clinics may seek to fine-tune open-weights LLMs, especially for report findings with more variable and challenging descriptions. Fine-tuning of LLMs via low-rank adaptation and 4-bit quantization was shown to improve performance over that of the fine-tuned BERT model when only 1000 or fewer manual annotations were available. When 2000 or more annotated reports were used, the performance differences were not relevant. Nevertheless, open-weights LLMs have advantages over BERT, as reducing the number of required manual annotations is critical; indeed, annotations increase the requirements for installing content extraction pipelines compared with "out of the box" zero-shot applications, which is especially relevant for clinics wanting to structure report databases with a variety of content extraction tasks and report types. However, the hardware requirements of lightweight BERT are drastically lower than those of, for example, Mistral-Large, which has over 1200 times more parameters. This suggests that for clinics without access to sophisticated multi–GPU hardware and limited financial resources, it may still be reasonable to invest in staff who can perform manual annotation and apply lightweight BERT models.

Model Performance Differences

In our study, Mistral-Large, with 123 billion parameters, showed promising zero-shot and fine-tuning results compared with models of similar or even larger sizes. Appendix S6 includes detailed discussion of the results using models with fewer parameters.

In our previous work, we compared BERT with rule-based and simpler deep learning–based NLP approaches using a nonpublic German dataset. This study provided a multilingual benchmark of 21 recent open- and closed-weights LLMs on a public English dataset and a nonpublic German dataset, having investigated zero- and one-shot prompting, as well as low-rank adaptation fine-tuning. A recent study published in *Radiology* (21) investigated the zero-shot prompting of the open-weights Vicuna 13B for English report content classification via English public datasets (Medical Information Mart for Intensive Care [MIMIC]-CXR, National Institutes of Health ChestX-ray14). LLM performance was compared with that of rule-based labels generated by CheXpert and a manually annotated test set of 100 reports. An evaluation of central

Table 6: Results on Pleural Effusion and Pneumothorax for Fine-tuning of Open-weights Large Language Models on Nonpublic German Reports

		Foundation Models										Fine-tuned from Foundation Models					
No. of Reports	BERT (0.1b)	Mistral-Large (123b)	Mixtral-8×22b	Mixtral-8×7b	Mistral-7b	Llama-3.1–70b	Llama-3.1–8b	Gemma-2–27b	Gemma-2–9b	Phi-3-medium (14b)	Phi-3-mini (3.8b)	Vicuna-13b	Bio-Mistral-7b	Meditron-70b	Meditron-7b	OpenBio-LLM-70b	OpenBio-LLM-8b
							Pleural Effusion										
10	57.5	96.7*	96.1*	93.5*	80.6*	98.1*	92.7*	97.3*	32.9	96.0*	86.3*	77.3*	85.0*	64.9*	70.6*	87.4*	91.6*
50	41.7	98.5*	97.7*	91.0*	94.5*	98.2*	94.3*	97.7*	37.2	96.0*	90.3*	90.6*	94.3*	92.6*	85.1*	97.7*	92.2*
100	53.1	98.3*	98.5*	95.8*	97.7*	98.7*	97.2*	97.2*	47.5	96.7*	89.2*	90.4*	96.5*	92.5*	86.5*	97.9*	94.7*
250	96.9	98.7*	98.9*	97.7	98.4*	97.8	98.1	98.0	54.6	98.4*	98.0	98.2*	98.5*	98.6*	97.4	98.8*	95.4
500	98.4	98.7	98.9	98.2	98.5	98.7	98.6	98.4	53.1	98.6	98.1	98.2	98.3	98.1	98.4	98.9	98.5
1000	98.7	98.9	99.0	98.7	98.7	98.8	98.8	98.7	51.9	98.6	98.5	98.6	98.8	99.1	98.7	98.9	98.8
2000	98.9	98.9	98.9	98.2	98.1	99.0	98.2	98.8	56.5	98.5	98.6	98.5	98.7	98.6	98.8	99.0	98.7
3500	98.8	99.1	98.9	98.6	98.7	98.7	98.2	98.6	55.8	98.6	98.7	98.6	98.8	99.0	98.8	98.5	98.7
7000	98.8	99.0	99.1	98.8	99.1	99.0	98.9	99.0	59.6	98.7	98.8	98.8	98.9	98.9	99.0	99.1	99.0
14 580	98.8	98.7	99.0	99.1	99.1	98.7	98.2	98.9	48.2	98.7	98.8	98.9	98.8	98.7	99.0	99.0	98.8
							Pneumothorax										
10	0.0	94.5*	91.1*	60.4*	78.7*	92.2*	89.0*	91.0*	4.3*	89.1*	84.3*	86.4*	81.8*	65.1*	34.5*	92.2*	86.2*
50	6.1	95.8*	89.4*	87.3*	84.5*	92.8*	92.9*	94.0*	5.8	86.7*	83.0*	73.5*	89.3*	91.7*	76.4*	93.6*	89.0*
100	12.8	94.2*	91.4*	90.6*	86.8*	92.1*	93.6*	94.7*	7.0	88.8*	82.4*	77.6*	84.0*	91.6*	72.9*	93.1*	90.3*
250	36.0	97.0*	94.0*	90.9*	91.9*	92.7*	93.0*	94.7*	9.3	89.7*	87.5*	85.7*	89.0*	94.7*	88.2*	91.0*	69.3*
500	81.9	96.4*	94.0*	91.4	92.4	94.2*	94.1*	94.1*	9.5	90.4	90.9	90.2	91.9	92.9*	90.9	91.6	94.7*
1000	87.0	97.0*	95.2	95.2	92.7	94.1	94.7	95.8*	8.4	92.5	93.0	91.2	93.4	93.3	92.9	95.9*	94.7
2000	91.0	97.6	96.3	94.7	94.5	95.2	94.6	92.4	8.8	93.1	90.9	92.4	94.6	94.0	92.9	93.6	95.8
3500	94.4	96.4	96.4	95.9	95.8	94.1	95.9	94.2	9.5	93.1	89.4	94.1	96.4	95.9	96.5	94.3	97.0
7000	97.0	97.6	96.5	97.6	98.2	97.6	97.0	95.8	10.9	96.4	93.6	95.9	97.6	95.8	97.6	95.9	98.2
14 580	95.7	97.1	97.1	98.2	97.6	96.4	95.2	94.8	8.2	94.7	97.6	95.3	97.6	97.0	96.5	97.6	97.6

Note.—BERT = bidirectional encoder representations from transformers. Foundation models are from Mistral AI, Meta AI, Google, and Microsoft. Labeling 500 reports required 5.5 hours of human annotation. The distributed data parallel fine-tuning of Llama-3.1–405b could not be investigated with the graphic processing unit setup available. The number of billion (b) parameters for the models is noted where this is not given by the model name.

* F1 sco es with nonoverlapping 95% CIs exceeding BERT.

venous catheter placement, which requires the interpretation of variable descriptive formulations, was not conducted. The authors were able to show that zero-shot prompts resulted in comparable performance to rule-based content extraction.

We acknowledge several limitations of our study. First, since LLMs are trained with large amounts of publicly available text data, it cannot be guaranteed that public report datasets are not part of the training corpus of some models, as many providers do not provide transparent information on included web data; this would compromise our findings. Recent work has shown the disadvantage of using public data for fair and conclusive benchmarks of LLMs compared with nonpublic data (18). Therefore, we also included nonpublic German chest radiography reports from intensive care units, for which it is certain that these reports were not part of the training corpus of any LLM and contained a high proportion of pathology findings.

Second, we extracted the presence of each finding within each report only as binary labels. However, in other scenarios, the information extraction task could be more complex, including evaluation of the uncertainty or severity of findings, relevant staging information, or dates that cannot be represented by binary classification. As the capabilities of LLMs to generate any text allow for the extraction of more nonbinary information by design (contrary to, eg, the classification head of BERT), we plan to further investigate other text extraction tasks in future studies.

Third, we did not investigate advanced prompting strategies, such as the chain of thought prompting, multiagent conversations, or retrieval augmented generation, with the inclusion of medical training documents as embedding vector databases (29–31). We also did not evaluate the application of multimodal vision-language models trained on radiologic reports and images. Compared with text-based analysis alone, multimodality could improve the results and should be investigated in future studies (32,33).

Fourth, we investigated only report content extracted from chest radiography examinations; therefore, the generalizability of the results to other cross-sectional imaging modalities or to other promising applications of LLMs in the areas of medical writing, clinical decision making, education, and data analysis should be investigated in future studies (34–36). Also, future work could compare the potential benefits of fine-tuning closed LLMs on external servers to those of fine-tuning open-weights LLMs within secured networks on more extensive public datasets.

Finally, as this is an exploratory study, we did not investigate the statistical significance of differences between models, which would also require accounting for multiple testing. The results should be validated in further hypothesis-based studies.

In conclusion, privacy-ensuring open-weights large language models (LLMs) have great potential for more efficient

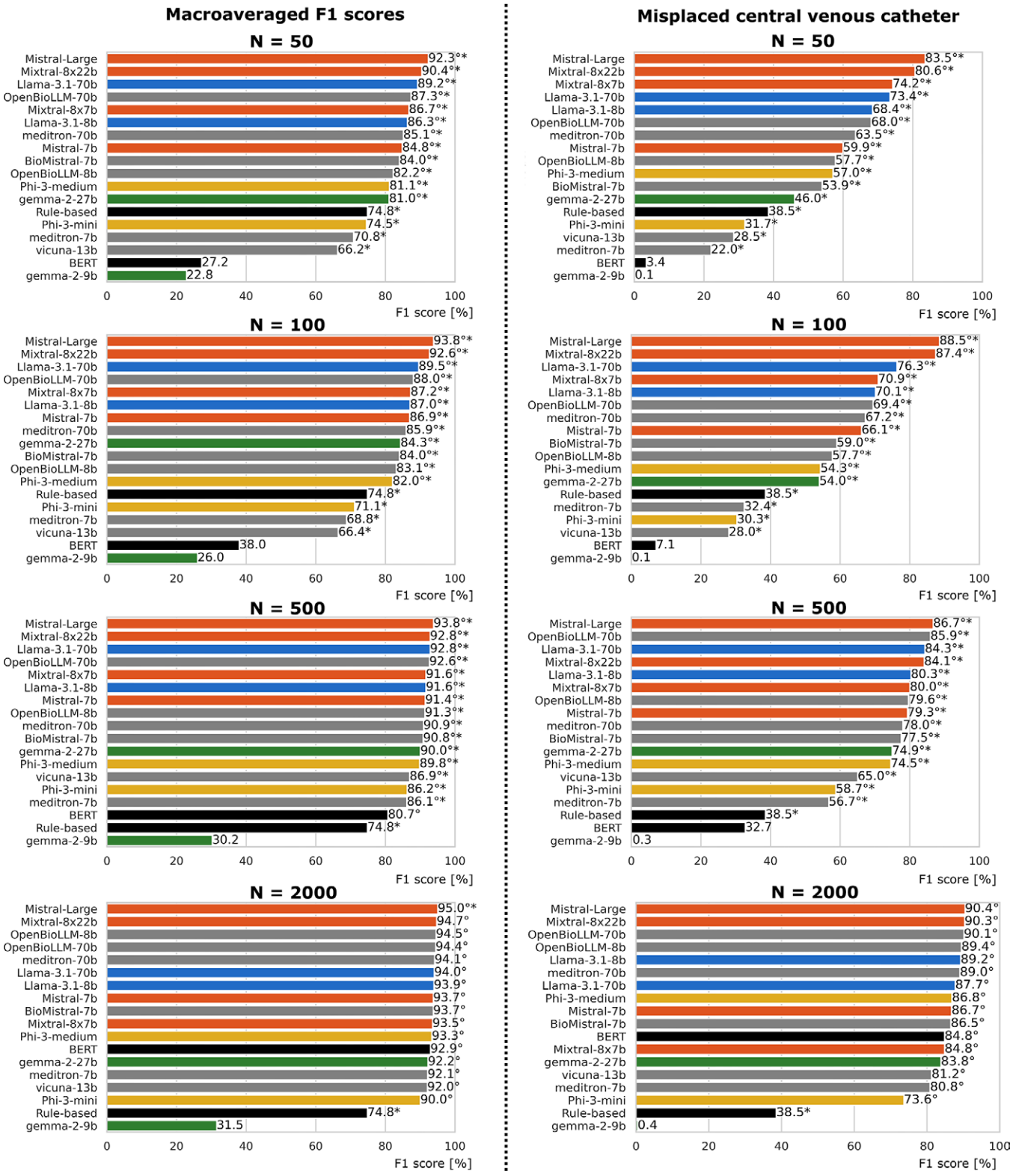


Figure 5: Bar plots of the results of fine-tuning open-weights large language models on nonpublic German reports. ° F1 scores with nonoverlapping 95% CIs exceeding those of the rule-based system. * F1 scores with nonoverlapping 95% CIs exceeding BERT. Only open-weights models were applied, as sharing reports with third parties by using closed-weights OpenAI models was not possible due to data protection.

"out-of-the-box" zero- or one-shot structuring of report databases compared with rule-based systems, especially for extracting more challenging text content. In addition, the investigated LLMs were particularly efficient compared with bidirectional encoder representations from transformers when fine-tuned with small amounts of manually annotated data.

Deputy Editor: Linda Moy
Scientific Editor: Shannyn Wolfe (AJE)

Acknowledgment: We want to thank Leonie Weinhold, PhD, for her valuable advice with respect to the statistical analysis of this exploratory study.

Author contributions: Guarantors of integrity of entire study, **S.N., A.M.S.**; study concepts/study design or data acquisition or data analysis/interpretation, all authors;

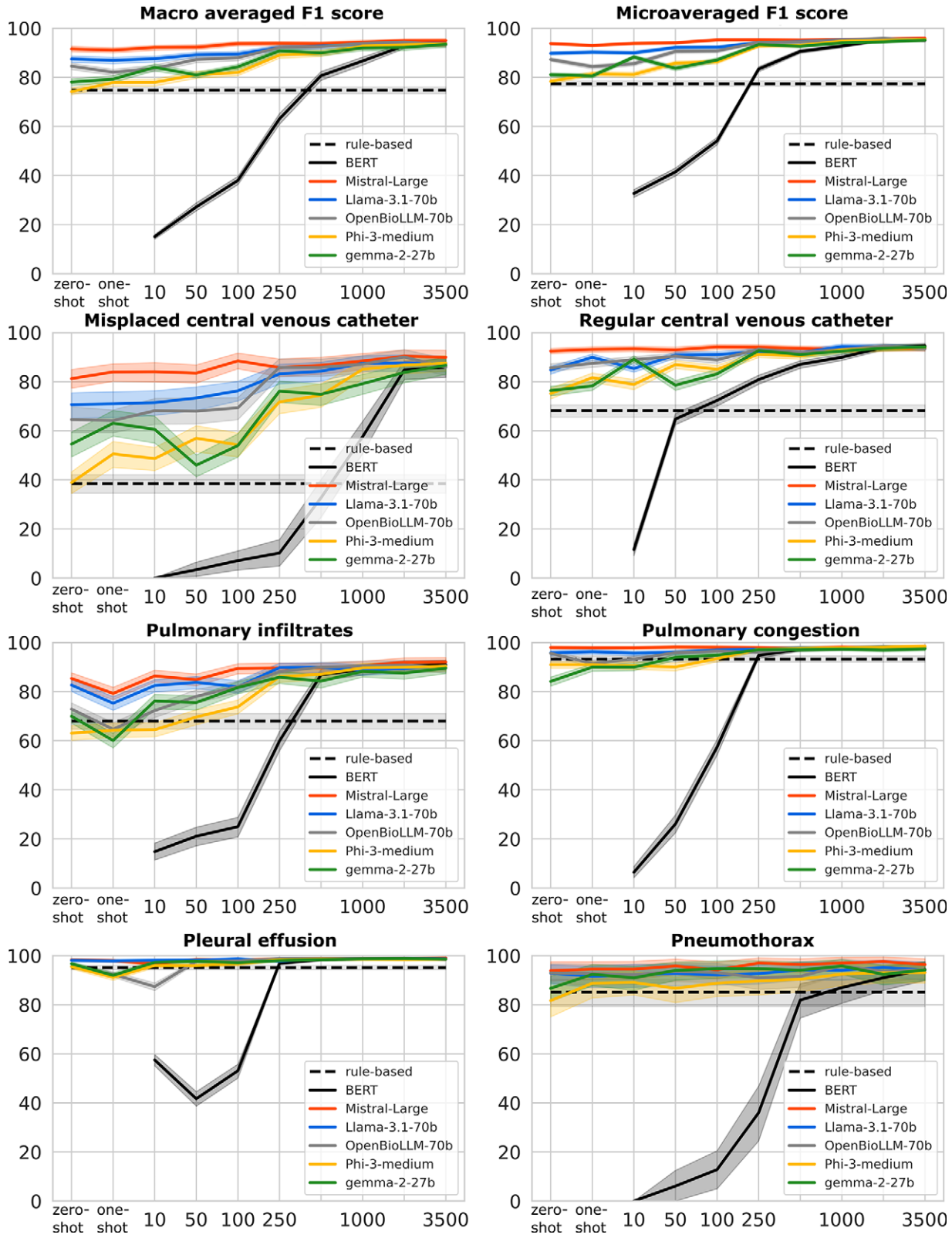


Figure 6: Line plots for the results of the zero- and one-shot prompting and the fine-tuning open-weights large language models on the nonpublic German reports. F1 scores are presented as percentages (y-axis) for zero- and one-shot prompting and for different numbers of manually annotated reports used (x-axis) for model development.

manuscript drafting or manuscript revision for important intellectual content, all authors; approval of final version of submitted manuscript, all authors; agrees to ensure any questions related to the work are appropriately resolved, all authors; literature research, **S.N., Y.C.L., A.I., D.K.**; clinical studies, **Y.C.L., A.I., W.B., D.K., C.C.P., J.A.L.**; experimental studies, **S.N.**; statistical analysis, **S.N., B.W., W.B.**; and manuscript editing, all authors

Data sharing: Data generated or analyzed during the study are available from the corresponding author by request.

Disclosures of conflicts of interest: **S.N.** Funded by the state of North Rhine-Westphalia (SIM-1-1). **B.W.** No relevant relationships. **Y.C.L.** Research grant from Siemens Healthcare. **M.T.** Funded by RACOON (NUM 2.0), which is supported

by the Federal Ministry of Education and Research of Germany (grant 01KX2121). **A.I.** Grants from the BONFOR Research Commission of the Medical Faculty Bonn and German Research Foundation under Germany's Excellence Strategy–EXC2151–390873048-Ernst und Berta Grimmke-Stiftung-Lfd. Nr.: 6/23; payment or honoraria for lectures, presentations, speakers bureaus, manuscript writing or educational events from Deutsche Röntgengesellschaft e.V.-C.T.I. GmbH. **B.S.** No relevant relationships. **W.B.** No relevant relationships. **D.K.** No relevant relationships. **C.C.P.** Speakers bureau payments from Guerbet and Julius Zorn. **J.A.L.** Received payments for activities related to the scientific advisory board for Bayer Healthcare; received payments for lectures from Bayer Healthcare, GE HealthCare, Novartis, Philips Healthcare, and Siemens Healthineers. **U.A.** No relevant relationships. **A.M.S.** No relevant relationships.

References

1. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med 2023;29(8):1930–1940.
2. Nobel JM, Kok EM, Robben SGF. Redefining the structure of structured reporting in radiology. Insights Imaging 2020;11(1):10.
3. European Society of Radiology (ESR). ESR paper on structured reporting in radiology. Insights Imaging 2018;9(1):1–7.
4. Jorg T, Halfmann MC, Arnhold G, et al. Implementation of structured reporting in clinical routine: a review of 7 years of institutional experience. Insights Imaging 2023;14(1):61.
5. Nowak S, Schneider H, Layer YC, et al. Development of image-based decision support systems utilizing information extracted from radiological free-text report databases with text-based transformers. Eur Radiol 2024;34(5):2895–2904.
6. Wornow M, Xu Y, Thapa R, et al. The shaky foundations of large language models and foundation models for electronic health records. NPJ Digit Med 2023;6(1):135.
7. Irvin J, Rajpurkar P, Ko M, et al. CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison. arXiv 1901.07031 [preprint] https://doi.org/10.48550/arXiv.1901.07031. Posted January 21, 2019. Accessed May 2024.
8. Nowak S, Biesner D, Layer YC, et al. Transformer-based structuring of free-text radiology report databases. Eur Radiol 2023;33(6):4228–4236.
9. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv 1810.04805 [preprint] https://doi.org/10.48550/arXiv.1810.04805. Posted October 11, 2018. Accessed May 2024.
10. ChatGPT. OpenAI. https://chat.openai.com. Accessed May 2024.
11. Brown T, Mann B, Ryder N, et al. Language models are few-shot learners. Adv Neural Inf Process Syst 2020;33:1877–1901.
12. Jiang AQ, Sablayrolles A, Roux A, et al. Mixtral of Experts. arXiv 2401.04088 [preprint] https://doi.org/10.48550/arXiv.2401.04088. Posted January 8, 2024. Accessed May 2024.
13. Jiang AQ, Sablayrolles A, Mensch A, et al. Mistral 7B. arXiv 2310.06825 [preprint] https://doi.org/10.48550/arXiv.2310.06825. Posted October 10, 2023. Accessed May 2024.
14. Huang W, Ma X, Qin H, et al. How Good Are Low-bit Quantized LLaMA3 Models? An Empirical Study. arXiv 2404.14047 [preprint] https://doi.org/10.48550/arXiv.2404.14047. Posted April 22, 2024. Accessed May 2024.
15. Abdin M, Jacobs SA, Awan AA, et al. Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone. arXiv 2404.14219 [preprint] https://doi.org/10.48550/arXiv.2404.14219. Posted April 22, 2024. Accessed May 2024.
16. Gemma Team; Mesnard T, Hardin C, et al. Gemma. Kaggle. https://doi.org/10.34740/KAGGLE/M/3301. Posted February 20, 2024. Accessed May 2024.
17. Hendrycks D, Burns C, Basart S, et al. Measuring Massive Multitask Language Understanding. arXiv 2009.03300 [preprint] https://doi.org/10.48550/arXiv.2009.03300. Posted September 7, 2020. Accessed May 2024.
18. Zhang H, Da J, Lee D, et al. A Careful Examination of Large Language Model Performance on Grade School Arithmetic. arXiv 2405.00332 [preprint] https://doi.org/10.48550/arXiv.2405.00332. Posted May 1, 2024. Accessed May 2024.
19. Demner-Fushman D, Kohli MD, Rosenman MB, et al. Preparing a collection of radiology examinations for distribution and retrieval. J Am Med Inform Assoc 2016;23(2):304–310.
20. OpenAI, Achiam J, Adler S, et al. GPT-4 Technical Report. arXiv 2303.08774 [preprint] https://doi.org/10.48550/arXiv.2303.08774. Posted March 15, 2023. Accessed May 2024.
21. Mukherjee P, Hou B, Lanfredi RB, Summers RM. Feasibility of Using the Privacy-preserving Large Language Model Vicuna for Labeling Radiology Reports. Radiology 2023;309(1):e231147.
22. Labrak Y, Bazoge A, Morin E, Gourraud PA, Rouvier M, Dufour R. BioMistral: A Collection of Open-Source Pretrained Large Language Models for Medical Domains. arXiv 2402.10373 [preprint] https://doi.org/10.48550/arXiv.2402.10373. Posted February 15, 2024. Accessed May 2024.
23. Chen Z, Cano AH, Romanou A, et al. MEDITRON-70B: Scaling Medical Pretraining for Large Language Models. arXiv 2311.16079 [preprint] https://doi.org/10.48550/arXiv.2311.16079. Posted November 27, 2023. Accessed May 2024.
24. Dettmers T, Pagnoni A, Holtzman A, Zettlemoyer L. QLoRA: Efficient Finetuning of Quantized LLMs. arXiv 2305.14314 [preprint] https://doi.org/10.48550/arXiv.2305.14314. Posted May 23, 2023. Accessed May 2024.
25. Hu EJ, Shen Y, Wallis P, et al. LoRA: Low-Rank Adaptation of Large Language Models. arXiv 2106.09685 [preprint] https://doi.org/10.48550/arXiv.2106.09685. Posted June 17, 2021. Accessed May 2024.
26. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine Learning in Python. arXiv 1201.0490 [preprint] https://doi.org/10.48550/arXiv.1201.0490. Posted January 2, 2012. Accessed May 2024.
27. Wan X, Sun R, Dai H, Arik SO, Pfister T. Better Zero-Shot Reasoning with Self-Adaptive Prompting. arXiv 2305.14106 [preprint] https://doi.org/10.48550/arXiv.2305.14106. Posted May 23, 2023. Accessed May 2024.
28. Liu NF, Lin K, Hewitt J, et al. Lost in the Middle: How Language Models Use Long Contexts. arXiv 2307.03172 [preprint] https://doi.org/10.48550/arXiv.2307.03172. Posted July 6, 2023. Accessed May 2024.
29. Wei J, Wang X, Schuurmans D, et al. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. arXiv 2201.11903 [preprint] https://doi.org/10.48550/arXiv.2201.11903. Posted January 28, 2022. Accessed May 2024.
30. Wu Q, Bansal G, Zhang J, et al. AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation. arXiv 2308.08155 [preprint] https://doi.org/10.48550/arXiv.2308.08155. Posted August 16, 2023. Accessed May 2024.
31. Rau A, Rau S, Zoeller D, et al. A Context-based Chatbot Surpasses Radiologists and Generic ChatGPT in Following the ACR Appropriateness Guidelines. Radiology 2023;308(1):e230970.
32. Bannur S, Hyland S, Liu Q, et al. Learning to Exploit Temporal Structure for Biomedical Vision-Language Processing. arXiv 2301.04558 [preprint] https://doi.org/10.48550/arXiv.2301.04558. Posted January 11, 2023. Accessed May 2024.
33. Chen Z, Varma M, Delbrouck JB, et al. CheXagent: Towards a Foundation Model for Chest X-Ray Interpretation. arXiv 2401.12208 [preprint] https://doi.org/10.48550/arXiv.2401.12208. Posted January 22, 2024. Accessed May 2024.
34. Shen Y, Heacock L, Elias J, et al. ChatGPT and Other Large Language Models Are Double-edged Swords. Radiology 2023;307(2):e230163.
35. Biswas S. ChatGPT and the Future of Medical Writing. Radiology 2023;307(2):e223312.
36. Salam B, Kravchenko D, Nowak S, et al. Generative Pre-trained Transformer 4 makes cardiovascular magnetic resonance reports easy to understand. J Cardiovasc Magn Reson 2024;26(1):101035.

3.2 **Nowak S***, Biesner D*, Layer Y, Theis M, Schneider H, Block W, Wulff B, Attenberger UI*, Sifa R*, Sprinkart AM*. Transformer-based structuring of free-text radiology report databases. European Radiology. 2023;33(6):4228–4236. DOI: 10.1007/s00330-023-09526-y

Objectives

To provide insights for on-site development of transformer-based structuring of free-text report databases by investigating different labeling and pre-training strategies.

Methods

A total of 93,368 German chest X-ray reports from 20,912 intensive care unit (ICU) patients were included. Two labeling strategies were investigated to tag six findings of the attending radiologist. First, a system based on human-defined rules was applied for annotation of all reports (termed "silver labels"). Second, 18,000 reports were manually annotated in 197 h (termed "gold labels") of which 10% were used for testing. An on-site pre-trained model (T_{mlm}) using masked-language modeling (MLM) was compared to a public, medically pre-trained model (T_{med}). Both models were fine-tuned on silver labels only, gold labels only, and first with silver and then gold labels (hybrid training) for text classification, using varying numbers (N: 500, 1000, 2000, 3500, 7000, 14,580) of gold labels. Macro-averaged F1-scores (MAF1) in percent were calculated with 95% confidence intervals (CI).

Results

$T_{mlm,gold}$ (95.5 [94.5–96.3]) showed significantly higher MAF1 than $T_{med,silver}$ (75.0 [73.4–76.5]) and $T_{mlm,silver}$ (75.2 [73.6–76.7]), but not significantly higher MAF1 than $T_{med,gold}$ (94.7 [93.6–95.6]), $T_{med,hybrid}$ (94.9 [93.9–95.8]), and $T_{mlm,hybrid}$ (95.2 [94.3–96.0]). When using 7000 or less gold-labeled reports, $T_{mlm,gold}$ (N: 7000, 94.7 [93.5–95.7]) showed significantly higher MAF1 than $T_{med,gold}$ (N: 7000, 91.5 [90.0–92.8]). With at least 2000 gold-labeled reports, utilizing silver labels did not lead to significant improvement of $T_{mlm,hybrid}$ (N: 2000, 91.8 [90.4–93.2]) over $T_{mlm,gold}$ (N: 2000, 91.4 [89.9–92.8]).

Conclusions

Custom pre-training of transformers and fine-tuning on manual annotations promises to be an efficient strategy to unlock report databases for data-driven medicine.

European Radiology (2023) 33:4228–4236
https://doi.org/10.1007/s00330-023-09526-y

European Radiology
EUROPEAN SOCIETY OF RADIOLOGY

IMAGING INFORMATICS AND ARTIFICIAL INTELLIGENCE

Transformer-based structuring of free-text radiology report databases

S. wak[1] · D. Biesner[2] · Y. C. Layer[1] · M. Theis[1] · H. Schneider[2] · W. Block[1] · B. Wulff[2] · U. I. Attenberger[1] · R. Sifa[2] · A. M. Spri art[1]

Received: 16 September 2022 / Revised: 5 January 2023 / Accepted: 3 February 2023 / Published online: 11 March 2023

Abstract

Objectives To provide insights for on-site development of transformer-based structuring of free-text report databases by investigating different labeling and pre-training strategies.

Methods A total of 93,368 German chest X-ray reports from 20,912 intensive care unit (ICU) patients were included. Two labeling strategies were investigated to tag six findings of the attending radiologist. First, a system based on human-defined rules was applied for annotation of all reports (termed "silver labels"). Second, 18,000 reports were manually annotated in 197 h (termed "gold labels") of which 10% were used for testing. An on-site pre-trained model (T_{mlm}) using masked-language modeling (MLM) was compared to a public, medically pre-trained model (T_{med}). Both models were fine-tuned on silver labels only, gold labels only, and first with silver and then gold labels (hybrid training) for text classification, using varying numbers (*N*: 500, 1000, 2000, 3500, 7000, 14,580) of gold labels. Macro-averaged F1-scores (MAF1) in percent were calculated with 95% confidence intervals (CI).

Results $T_{mlm,gold}$ (95.5 [94.5–96.3]) showed significantly higher MAF1 than $T_{med,silver}$ (75.0 [73.4–76.5]) and $T_{mlm,silver}$ (75.2 [73.6–76.7]), but not significantly higher MAF1 than $T_{med,gold}$ (94.7 [93.6–95.6]), $T_{med,hybrid}$ (94.9 [93.9–95.8]), and $T_{mlm,hybrid}$ (95.2 [94.3–96.0]). When using 7000 or less gold-labeled reports, $T_{mlm,gold}$ (*N*: 7000, 94.7 [93.5–95.7]) showed significantly higher MAF1 than $T_{med,gold}$ (*N*: 7000, 91.5 [90.0–92.8]). With at least 2000 gold-labeled reports, utilizing silver labels did not lead to significant improvement of $T_{mlm,hybrid}$ (*N*: 2000, 91.8 [90.4–93.2]) over $T_{mlm,gold}$ (*N*: 2000, 91.4 [89.9–92.8]).

Conclusions Custom pre-training of transformers and fine-tuning on manual annotations promises to be an efficient strategy to unlock report databases for data-driven medicine.

Key Points

- *On-site development of natural language processing methods that retrospectively unlock free-text databases of radiology clinics for data-driven medicine is of great interest.*
- *For clinics seeking to develop methods on-site for retrospective structuring of a report database of a certain department, it remains unclear which of previously proposed strategies for labeling reports and pre-training models is the most appropriate in context of, e.g., available annotator time.*
- *Using a custom pre-trained transformer model, along with a little annotation effort, promises to be an efficient way to retrospectively structure radiological databases, even if not millions of reports are available for pre-training.*

Keywords Radiology · Deep learning · Natural language processing · Intensive care units · Thorax

Abbreviations

CI	Confidence interval
CVC	Central venous catheter
ICU	Intensive care unit
MAAUC	Macro-averaged area under the receiver operating characteristic curve
MAF1	Macro-averaged F1-score
ML	Machine learning
MLM	Masked-language modeling
NLP	Natural language processing
TFIDF	Term frequency–inverse document frequency

S. Nowak and D. Biesner contributed equally as joint first authors. U.I. Attenberger, R. Sifa, and A.M. Sprinkart contributed equally as joint last authors.

✉ S Nowak
sebastian.nowak@ukbonn.de

[1] Department of Diagnostic and Interventional Radiology, University Hospital Bonn, Venusberg-Campus 1, 53127 Bonn, Germany

[2] Fraunhofer Institute for Intelligent Analysis and Information Systems IAIS, Sankt Augustin, Germany

Introduction

Structured reporting, i.e., the use of IT-based systems for importing and arranging medical content in radiological reports, not only has the potential to have a positive impact on patient care by enhancing the quality of radiologists' practices, it is also beneficial for the development of image-based artificial intelligence systems by helping to compile large retrospective patient collectives with diseases of interest [1, 2]. To this day, most radiological reports are in free-text and not in structured format. Even if a clinic would introduce structured reporting, retrospective assembling of large image collectives is labor-intensive, as the corresponding reports of recent years remain unstructured. Therefore, there is a need for automatic natural language processing (NLP) systems that categorize free-text reports in a set of predefined labels, thereby unlocking the corresponding image database for the development of an artificial intelligence–based diagnostic decision system.

To achieve automated analysis of medical text data, various methods with different levels of complexity have been proposed. For example, simple systems based on human-defined rules have been applied to automatically annotate the occurrence of findings in English chest X-ray reports [3]. These rule-based systems have the advantage that the method itself does not require any manually annotated reports. However, the creator of such a system must have considerable expert knowledge about general information, content, and wording of the reports. Moreover, there may still be findings whose appearance and description are subject to great variability, making the establishment of comprehensive rules a difficult task.

On the other hand, there are machine learning (ML)–based methods that have the disadvantage of requiring a large amount of time-consuming manual annotated reports for training. In recent years, transformer models based on the self-attention mechanism have emerged as the state-of-the-art ML–based NLP method, also for medical text data [4–9]. The required amount of annotated data to train a transformer can be reduced by transfer learning, i.e., utilizing the fundamental text comprehension skills of a model that has already been pre-trained on different large public datasets and/or for another task.

In contrast to radiological image datasets, where the appearance of a finding, e.g., a pneumothorax, is independent of the country, the description of the finding in a radiological report can differ substantially, which is most obvious if the countries do not share a common language. For NLP, this limits the development and application of pre-trained models compared to computer vision applications, as, e.g., a model pre-trained on English reports cannot be directly applied to German texts. Moreover, unlike cross-sectional radiological images, text-based medical data contain sensitive information directly linked to personal data. This makes the public sharing of medical text data in compliance with data protection laws highly problematic in many countries [10]. Even when sharing pre-trained parameters of a transformer model trained on sensible data, it also cannot be ensured that data protection laws are met, as it has been shown that information from the training data can be extracted from large pre-trained transformers [11]. There are efforts to automate the de-identification of German text data using ML methods, from medical and other domains. However, currently, these methods cannot guarantee 100% accuracy [10, 12]. Consequently, efficient development of NLP models for structuring radiological reports on-site is of great interest.

Several approaches have already been presented for the on-site development of transformer models based on radiological text data. Two studies that are based on several hundred thousand English-language reports propose to employ a publicly available transformer pre-trained on medical text and to fine-tune that model in a two-step hybrid label approach. With the hybrid approach, the pre-trained model is first adapted to a high number of rule-based annotated text (termed "silver labels") and then to only a very limited number of manual annotations (termed "gold labels") [7, 8]. In contrast, another study proposes a custom pre-training of the transformer by MLM and next-sentence prediction using millions of radiology reports and subsequent fine-tuning to only a few gold-labeled reports [9].

However, for clinics seeking to develop NLP models on-site, it remains unclear which of those pre-training and labeling strategies are most appropriate for structuring their free-text radiological report database. First, it is not clear if a custom pre-training of transformer models is also beneficial with a significantly lower number of reports than in above-mentioned studies. Second, it is not clear whether the effort required to create a rule-based system for silver label generation in hybrid training is worthwhile compared to utilizing more gold labels by investing more annotator time. Therefore, the goal of this work is to provide insight and guidance for efficient retrospective structuring of radiological databases by systematically evaluating the performance of publicly available and custom pre-trained text-based transformers with respect to different labeling strategies and human annotation effort.

Materials and methods

Dataset and annotation

With institutional review board approval (AZ 411/21), written informed consent was waived, and approved data processing took place on the basis of the Health Data Protection Act North Rhine-Westphalia (GDSG NW) §6 (2) state law NRW. The retrospective dataset includes 93,368

free-text chest X-ray reports of 20,912 ICU patients (age: 62.7 ± 21.4, 8081 female) from University Hospital Bonn that were extracted consecutively from the radiological information system dating between December 2015 and July 2021.

First, 35 labels including not only findings but also further information, e.g., on indication, were defined for systematic annotation of the reports (see Supplement S1). Information on the interpretation of findings was not assessed from the reports. Under the supervision of a radiology resident (Y.C.L.), two medical research assistants labeled this information using the open-source software doccano [13]. The medical research assistants were trained to assign the correct labels based on the context of the free-text reports and not just on individual words. In case of ambiguity, they were instructed to consult the supervisor. Manually annotated gold labels were curated for 18,000 reports including only reports with unique admission numbers. A subset of six findings that are frequently raised during a patient's ICU stay were selected from the entire label set. This selection was based on frequency of the finding and their clinical relevance and was made prior to NLP development. The NLP models were developed to predict the occurrence of these labels based on the report text via multi-label classification. These labels and their relative appearances within the gold-labeled reports are pulmonary infiltrates (20.0%), pleural effusion (45.6%), pulmonary congestion (34.0%), pneumothorax (3.8%), regular position of the central venous catheter (CVC) (45.8%), and misplaced position of the CVC (8.4%).

The 18,000 gold-labeled reports were randomly split into training set A (14,580), validation set B (1620), and test set C (1800). Additional 500 reports (test set D) were labeled by the radiology resident and both research assistants independently to determine the agreement between the annotators. This dataset was also used for the final test of the best NLP model with annotations of the radiology resident. Moreover, silver labels were created for a total of 91,068 reports applying a rule-based model which is described below. These are all available reports except the 1800 and 500 texts of the gold-labeled hold-out test sets C and D. Silver-labeled data were split into training set E (81,961) and validation set F (9107). Figure 1 shows an overview of the entire study.

Rule-based model

A set of rules was defined to automatically annotate the free-text radiological reports. In short, the algorithm searches for specific terms, negations, and descriptions of uncertainty or applies further text-based rules in the "findings" section of the report. Detailed descriptions of the rule-based system can be found in Supplement S2.

Baseline NLP model

As a baseline NLP approach, we trained a term frequency–inverse document frequency (TFIDF) model on the training text and fitted a one-layer fully connected neural network to the labeled training data [14]. Training details can be found in Supplement S3.

Transformer-based model

We applied BERT as an established transformer model that has also been used in other work on medical text analysis [5, 7–9]. See Supplement S4 for details on the model architecture.

To investigate the impact of different pre-training strategies, we employed (i) a publicly available BERT language model that was pre-trained on (not annotated) German legal documents, Wikipedia, and news articles and then further adapted to medical articles and texts scraped from the web (T_{med}) [15–17] and (ii) created a custom pre-trained BERT language model (T_{mlm}) by applying MLM on the texts of train set E. To demonstrate the general effect of pre-training, we also trained a model from scratch for classification on gold-labeled text data without any pre-training ($T_{rand,gold}$).

To examine the difference between different label strategies, three experiments were performed with the two pre-trained models. First, the pre-trained models were fine-tuned on the 14,580 gold labels of training set A only ($T_{med,gold}$, $T_{mlm,gold}$). Second, both pre-trained models were fine-tuned on silver-labeled training set E ($T_{med,silver}$, $T_{mlm,silver}$). Third, the models fine-tuned on silver-labeled training set E ($T_{med,silver}$, $T_{mlm,silver}$) were subsequently fine-tuned on training set A in a hybrid training ($T_{med,hybrid}$, $T_{mlm,hybrid}$). To investigate the effect of the number of available gold labels for fine-tuning, the models were also trained with limited numbers of gold-labeled reports of train set A (500, 1000, 2000, 3500, 7000 reports). All models were tested on test set C, and the best model was additionally tested on test set D.

When fine-tuning the models for text classification, we applied the following concepts. As proposed in previous studies, we fine-tuned all pre-trained models for text classification in two steps: First, frozen pre-trained language model parameters were used to adapt the new classification head and then all parameters were trained, but with layer-specific learning rates with maximum values increasing linearly from 10^{-9} to 10^{-6} from the first to the last layer [18–20]. Since the threshold for binarization of the predictions after sigmoid activation is not intrinsically set in multi-label classification, class-specific thresholds were determined by identifying the thresholds with the highest F1-scores on the training data [21]. Also, oversampling and loss weighting according to the occurrence of the classes within the training data were used to

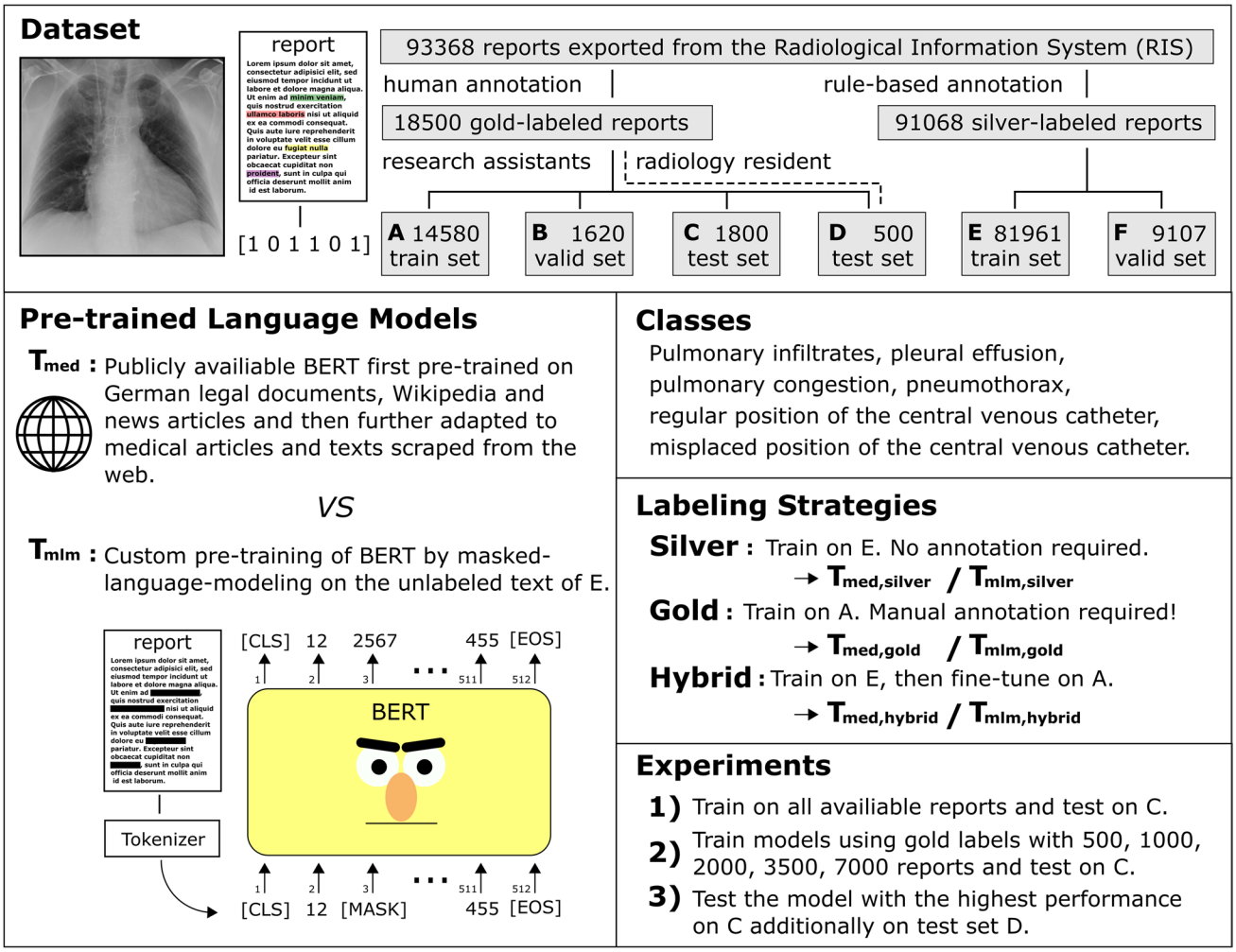


Fig. 1 Overview of the presented study. The dataset of the presented study includes a total of 93,368 free-text chest X-ray reports of intensive care unit patients. For a subset of the dataset, human annotations were generated for the occurrence of six findings within the reports to create gold-labeled training, validation, and test datasets. Furthermore, a rule-based system was applied for silver label generation. The use of an on-site pre-trained model using masked-language modeling (T_{mlm}) was compared to a public, medically pre-trained model (T_{med}) when adapting to silver labels only, gold labels only, and to first with silver and then gold labels (hybrid). To also give insights into which pre-training and labeling strategy is most appropriate in context of available human annotation time, the models were developed using varying numbers of gold-labeled reports

Table 1 Model performances for different pre-training and labeling strategies using all training data

			Silver		Gold				Hybrid	
Class	SP	RB	T_{med}	T_{mlm}	TFIDF	T_{rand}	T_{med}	T_{mlm}	T_{med}	T_{mlm}
Infiltrates	352	69.7	69.3	69.2	79.8	80.3	92.9	92.9	**93.6**	92.0
Congestion	611	94.3	94.1	94.2	88.7	88.2	**98.1**	**98.1**	**98.1**	97.9
Effusion	818	95.0	94.6	94.7	88.7	91.0	**98.8**	**98.8**	**98.8**	**98.8**
Pneumothorax	65	87.8	87.1	87.0	75.2	79.7	96.1	96.0	**98.5**	98.4
Regular CVC	825	67.0	67.8	67.9	89.8	90.7	93.4	**95.4**	94.9	95.0
Misplaced CVC	151	36.9	37.3	38.1	77.2	75.7	88.8	**91.7**	85.6	89.3
Macro average	2822	75.1	75.0	75.2	83.2	84.3	94.7	**95.5**	94.9	95.2
Micro average	2822	77.3	77.3	77.5	87.1	87.7	95.7	**96.5**	96.1	96.1

F1-scores (%) observed for the hold-out test set of 1800 gold-labeled reports for the rule-based (RB) system, the TFIDF approach and the transformer models trained with all 14,580 gold-labeled training data

The support (SP), i.e., the number of positive samples, is given for each class

For each class, the highest F1-scores are indicated in bold

compensate for class imbalance. Classes that occurred in less than 25% of the training reports were duplicated until they accounted for at least 25%.

The pre-trained models fine-tuned for text classification on 14,580 and 7000 gold-labeled reports were trained for 75 epochs. Models that were fine-tuned on less than 7000 reports were trained with the same amount of optimization steps as the models trained on 7000 reports to ensure convergence. The pre-trained models that were fine-tuned on 81,961 silver-labeled reports were trained for 25 epochs.

For the custom pre-training via MLM, first, the BERT model was trained on 81,961 reports with a maximum learning rate of 10^{-4} and 15% of tokens masked for 150 epochs. After that, the model was further trained for 150 epochs with a maximum learning rate of 10^{-5} and 15% of whole words masked within the text. For custom pre-training, no weight decay was used.

For all models, the "bert-base-german-cased" tokenizer of the Huggingface's transformer library was used and all models were trained using the Adam optimizer with decoupled weight decay regularization of 0.01, a learning rate

Table 2 Model performances for different pre-training and labeling strategies using different numbers of gold-labeled training data

N		Gold			Hybrid		Gold			Hybrid	
		TFIDF	T_{med}	T_{mlm}	T_{med}	T_{mlm}	TFIDF	T_{med}	T_{mlm}	T_{med}	T_{mlm}
		Macro averaged					Micro averaged				
500	(3.4%)	34.9	59.8	70.9*	86.9†	**90.4***†	57.9	81.6	85.0*	92.7†	**93.9**†
1000	(6.9%)	44.7	64.5	85.6*	**91.5**†	87.6	66.3	82.3	91.4*	**94.4**†	92.6
2000	(13.7%)	58.2	84.1	91.4*	89.1	**91.8**	74.3	91.6	94.0*	93.0	**94.3***
3500	(24.0%)	68.6	88.5	**93.5***	91.6	93.0	79.6	93.3	**95.6***	94.1	95.1
7000	(48.0%)	77.4	91.5	**94.7***	92.1	94.1	84.3	94.4	**96.2***	94.4	95.6*
14,580	(100%)	83.2	94.7	**95.5**	94.9	95.2	87.1	95.7	**96.5**	96.1	96.1
		Misplaced CVC					Congestion				
500	(3.4%)	2.6	33.2	33.3	53.4†	**73.6***†	37.4	87.5	94.2*	**97.6**†	97.5†
1000	(6.9%)	16.8	44.0	65.4*	77.1	**77.4**†	55.0	90.2	97.5*	**97.9**	95.9
2000	(13.7%)	45.0	66.7	81.8*	72.2	**82.9***	68.6	97.1	**98.1**	97.2	97.1
3500	(24.0%)	54.3	74.7	**87.0***	78.0	85.7	76.9	97.8	**98.3**	97.1	97.0
7000	(48.0%)	68.2	80.4	**88.9***	80.9	88.1	83.2	98.4	**98.6**	97.9	97.9
14,580	(100%)	77.2	88.8	**91.7**	85.6	89.3	88.7	**98.1**	**98.1**	**98.1**	97.9
		Regular CVC					Effusion				
500	(3.4%)	77.2	85.6	81.7	90.9†	**92.2**†	68.2	88.8	92.1*	**98.3**†	97.9†
1000	(6.9%)	80.2	83.9	89.2*	**92.7**†	91.9	76.4	89.4	96.5*	**98.5**†	95.8
2000	(13.7%)	85.2	89.9	92.5	92.4	**92.8**	80.4	97.5	97.1	97.9	**98.0**
3500	(24.0%)	87.2	90.9	**94.6***	92.3	94.2	83.7	**98.5**	98.3	98.4	98.1
7000	(48.0%)	88.7	91.7	**95.0***	93.1	94.5	87.7	**98.8**	98.7	98.2	98.6
14,580	(100%)	89.8	93.4	**95.4**	94.9	95.0	88.7	**98.8**	**98.8**	**98.8**	**98.8**
		Infiltrates					Pneumothorax				
500	(3.4%)	24.2	63.6	79.4*	88.0†	**90.3**†	0.0	0.0	44.9*	**93.0**†	91.2†
1000	(6.9%)	36.6	62.3	85.4*	**90.6**†	89.9	3.0	16.9	79.3*	**92.2**†	74.8
2000	(13.7%)	53.0	84.7	89.1	86.5	**90.2**	16.7	68.7	89.8*	88.7	**90.0**
3500	(24.0%)	67.9	88.0	91.2	90.5	**91.9**	41.5	81.0	91.7	**93.7**	90.9
7000	(48.0%)	77.2	90.7	**92.6**	89.4	91.3	59.6	89.1	**94.4**	93.1	**94.4**
14,580	(100%)	79.8	92.9	92.9	**93.6**	92.0	75.2	96.1	96.0	**98.5**	98.4

F1-scores (%) observed on all classes for the hold-out test set of 1800 gold-labeled reports for the experiments on training with the different numbers (*N*) of the 14,580 gold-labeled training reports. The highest F1-scores of a class at a given *N* are highlighted by bold font. Approximately 5.5 h of work was performed to annotate 500 reports

*Significantly higher F1-scores comparing to all models trained with the same label strategy (gold or hybrid), independent of the model (T_{med}, T_{mlm}, TFIDF)

†Significantly higher F1-scores of a hybrid or gold-trained model respectively compared to all models trained with the other label strategy

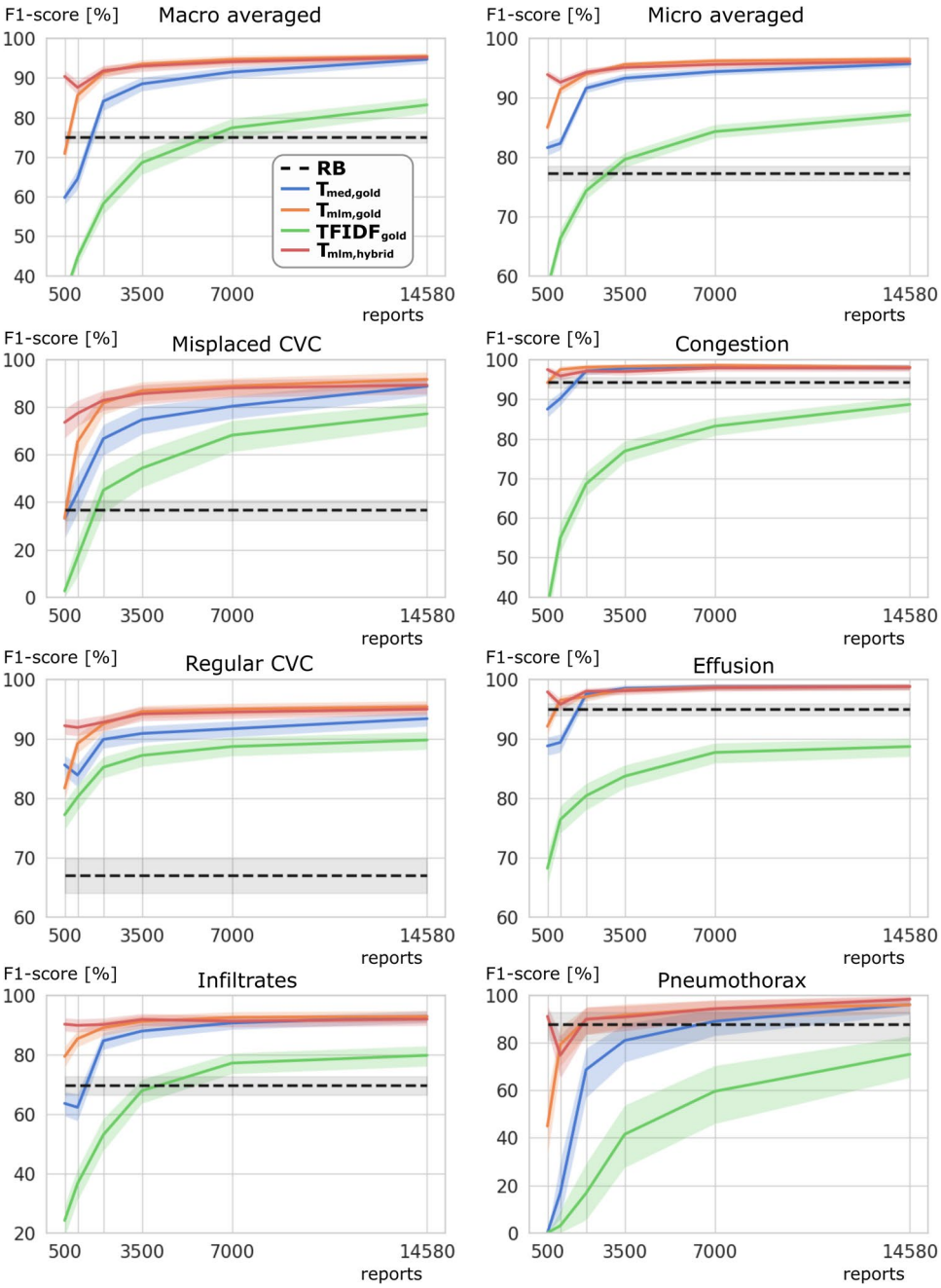


Fig. 2 Model performances for different numbers of gold-labeled reports. F1-scores in % (*y*-axis) are displayed for the rule-based (RB) system in black, as well as for $T_{med,gold}$ (blue), $T_{mlm,gold}$ (orange), $TFIDF_{gold}$ (green), and $T_{mlm,hybrid}$ (red) using various numbers of gold-labeled reports for training (*x*-axis)

scheme with warmup until 10% of all training steps and subsequent cosine decay, drop-out of 0.1, mixed precision, random seed 42, and a batch size of 24 on an NVIDIA RTX 3090 or NVIDIA TITAN RTX using PyTorch v1.8.1 and the transformers library v4.13.0 [15, 22]. The performance metrics were calculated using scikit-learn v0.24.2, and 95% CIs were calculated by bootstrapping with 1000 resamples for the text-classification models [23]. Performance differences were considered significant based on non-overlapping CIs.

Results

Time for manual annotation of 18,000 radiological reports was 197 h (39.4 s per report). The two medical research assistants' annotations showed high agreement with the radiology resident's annotations, with mean accuracy of 97.4% and 97.3% and MAF1 in percent of 92.9 and 93.5, respectively.

In the custom pre-training of BERT (T_{mlm}), an accuracy of 88.6% was achieved after 3.3 days of training to predict

15% masked tokens and subsequently an accuracy of 78.4% was achieved after again 3.3 days of training to predict 15% masked whole words on test data set C.

Table 1 shows the performance of all examined models trained with all available silver- and/or gold-labeled text classification data. The highest performance was observed for $T_{mlm,gold}$ with a MAF1 in percent of 95.5 (CI: 94.5–96.3), which was significantly higher than that of the rule-based system (75.1 [73.6–76.5]), $T_{med,silver}$ (75.0 [73.4–76.5]), $T_{mlm,silver}$ (75.2 [73.6–76.7]), $TFIDF_{gold}$ (83.2 [81.3–85.1]), and $T_{rand,gold}$ (84.3 [82.5–86.0]). However, the performance was not significantly higher than that of $T_{med,gold}$ (94.7 [93.6–95.6]), $T_{med,hybrid}$ (94.9 [93.9–95.8]), and $T_{mlm,hybrid}$ (95.2 [94.3–96.0]).

Table 2 and Fig. 2 show the performance of the examined models for all classes when using lower numbers of gold-labeled data. The classification of the description of a misplaced CVC was found to be the most challenging class with the lowest F1-scores for each method. Considering models trained exclusively with gold-labeled reports, $T_{mlm,gold}$ showed significantly higher MAF1 as well as misplaced CVC F1-scores than $TFIDF_{gold}$ and $T_{med,gold}$ when only 1000 to 7000 gold-labeled reports were provided for training. The hybrid models adapted on only 500 gold labels ($T_{mlm,hybrid}$ 90.4 [89.0–91.9], $T_{med,hybrid}$ 86.9 [85.1–88.5]) achieved already significantly higher MAF1 than the rule-based system (75.1 [73.6–76.5]). Considering both models trained with a hybrid label scheme, no significant differences in MAF1 and F1-scores for misplaced CVC were observed for $T_{med,hybrid}$ and $T_{mlm,hybrid}$ trained with 1000 or more gold labels. When using 2000 or more gold-labeled reports (22 h of annotation), the previous use of silver labels in hybrid training of the BERT models ($T_{mlm,hybrid}$ 91.8 [90.4–93.2], $T_{med,hybrid}$ 89.1 [87.6–90.6]) did not provide a significant improvement of MAF1 over the BERT models ($T_{mlm,gold}$ 91.4 [89.9–92.8], $T_{med,gold}$ 84.1 [81.7, 86.0]) trained directly on the gold-labeled reports. Due to less than 4% positive pneumothorax cases leading to wide CIs, F1-scores for pneumothorax of the rule-based system (87.8 [81.1–93.0]) are only significantly lower compared with $T_{mlm,hybrid}$ (98.4 [95.9–100.0]) and $T_{med,hybrid}$ (98.5 [96.0–100.0]) trained with all gold labels.

Table 3 shows further performance metrics for the best model, $T_{mlm,gold}$, which was pre-trained using MLM and fine-tuned to 14,580 gold labels that showed the highest MAF1 (95.5, CI: 94.5–96.3) and macro-averaged area under the receiver operating characteristic curve (MAAUC: 97.1, CI: 96.3–97.8) for test set C with 1800 cases, as well as for test set D with 500 cases (MAF1: 93.5, CI: 91.0–95.3; MAAUC: 95.7, CI: 93.8–97.1). Due to space limitations, CIs for each single value in Tables 1–3 can be found in Supplement S5.

Discussion

The current study investigates efficient on-site development of NLP methods in the context of different pre-training and labeling strategies for structuring and thus unlocking radiological databases for data-driven medicine using German ICU chest X-ray reports. The work provides clinics seeking to develop NLP models on-site with insights and guidance on which strategy is preferable for their specific project in the context of available annotator and developer time and the complexity of the information to be extracted. Methods for training transform-based NLP models will be provided upon reasonable request (https://qilab.de).

The results show that when training with a large set of silver labels without the use of gold labels, the pre-trained BERT models achieved comparable performance to the rule-based system and are limited by the quality of the silver labels. By using a publicly available, medically pre-trained BERT with a hybrid label approach that was first adapted on all silver labels and then fine-tuned on a small set of gold-labeled reports, significantly higher performance can

Table 3 Detailed model performance of $T_{mlm,gold}$ for both test sets

Class	1800 test set						500 test set					
	SP	AC	PR	RC	F1	AUC	SP	AC	PR	RC	F1	AUC
Infiltrates	352	97.2	91.2	94.6	92.9	96.2	105	94.6	86.1	88.6	87.3	92.4
Congestion	611	98.7	97.6	98.7	98.1	98.7	176	97.8	97.1	96.6	96.9	97.5
Effusion	818	98.9	98.5	99.0	98.8	98.9	241	99.4	100	98.8	99.4	99.4
Pneumothorax	65	99.7	100	92.3	96.0	96.2	20	99.4	100	85.0	91.9	92.5
Regular CVC	825	95.8	95.8	95.0	95.4	95.8	211	95.8	93.6	96.7	95.1	95.9
Misplaced CVC	151	98.6	88.8	94.7	91.7	96.8	50	98.0	87.0	94.0	90.4	96.2
Macro average	2822	98.1	95.3	95.7	95.5	97.1	803	97.5	94.0	93.3	93.5	95.7
Micro average	2822	97.7	96.1	96.8	96.5	97.4	803	97.4	95.1	95.8	95.4	96.8

F1-scores and AUC in % for each class on both test sets for $T_{mlm,gold}$ trained with all available data

The support (SP), i.e., the number of positive samples is given for each class and both test sets

AC accuracy, *RC* recall, *PR* precision

be achieved compared to the rule-based system. Both findings are in line with previous studies that used silver-labeled chest X-ray reports in the English language by CheXpert and further trained on 1000 manually curated reports or sentences [3, 7, 8]. However, when utilizing 2000 or more gold-labeled reports that were generated in only 22 h of annotation, the hybrid label approach did not provide a significant improvement over training the publicly available BERT model directly with the gold labels. The results of this work also show that the custom pre-training of BERT with only 81,961 reports can not only achieve high MLM accuracy compared to previous studies with millions of reports [9], but also demonstrate that this model achieves significantly higher performance than the publicly available pre-trained BERT model when further trained for report classification with 7000 or less gold labels.

The performance of the rule-based system that generates the silver labels varies strongly between the extracted classes. This can be explained by the fact that some information from the radiological reports have more variable attributes and are more difficult to identify by rules and standard formulations. In contrast to previous studies, we extracted information about a regular or misplaced position of the CVC from the findings. Especially, the formulations for a misplaced CVC appeared to be difficult to recognize with simple rules, in contrast to, e.g., pleural effusion and pulmonary congestion. Through further effort by clinicians and technicians, certainly more special cases could be covered by advanced rules, in order to further develop the simple rule-based labeler. However, this would require extensive reading and studying of the reports, during which gold labels could already be generated. With regard to the results of this study, it is therefore questionable whether for information with variable attributes and descriptions, the effort of developing advanced rules for a rule-based labeler is worthwhile compared to generating more gold labels with subsequent training of custom pre-trained transformers.

A limitation of the study is that annotation of the contents of the radiology reports was performed by medical research assistants under the supervision of a radiology resident. Because the annotators did not have to interpret imaging, but were simply required to identify and mark the statements of the attending radiologist within the report, we judged that annotation was not required to be conducted by board-certified radiologists. The high agreement of the different annotators confirmed this judgement, which minimized the cost of annotation and allowed for capturing a larger set of reports.

Conclusion

In conclusion, we find that an on-site custom pre-training of text-based transformers with subsequent adaptation to manually curated gold labels promises to be an efficient strategy to unlock radiological report databases for data-driven medicine.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00330-023-09526-y.

Funding Open Access funding enabled and organized by Projekt DEAL. S.R., B.W., D.B, and H.S. are affiliated with the Competence Center for Machine Learning Rhine-Ruhr, which is funded by the Federal Ministry of Education and Research of Germany (grant no. 01|S18038B). The authors gratefully acknowledge this support. The funders had no influence on the conceptualization and design of the study, data analysis and data collection, preparation of the manuscript, and the decision to publish.

Declarations

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

Conflict of interest The authors of this manuscript declare no relationships with any companies whose products or services may be related to the subject matter of the article.

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

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

Ethical approval Institutional Review Board approval was obtained by the local ethics committees at the Medical Faculty of the Rheinische Friedrich-Wilhelms-Universität Bonn (AZ 411/21).

Methodology
- retrospective
- experimental study
- performed at one institution

References

1. Nobel JM, Kok EM, Robben SG (2020) Redefining the structure of structured reporting in radiology. Insights Imaging 11(1):1–5

2. European Society of Radiology (ESR) (2018) ESR paper on structured reporting in radiology. Insights Imaging 9:1–7
3. Irvin J, Rajpurkar P, Ko M et al (2019) Chexpert: a large chest radiograph dataset with uncertainty labels and expert comparison. In Proceedings of the AAAI conference on artificial intelligence 33(1):590–597
4. Vaswani A, Shazeer N, Parmar N et al (2017) Attention is all you need. In Advances in neural information processing systems 30
5. Devlin J, Chang MW, Lee K, Toutanova K (2018) Bert: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805
6. Wahab A, Sifa R (2021) Dibert: Dependency injected bidirectional encoder representations from transformers. In 2021 IEEE Symposium Series on Computational Intelligence 1–8
7. Smit A, Jain S, Rajpurkar P, Pareek A, Ng AY, Lungren MP (2020) CheXbert: combining automatic labelers and expert annotations for accurate radiology report labeling using BERT. arXiv preprint arXiv:2004.09167
8. McDermott MB, Hsu TMH, Wenig WH, Ghassemi M, Szolovits P (2020) Chexpert++: approximating the chexpert labeler for speed, differentiability, and probabilistic output. In Proceedings of the 5th Machine Learning for Healthcare Conference 913–927
9. Bressem KK, Adams LC, Gaudin RA et al (2020) Highly accurate classification of chest radiographic reports using a deep learning natural language model pre-trained on 3.8 million text reports. Bioinformatics 36(21):5255–5261
10. Richter-Pechanski P, Amr A, Katus HA, Dieterich C (2019) Deep learning approaches outperform conventional strategies in de-identification of German medical reports. In GMDS 101–109
11. Carlini N, Tramer F, Wallace E et al (2021) Extracting training data from large language models. In 30th USENIX Security Symposium 2633–2650
12. Biesner D, Ramamurthy R, Stenzel R et al (2022) Anonymization of German financial documents using neural network-based language models with contextual word representations. Int J Data Sci Anal 13(2):151–161
13. Nakayama H, Kubo T, Kamura J, Taniguchi Y, Liang X (2018) doccano: Text annotation tool for human. Available via https://github.com/doccano/doccano. Accessed 28 Jul 2022
14. Baeza-Yates R, Ribeiro-Neto B (1999) Modern information retrieval, 2nd edn. ACM Press, New York
15. Wolf T, Debut L, Sanh V et al (2019) Huggingface's transformers: state-of-the-art natural language processing. arXiv preprint arXiv:1910.03771
16. Deepset (2021) German BERT. Available via https://huggingface.co/bert-base-german-cased. Accessed 28 Jul 2022
17. Shrestha M (2021) German Medical BERT. Available via https://huggingface.co/smanjil/German-MedBERT. Accessed 28 Jul 2022
18. Sun C, Qiu X, Xu Y, Huang X (2019) How to fine-tune bert for text classification? In 2019 China national conference on Chinese computational linguistics 194–206
19. 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(11):8807–8815
20. 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(1):1–8
21. Gan L, Yuen B, Lu T (2019) Multi-label classification with optimal thresholding for multi-composition spectroscopic analysis. Mach Learn Knowl Extr 1(4):1084–1099
22. Loshchilov I, Hutter F (2017) Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101
23. Pedregosa F, Varoquaux G, Gramfort A et al (2011) Scikit-learn: machine learning in Python. J Mach Learn Res 12:2825–2830

3.3 **Nowak S***, Schneider H*, Layer YC, Theis M, Biesner D, Block W, Wulff B, Attenberger UI, Sifa R*, Sprinkart AM*. Development of image-based decision support systems utilizing information extracted from radiological free-text report databases with text-based transformers. European Radiology. 2024;34(5):2895-2904. DOI: 10.1007/s00330-023-10373-0

Objectives

To investigate the potential and limitations of utilizing transformer-based report annotation for on-site development of image-based diagnostic decision support systems (DDSS).

Methods

The study included 88,353 chest X-rays from 19,581 intensive care unit (ICU) patients. To label the presence of six typical findings in 17,041 images, the corresponding free-text reports of the attending radiologists were assessed by medical research assistants ("gold labels"). Automatically generated "silver" labels were extracted for all reports by transformer models trained on gold labels. To investigate the benefit of such silver labels, the image-based models were trained using three approaches: with gold labels only (M_G), with silver labels first, then with gold labels ($M_{S/G}$), and with silver and gold labels together (M_{S+G}). To investigate the influence of invested annotation effort, the experiments were repeated with different numbers (N) of gold-annotated reports for training the transformer and image-based models and tested on 2099 gold-annotated images. Significant differences in macro-averaged area under the receiver operating characteristic curve (AUC) were assessed by non-overlapping 95% confidence intervals.

Results

Utilizing transformer-based silver labels showed significantly higher macro-averaged AUC than training solely with gold labels (N = 1000: M_G 67.8 [66.0–69.6], $M_{S/G}$ 77.9 [76.2–79.6]; N = 14,580: M_G 74.5 [72.8–76.2], $M_{S/G}$ 80.9 [79.4–82.4]). Training with silver and gold labels together was beneficial using only 500 gold labels (M_{S+G} 76.4 [74.7–78.0], $M_{S/G}$ 75.3 [73.5–77.0]).

Conclusions

Transformer-based annotation has potential for unlocking free-text report databases for the development of image-based DDSS. However, on-site development of image-based DDSS could benefit from more sophisticated annotation pipelines including further information than a single radiological report.

Nowak *et al. European Radiology* (2024) 34:2895–2904
https://doi.org/10.1007/s00330-023-10373-0

EUROPEAN SOCIETY OF RADIOLOGY
European Radiology

IMAGING INFORMATICS AND ARTIFICIAL INTELLIGENCE

Development of image-based decision support systems utilizing information extracted from radiological free-text report databases with text-based transformers

Sebastian Nowak[1*†], Helen Schneider[2†], Yannik C. Layer[1], Maike Theis[1], David Biesner[2], Wolfgang Block[1], Benjamin Wulff[2], Ulrike I. Attenberger[1], Rafet Sifa[2†] and Alois M. Sprinkart[1†]

Abstract

Objectives To investigate the potential and limitations of utilizing transformer-based report annotation for on-site development of image-based diagnostic decision support systems (DDSS).

Methods The study included 88,353 chest X-rays from 19,581 intensive care unit (ICU) patients. To label the presence of six typical findings in 17,041 images, the corresponding free-text reports of the attending radiologists were assessed by medical research assistants ("gold labels"). Automatically generated "silver" labels were extracted for all reports by transformer models trained on gold labels. To investigate the benefit of such silver labels, the image-based models were trained using three approaches: with gold labels only (M_G), with silver labels first, then with gold labels ($M_{S/G}$), and with silver and gold labels together (M_{S+G}). To investigate the influence of invested annotation effort, the experiments were repeated with different numbers (N) of gold-annotated reports for training the transformer and image-based models and tested on 2099 gold-annotated images. Significant differences in macro-averaged area under the receiver operating characteristic curve (AUC) were assessed by non-overlapping 95% confidence intervals.

Results Utilizing transformer-based silver labels showed significantly higher macro-averaged AUC than training solely with gold labels ($N = 1000$: M_G 67.8 [66.0–69.6], $M_{S/G}$ 77.9 [76.2–79.6]; $N = 14{,}580$: M_G 74.5 [72.8–76.2], $M_{S/G}$ 80.9 [79.4–82.4]). Training with silver and gold labels together was beneficial using only 500 gold labels (M_{S+G} 76.4 [74.7–78.0], $M_{S/G}$ 75.3 [73.5–77.0]).

Conclusions Transformer-based annotation has potential for unlocking free-text report databases for the development of image-based DDSS. However, on-site development of image-based DDSS could benefit from more sophisticated annotation pipelines including further information than a single radiological report.

Clinical relevance statement Leveraging clinical databases for on-site development of artificial intelligence (AI)–based diagnostic decision support systems by text-based transformers could promote the application of AI in clinical practice by circumventing highly regulated data exchanges with third parties.

†Sebastian Nowak, Helen Schneider, Rafet Sifa, and Alois M. Sprinkart contributed equally.

*Correspondence:
Sebastian Nowak
sebastian.nowak@ukbonn.de
Full list of author information is available at the end of the article

Key Points • *The amount of data from a database that can be used to develop AI-assisted diagnostic decision systems is often limited by the need for time-consuming identification of pathologies by radiologists.*

• *The transformer-based structuring of free-text radiological reports shows potential to unlock corresponding image databases for on-site development of image-based diagnostic decision support systems.*

• *However, the quality of image annotations generated solely on the content of a single radiology report may be limited by potential inaccuracies and incompleteness of this report.*

Keywords Radiology, Deep learning, Intensive care units, Thorax

Introduction

The application of AI-based DDSS has demonstrated the potential to increase efficiency and reading accuracy, thereby improving patient care [1–3]. The development of image-based DDSS requires a significant amount of training images for which it is known whether the disease of interest is present or not. If these annotations are not available for an image database, the number of images that can be used for DDSS development is limited by the need for time-consuming and costly image evaluation by annotators with considerable domain knowledge [1]. A further challenge is that medical data is subject to strict privacy regulations in most countries, making it difficult to share medical images for creating large international databases [4]. As a result, there is potential for local development of image-based DDSS in radiology clinics, as no data exchange in compliance with privacy regulations is required and diagnoses and findings are already made by radiology experts during clinical routine and documented in radiology reports.

These reports are commonly in free-text format, as many clinics have not integrated structured reporting into their daily routine [5]. To retrospectively identify a cohort of patients with a disease of interest from a report database, and thereby create labels for image-based DDSS development, it is necessary to assess the content of the reports in a fixed set of labels. Although the time-consuming and expert knowledge requiring reporting of images does not have to be repeated, retrospective assessment of the content of thousands of radiological reports to identify patient cohorts continues to involve considerable effort. To overcome this burden, various labeling and model pretraining strategies have been proposed to develop state-of-the-art transformer-based natural language processing (NLP) methods to classify the content of single radiological reports that can be used for retrospective structuring of chest X-ray report databases [6–8]. In a recent study, we investigated the potential of these different approaches for retrospective structuring of chest X-ray reports of ICU patients with respect to initial human annotation time required for subsequent NLP developments [9].

The results of a recent conference paper, in which the authors used X-ray images and English reports from the CheXpert dataset, indicate an advantage of transformers over rule-based systems in creating report content annotations for training image-based DDSS [10]. In another study using in-house chest X-ray examinations from a German university hospital, transformer-based annotations were also successfully used to develop image-based models [2]. Although manual report content was captured in "gold labels" for performance evaluations in these studies, the image-based DDSS were primarily trained with automatically generated "silver labels" from transformers. However, when a clinic develops a transformer to classify report content for on-site database structuring, manual annotations are typically performed. These are then also available as gold labels for subsequent training of the image-based DDSS. Therefore, in a realistic scenario, the development of transformer models has to be considered together with the subsequent development of image models.

The aim of this exploratory study is to gain insight into the potential and limitations of using manually created gold labels and transformer-based silver annotations of the contents of radiological reports for subsequent on-site development of image-based AI models for DDSS, also with respect to manual report annotation effort.

Material and methods

Overview

Radiological report content annotations generated in a previous study on transformer-based structuring of free-text radiology databases were used to label the corresponding ICU chest X-ray images for the development of DDSS systems [9]. Figure 1 illustrates the overall concept of the study and provides an overview of the different data sources and datasets used, as well as an overview of the different experiments conducted.

Dataset

With institutional review board approval (AZ 411/21), written informed consent was waived. Approved data processing took place based on the health data

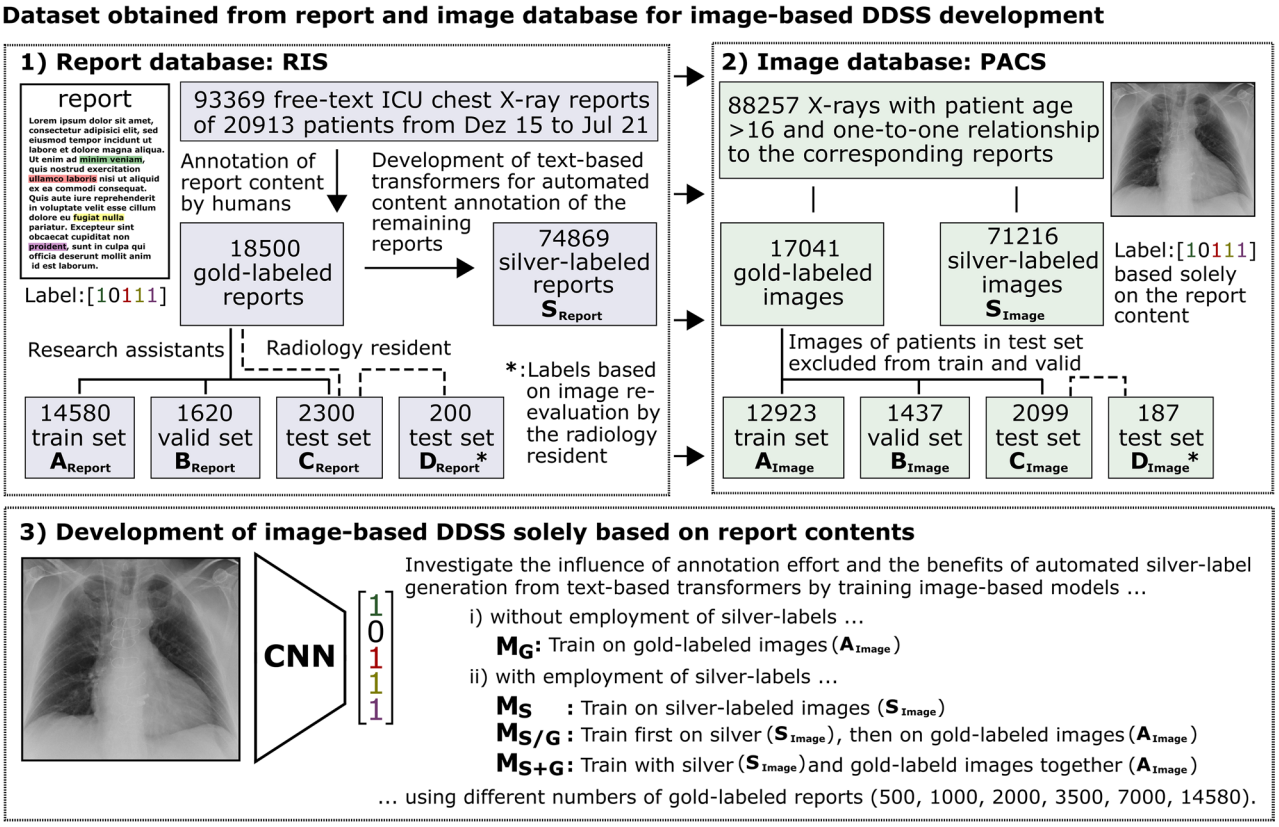


Fig. 1 Overview of the entire study. (1) Report contents of chest X-ray examinations from intensive care unit (ICU) patients were exported from the radiology information system (RIS). For a portion of the exported reports, the text content was manually annotated ("gold labels") and divided into a training (A_{Report}), validation (B_{Report}), and test (C_{Report}) subset. Text-based transformer models that automatically "silver label" the content of the remaining reports were developed using the gold-labeled reports (S_{Report}). The report annotation and development of the transformers shown in (1) was conducted in a previous study. For the current study, the corresponding images of 200 reports of the C_{Report} subset were re-evaluated to create image-based gold labels for testing and to assess the disagreement with the report content (D_{Report}). (2) Images of patients older than 16 years with a clear one-to-one relationship to their associated report were exported from the Picture Archiving and Communication System (PACS). Consequently, the corresponding images to the different report datasets were available that have report content-based gold or silver labels (A_{Image}, B_{Image}, C_{Image}, S_{Image}) or image-based gold labels (D_{Image}). (3) These datasets were used to explore different approaches for leveraging report content for the development of image-based DDSS

protection act North Rhine-Westphalia (GDSG NW) §6 (2) state law NRW. The initial cohort includes 93,368 chest X-ray examinations with reports in German language of 20,913 ICU patients of the University Hospital Bonn from December 2015 to July 2021. The chest X-ray examinations were requested from various ICUs of our clinic (24% from anesthesiological, 24% from cardio-surgical, 20% from surgical, 11% from cardiological, 8% from neurological, 7% from internal medicine, 3% from oncological, and 3% from pediatric ICUs). In a previous study, two trained medical research assistants manually annotated the content of 18,000 chest X-ray reports under the supervision of a radiology resident with a mean annotation time of 39.4 s per report [9]. In these manually assessed reports, common indications were "position of medical devices" (45%) or presence of "pleural infiltrates" (39%), "pneumothorax" (38%), "pleural effusion" (30%), and "congestion" (22%). Additional 500 reports were annotated by the radiology resident and independently by the trained medical research assistants to assess inter-reader variability (mean accuracy of agreement: 97.4% and 97.3%, mean Cohen's kappa: 0.92 and 0.91) [9]. These manually generated annotations are referred to as "gold labels." This gold-labeled data set was randomly split into 14,580 training (A_{Report}), 1620 validation (B_{Report}), and 2300 hold-out test reports (C_{Report}). The test set includes the 500 annotations from the radiology resident. For 200 reports of the test set that were annotated by the medical research assistants, the radiology resident reinterpreted imaging to assess overall label quality and to serve as an additional image labeled test set (D_{Report}). In addition to these gold-labeled reports,

automatically generated "silver labels" (S_{Report}) were created by text-based transformer models (see Fig. 1). Detailed information about the annotation process can be found in supplement S5 and details on the development of the employed NLP algorithms can be found in the previous open-access study [9].

In 91,461 out of 93,368 examinations, a DICOM query of the picture archiving and communication system of the clinic returned only a single image object for the accession number associated with the report. Based on the unique one-to-one relationship between the report and the image, automatic export of the relevant image was performed while the remaining studies were excluded. Subsequently, patients younger than 16 years of age were excluded since the proportion and anatomy of not full-grown patients is different from that of full-grown patients. This resulted in a dataset with 88,257 images, 17,041 with gold labels and 71,216 with silver labels (S_{Image}). No images were excluded due to quality aspects so that the data set reflects a realistic representation of clinical routine images. Furthermore, it was ensured that no images from other examination days of a patient from the test and validation cohort were in the training set. If there were several images of a patient acquired on different examination days within the test or validation cohort, one image was randomly selected. This resulted in a total of 12,923 training (A_{Image}), 1437 validation (B_{Image}), and 2099 test (C_{Image}) images that had corresponding gold-labeled reports and 187 images from the test set with image-based gold labels (D_{Image}). Based on these silver and gold annotated images, DDSS models were developed for the detection of pulmonary infiltrates, pleural effusion, pulmonary congestion, pneumothorax, and misplaced position of the central venous catheter (CVC).

Pre-processing

An algorithm was applied to perform a rectangular crop of image areas outside the radiation field that were caused by acquiring the image with portable X-ray equipment in supine position. Details can be found in supplement S1. The cropped images were resized to 512×512 pixels. Then, a standard U-Net model segmented the lung to allow for computation of mean and variance within the lung mask for *z*-score normalization of the image values [11]. More information on the development of the lung segmentation U-Net used for pre-processing can be found in supplement S2. During training of the DDSS models, image augmentation methods were applied, which are described in detail in supplement S3. During training, all classes were up-sampled to at least 20% to avoid class imbalance in multi-label classification.

Experiments

A DenseNet-121 Convolutional Neural Network with ImageNet pre-trained weights from the PyTorch torchvision library was used as established model for processing lung diseases in chest X-rays [12, 13]. To investigate the benefits of automatically transformer-generated silver labels, the model was trained with four approaches: (i) ith gold labels only (M_G), (ii) with silver labels only (M_S), (iii) first with silver then with gold labels ($M_{S/G}$), and (iv) with silver and gold labels together (M_{S+G}).

To investigate these approaches with respect to different amounts of invested human annotation effort in an end-to-end manner, the development of transformers for silver label generation and the development of image-based DDSS using approaches i, ii, iii, and iv were repeated using different amounts of gold-labeled reports (*N*: 500, 1000, 2000, 3500, 7000, 14,580).

Binary cross entropy loss, AdamW optimizer, a one cycle learning rate schedule with a maximum learning rate of 0.01, a weight decay of 0.01, and a batch size of 128 was used for training [14]. While fine-tuning the $M_{S/G}$ model on gold labels after training with silver labels, the maximum learning rate was reduced by a factor of 10^{-1} per dense block from the last to the first block, as commonly done when applying pre-trained weights [15, 16]. Detailed information on model architecture and training can be found in supplement S4. Model performance was assessed by single and macro-averaged AUC with 95% confidence intervals calculated by bootstrapping with 1000 resamples using torchmetrics v0.10.3. Non-overlapping CIs are interpreted as significant differences [17].

The report content classifying Bidirectional Encoder Representations from Transformers (BERT) models was developed in a previous study by pre-training the transformer with the unsupervised learning technique "masked language modeling" and subsequent fine tuning to gold-labeled reports [9]. Detailed information on the training and hyperparameters used can be found in the previous open-access study on on-site development of transformers in radiological clinics [9].

Results

The main findings of the results are the following:

- The use of transformer-based silver labels is beneficial for the development of image-based DDSS of ICU chest X-ray examinations.
- Separated training with silver and then gold labels is advantageous if more than 2000 gold labels are available.
- There are differences between labels based on report content and labels based on image reinterpretation.

Table 1 shows the number of positive cases for the different pathological findings for all datasets used. The three classes with the lowest number of positive cases in the gold label dataset were pneumothorax (429), misplaced CVC (1071), and infiltrates (2560), and the two classes with the highest number of positive cases were congestion (4423) and effusion (6063).

Table 2 and Fig. 2 show the diagnostic performance of the examined DDSS models evaluated on the test images with report-based labels for all classes and various numbers of gold-labeled reports. For all subsets with 1000 or more of gold-labeled reports employed, significantly higher macro-averaged and misplaced CVC AUC scores were observed for the DDSS models employing transformer generated silver labels (M_S, M_{S+G}, and $M_{S/G}$) compared to the DDSS model trained solely on gold-labeled images (M_G). For pleural effusion, M_S, M_{S+G}, and $M_{S/G}$ performed significantly better than M_G when 3500 or a lower number of gold-labeled reports were available. The same observation was made for pulmonary infiltrates when only 2000 or fewer gold-labeled reports were available. M_{S+G} performed better than M_S and $M_{S/G}$ when using only 500 gold-labeled reports for the three findings pneumothorax, misplaced CVC, and pulmonary infiltrates, which had the lowest number of positive cases. Table 2 additionally lists the diagnostic performance on the test data set with image-based labels (D_{Image}). It was observed that for macro-average, misplaced CVC AUC, M_{S+G} had higher values than M_S and $M_{S/G}$ when 2000 or fewer gold-labeled reports were available and $M_{S/G}$ had higher values than M_S and M_{S+G} when more than 2000 gold-labeled reports were used.

Interestingly, the macro-averaged AUC of the models evaluated on the test set with image-based labels were higher than the macro-averaged AUC of the same models evaluated on the report-based labeled test set. For pulmonary congestion, AUC values of all M_{S+G} and $M_{S/G}$ models evaluated on the dataset with image-based labels were significantly higher than the same models tested on the report-based labels. Detailed metrics for $M_{S/G}$ for which the highest macro-averaged AUC values were observed in both the report- and image-labeled test sets can be found in Table 3.

Table 4 shows the agreement between the labels based on report content and the labels based on image reassessment of the gold-labeled test set (D_{Image}). When comparing report content annotation with image reevaluation, the lowest AUC (93.5%, 95.5%) and accuracy values (93.0%, 93.6%) were observed for pulmonary infiltrates and congestion. For pulmonary infiltrates, sensitivity was 100% and specificity 91.0%, and for pulmonary congestion, sensitivity was 89.3% and specificity 97.6%.

Discussion

In this study, we investigated the potential and limitations of extracting findings from radiology reports, also employing text-based transformers, to annotate the corresponding images for on-site development of image-based DDSS. In many countries, such as Germany, data protection regulations strictly restrict the exchange of radiological reports and images that contain personal data closely linked to sensitive medical information with third parties (e.g., AI companies). The opportunity to develop these systems using unstructured, retrospectively collected data on-site in radiology clinics could drive the development and ultimately the application of specialized AI models in routine clinical practice. These AI applications could, for example, provide an initial assessment immediately after image acquisition by the technical assistants and therefore could contribute to faster detection and treatment of emergencies.

For the following reasons, we considered ICU chest X-ray examinations suitable for investigating this subject. With ICU chest X-ray examinations, there is usually

Table 1 Number of positive cases for all silver- and gold-labeled training images (S_{Image}, A_{Image}) and the gold-labeled validation (B_{Image}) and test subsets (C_{Image}, D_{Image}) used in this study. To investigate the influence of human annotation effort, the experiments were repeated with subsets of the gold-labeled training set A_{Image} with different numbers (*N*) of images

Datasets	S_{Image}	A_{Image}						B_{Image}	C_{Image}	D_{Image}
Label type	Silver	Gold								
Purpose	Training	Training with various *N* of gold-labels						Valid	Test	Test
Number of images	56,797	12,923	6206	3096	1773	877	450	1437	2099	187
Findings	Number of positive cases in dataset splits									
Misplaced CVC	3766	1071	504	253	154	70	37	108	180	44
Effusion	36,922	6063	2868	1428	798	396	200	680	1004	113
Infiltrates	22,291	2560	1226	619	369	192	111	301	729	103
Congestion	17,360	4423	2105	1090	625	292	151	500	424	54
Pneumothorax	2450	429	210	122	71	34	15	51	74	34

Table 2 Area under the receiver operating characteristic curve (AUC) in % observed for the hold-out test set of 2099 images that were labeled by report content and for the hold-out test set of 187 images that were labeled by re-evaluating imaging. The image-based models were trained on report-based labels with four different approaches: solely on gold labels (M_G), solely on silver labels (M_S), first with silver, then with gold labels ($M_{S/G}$) and with silver and gold labels together (M_{S+G}). The transformer and image-based models were trained with various numbers (*N*) of gold-labeled reports and images to investigate the influence of annotation effort on DDSS model performance. For M_S, solely silver-labeled images were used generated by the transformer trained with *N* gold labels. The highest performances of the models trained with the same number of gold labels are indicated by bold font for both test sets. Significant differences between the AUCs of M_G and M_S or M_G and M_{S+G} or M_G and $M_{S/G}$ are indicated by * and between the AUCs of the same model (M_G/M_S/M_{S+G}/$M_{S/G}$) tested on report- or image-based labels with †

Number of gold labels used		Test-set labeled by report content (*N* = 2099)								Test-set labeled by image content (*N* = 187)							
		M_G	M_S	M_{S+G}	$M_{S/G}$	M_G	M_S	M_{S+G}	$M_{S/G}$	M_G	M_S	M_{S+G}	$M_{S/G}$	M_G	M_S	M_{S+G}	$M_{S/G}$
Reports	Images	AUC macro-averaged				Misplaced CVC				AUC macro-averaged				Misplaced CVC			
14,580	12,935	74.5	79.7*	78.8*	**80.9***	63.1	73.5*	77.3*	**77.7***	75.8	84.6*	82.4	**84.8***	61.3	81.8*	79.3*	**83.4***
7000	6206	73.4	78.1*	78.2*	**79.2***	64.3	73.4*	70.5	**74.1***	76.5	82.1*	82.0	**82.8**	68.8	76.4	73.6	**76.7**
3500	3096	71.8	78.3*	**79.2***	78.5*	63.1	71.9*	**74.5***	72.6*	75.7	82.9*	81.8	**83.0***	65.4	77.7	73.1	**77.9**
2000	1773	71.5	77.4*	**78.5***	**78.5***	63.4	71.3*	73.2*	**74.3***	73.5	79.9	**81.5***	81.1*	67.4	71.7	**75.9**	75.6
1000	877	67.8	77.5*	77.3*	**77.9***	59.7	68.6*	69.8*	**69.6***	69.5	80.3*	**82.8***†	80.2*	57.5	69.9	**76.0**	69.5
500	450	68.5	75.1*	**76.4***	75.3*	57.7	65.7	**69.2***	67.4*	68.9	78.9*	**80.1***	76.9*	58.9	72.5	**76.7**	69.7
Reports	Images	Pleural effusion				Pulmonary congestion				Pleural Effusion				Pulmonary congestion			
14,580	12,935	83.8	86.1	85.7	**86.4**	72.5	73.5	**75.2**	74.5	84.5	87.9	**88.6**	87.5	81.1	81.7	**84.8**†	83.9†
7000	6206	83.6	84.5	**85.9**	85.8	72.9	74.2	74.3	**74.4**	84.1	85.5	**87.7**	86.6	81.9†	84.8†	84.3†	**84.8**†
3500	3096	82.2	85.7*****	**86.1***	85.7*	69.3	74.4*	**74.8***	74.4*	82.2	88.2	87.1	**88.5**	81.9†	83.4†	83.0†	**83.9**†
2000	1773	81.1	85.8*****	**86.2***	85.6*	70.7	73.9	73.3	**74.4**	81.3	86.7	**87.8**	87.6	80.6†	82.3†	82.2†	**83.5**†
1000	877	79.8	86.3*****	85.9*	**86.2***	69.2	74.3*	73.5	**74.6***	79.1	**87.2**	86.8	86.8	81.6†	83.8†	**84.3**†	83.9†
500	450	80.4	84.4*****	84.4*	**84.8***	68.1	72.7	71.4	**72.9***	79.4	85.5	82.1	**86.3**	76.5	**85.4**†	81.8†	85.0†
Reports	Images	Pulmonary infiltrates				Pneumothorax				Pulmonary infiltrates				Pneumothorax			
14,580	12,935	80.6	82.3	**82.2**	81.9	72.5	83.4	73.9	**84.0**	73.3	**81.3**	79.1	77.3	79.1	90.3	80.2	**91.9**
7000	6206	78.5	81.4	**82.6**	81.7	67.6	77.2	77.5	**79.8**	76.3	**79.4**	79.2	77.8	71.2	84.7	85.2	**88.0***
3500	3096	78.7	81.2	**82.0**	81.1	65.8	78.6	**78.8***	78.5	78.7	76.1	**77.5**	76.0	70.2	**89.2***	88.0*	88.8*
2000	1773	74.1	81.8*	**82.4***	81.6*	68.1	74.0	**77.4**	76.7	68.2	**79.4**	77.5	77.8	69.9	79.1	**83.8**	81.1
1000	877	70.3	80.9*	**83.1***	81.6*	59.9	77.5*	74.2*	**77.6***	63.6	74.5	**81.3***	75.6	65.9	**86.3***	85.5*	85.3*
500	450	72.4	79.1*	**80.7***	78.6*	63.9	73.6	**76.3***	72.8	69.0	**75.8**	74.8	72.2	60.8	75.3	**84.9***	71.4

a clear one-to-one relationship between the report and the image, without the report describing multiple images of an imaging series. The image data is two-dimensional, which makes the development of DDSS less complex. The images of ICU patients frequently present severe pathologies, which reduces class imbalance for training of DDSS. Lastly, rapid identification of pathologies is essential in these critically ill patients, which makes DDSS of high interest [18]. However, ICU chest X-ray examinations are in principle more demanding to analyze than regular chest X-rays. One reason for this is that ICU patients suffer from a variety of serious conditions and may receive a variety of treatments. ICU patients may be mechanically ventilated; there may be tubes, catheters, and other medical devices that can alter, obscure, or distort the anatomy of the lungs. Another reason is a frequently limited image quality. ICU X-rays of critically ill patients are typically acquired with portable X-ray scanners in lying position, which can induce gravity related alterations in location and appearance of organs and tissues. Also, the condition of the patient and the medical equipment may not allow ICU patients to be positioned accurately perpendicular to the X-ray beam resulting in further image distortion.

Despite these particular challenges, the image-based model utilizing both manual and transformer-based report content labels showed a macro-average AUC of 84.8% on the image-labeled test set. This indicates the potential of transformers for unlocking the content of free-text reports of radiological report databases to ultimately develop image-based DDSS without the need for image re-evaluations. The investigation of the performance of the models developed with different numbers of gold-labeled reports demonstrated that it is beneficial to train with silver and gold labels together when only 2000 or fewer reports have been annotated by humans. If more reports can be annotated, separated training with silver and then gold labels appeared preferable in our study compared to training with a mixture

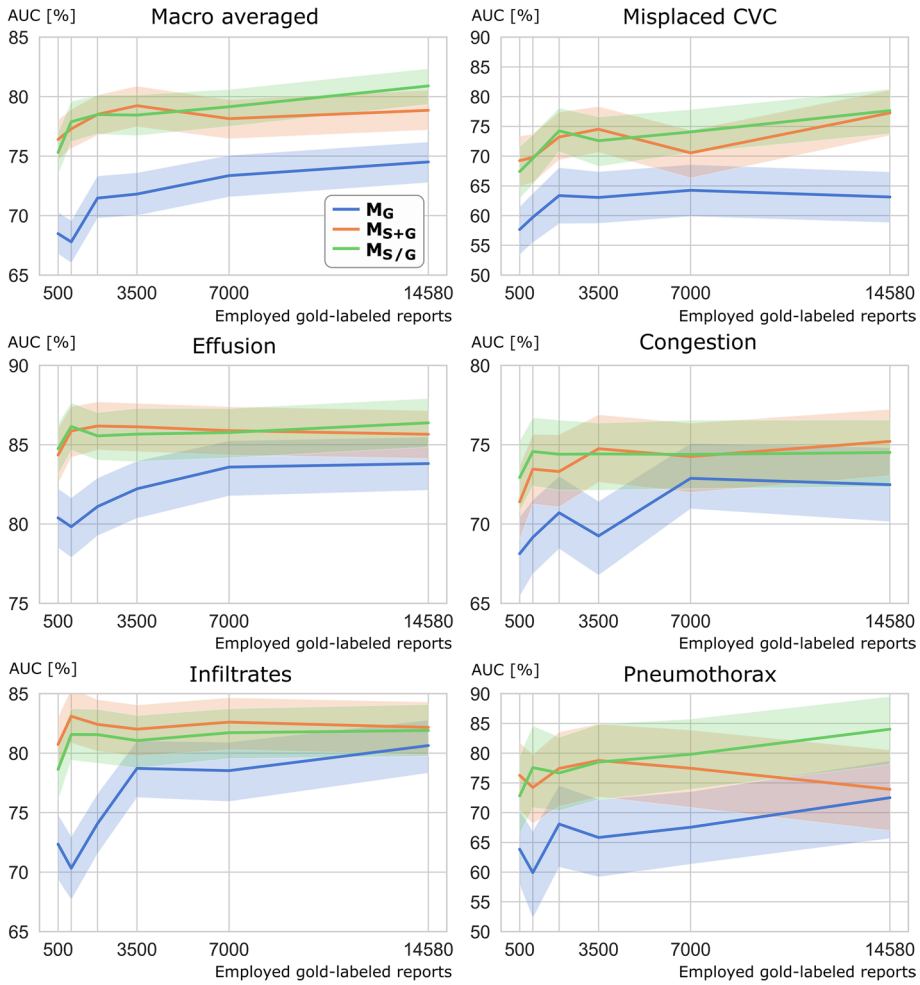


Fig. 2 Area under the receiver operating characteristic curve (AUC) of the image-based DenseNet models for various levels of human annotation effort, represented as different numbers of employed manually labeled reports on the *x*-axis. Note that the transformer models for report content classification (silver labels generation) were also employing the same varying amounts of manually gold-labeled reports so that the end-to-end effect of different amounts of human annotation effort can be assessed. CVC, central venous catheter; M_G: model trained on solely report-based gold labels; M_{S+G}: model trained on report-based silver and gold labels together; $M_{S/G}$: model trained first on report-based silver labels, then on gold labels

Table 3 Detailed metrics for the receiver operating characteristic analysis of the best model $M_{S/G}$ trained with all available data on both test sets with report and image-based labels. The area under the receiver operating characteristic curve (AUC) in % is given per class. Also, sensitivity and specificity in % are given per class for binary classifications. Thresholds were calculated by the Youden-Index on the training set and applied to the test set

	Test-set labeled by report content			Test-set labeled by image content		
Classes	**AUC**	**Sensitivity**	**Specificity**	**AUC**	**Sensitivity**	**Specificity**
Misplaced CVC	77.6	64.4	74.1	83.4	70.5	76.9
Pleural Effusion	86.4	71.9	83.3	87.5	73.5	83.8
Pulmonary Congestion	74.5	57.8	75.9	83.9	60.2	88.1
Pulmonary Infiltrates	81.9	78.3	70.7	77.3	77.8	66.9
Pneumothorax	84.0	87.8	60.6	91.8	97.1	60.1
Overall	80.9	72.0	72.9	84.8	75.8	75.2

Table 4 Accuracy, sensitivity, specificity, area under the receiver operating characteristic curve (AUC), precision, and F1-Score between report-based generated labels from medical research assistants and image-based labels from a radiology resident. A total of 187 images were considered during the evaluation

Class	Accuracy	Sensitivity	Specificity	AUC	Precision	F1-score
Misplaced CVC	97.9	100.0	97.2	98.6	91.7	95.7
Pleural effusion	95.7	95.6	95.9	95.8	97.3	96.4
Pulmonary Congestion	93.0	89.3	97.6	93.5	97.9	93.4
Pulmonary infiltrates	93.6	100.0	91.0	95.5	81.8	90.0
Pneumothorax	98.9	100.0	98.7	99.3	94.4	97.1
Overall	95.8	97.0	96.1	96.5	92.6	97.0

of gold- and silver-labeled images. This is in line with the observation on the two test datasets that the model trained with only silver-labeled images performed better than the model trained with mixed label types when 14,580 gold-labeled reports were available to train the silver label generating transformer.

In addition to the report-based labeled test set, we also generated an image-based labeled test to investigate discrepancies between report content and image findings that potentially pose a limitation to the use of manual and transformer generated report-based labels for on-site DDSS development. Interestingly, it was observed that all models demonstrate higher macro-averaged AUC values when evaluated on the test set with image-based labels compared to evaluation of the same models on the report-based labeled test set. A previous conference paper already discussed potential reasons that can lead to discrepancies between report-content and image findings [19]:

i) Findings that are not of high relevance to the current clinical condition of the patient might not be mentioned in the report, although they may be present within imaging.
ii) Findings within a report may be based on information that is not content of the report, e.g., information from reports from previous examinations or clinical/laboratory parameters.
iii) Borderline image findings could yet be remarked by the attending radiologist for assurance and consequently be considered equally as definite findings for the DDSS training.
iv) And lastly, the radiologist might have made an error during the reporting. Also, further errors may occur during the subsequent annotation of the report content by the human annotators and/or by the transformers.

To assess the overall label discrepancies potentially caused by the above-listed reasons, the results of the image reassessments were compared with the gold labels based on the report content. This revealed high specificity combined with lower sensitivity for pulmonary congestion; i.e., congestions present within imaging were occasionally not mentioned in the report. However, it was rare that the image reader disagreed after re-evaluation when the pathology was mentioned in the report. One could speculate that minor congestions that were not of major importance for the current clinical question were occasionally not reported, as also described in above-described scenario i. terestingly, both models pre-trained with silver labels showed significantly higher AUC values for pulmonary congestions when evaluated on the test subset with image-based labels compared to the test subset with report-based labels. This indicates that despite the observed limited sensitivity of the report content for pulmonary congestion, the DDSS models learned to correctly detect the pathology also in some cases where it was not mentioned in the corresponding reports of the test subjects.

For pulmonary infiltrates, high sensitivity with lower specificity was observed when comparing report content with image re-evaluation. This implies that the reader who re-assessed infiltrates solely on imaging occasionally disagreed with the occurrence of the pathology in the report. However, when the image reader identified infiltrates, this consistently agreed with the report content.

The more frequent recognition of infiltrates in the report texts compared a to re-evaluation of the images may result from additional information available to the attending radiologist at the time of reporting, but which is not content of the report text, as described in scenario ii. or example, recent inflammatory laboratory values and results of previous clinical examinations or previous radiological reports may have encouraged the examiner to describe a lesion as an infiltrate. The more frequent inclusion of infiltrates in the report texts may also be caused by the difficulty identifying a lesion as pulmonary infiltrate on ICU images with patients in lying position. This may increase the number of borderline cases that could still be mentioned in the report by the attending radiologist, as described above in scenario iii.

Other work propose the following approaches to address this challenge of imprecise direct mapping of report and image content. Similar to the current study, one study proposes to first train an image-based deep learning model with labels that are derived from the content of the corresponding reports [19]. The authors claim that the class probabilities provided by this image-based model are more precise labels for the development of text-based transformers in comparison to the initial labels derived from the report content. A follow-up study shows that the labels of this improved transformer also lead to higher performance of the image-based DDSS [10]. Another paper proposes a more sophisticated approach for the annotation of chest X-ray images based on report content by also assessing a second report of a recent CT scan [2]. If the contents of both reports agree, the authors assume that the X-ray report text is accurate. To reduce noise in the dataset caused by imprecise report texts, the authors also propose to first train an image-based model on the noisy data. Then, some image-based labels are manually created by reviewing cases for which the prediction of this model strongly disagrees with the report content label. This more sophisticated approach, involving annotation of two reports and reviewing of imaging, showed promising results in improving the quality of the labels. However, the scope of eligible patients is limited, as imaging and reporting must be available for both modalities and the manual re-evaluation of images requires costly time of radiological experts. Other work presented algorithmic approaches to increase robustness to noisy labels during training of an image-based deep learning model. For example, one paper proposes to extend the loss function to allow the model to ignore cases during training that are strong outliers due to inaccurate labels [20]. This warrants further studies investigating the utility of more time-consuming labeling approaches versus the use of algorithmic approaches to handle the noise of labels extracted directly from report contents for on-site DDSS development in radiology departments.

The use of transformer-based report content annotation for DDSS developments has a further limitation that is not apparent from the study results. Unlike the ICU chest X-ray examination used in this study, the report content of, e.g., MRI examinations are based on multiple imaging sequences. Therefore, further considerations are required when applying the concept to other imaging modalities.

Conclusion

The results show that report content extraction by transformers could aid in unlocking unstructured retrospective routine data in radiological clinics for on-site DDSS development. However, noisy labels caused by imperfect report and image content mapping pose challenges to the presented approach. Therefore, on-site development of image-based DDSS could potentially benefit from more sophisticated annotation pipelines that include information beyond the corresponding radiological report and from algorithmic approaches to handle noisy labels. Moreover, the application of the approach of employing report contents for training of image-based DDSS should be further investigated for imaging examinations where the report is based on multiple images.

Abbreviations

AUC	Area under the receiver operating characteristic curve
AI	Artificial intelligence
CVC	Central venous catheter
DDSS	Diagnostic decision support systems
ICU	Intensive care unit
NLP	Natural language processing

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1007/s00330-023-10373-0.

Below is the link to the electronic supplementary material.Supplementary file1 (PDF 299 KB)

Funding

Open Access funding enabled and organized by Projekt DEAL. R.S., B.W., D.B, and H.S. are affiliated with the Competence Center for Machine Learning Rhine-Ruhr, which is funded by the Federal Ministry of Education and Research of Germany (grant no. 01|S18038B). S.N. was funded by RACOON (NUM), which is supported by the Federal Ministry of Education and Research of Germany (grant no. 01KX2121). The authors gratefully acknowledge this support. The funders had no influence on the conceptualization and design of the study, data analysis and data collection, preparation of the manuscript, and the decision to publish.

Declarations

Guarantor

The scientific guarantor of this publication is PD Dr.-Ing. Alois Martin Sprinkart.

Conflict of interest

The authors of this manuscript declare no relationships with any companies, whose products or services may be related to the subject matter of the article.

Statistics and biometry

No complex statistical methods were necessary for this paper.

Informed consent

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

Ethical approval

Institutional Review Board approval was obtained by the local Ethics Committees at the Medical Faculty of the Rheinische Friedrich-Wilhelms-Universität Bonn (AZ 411/21).

Study subjects or cohorts overlap

The present study is a follow-up to a study published in European Radiology on the transformer-based structuring of free-text radiology databases (Nowak, S., Biesner, D., Layer, Y.C. et al Eur Radiol (2023). https://doi.org/10.1007/s00330-023-09526-y). The manual and transformer-based report annotations

generated in the previous study were used in the current study to annotate the corresponding ICU radiographs to investigate the value of these report-based annotations for the development of image-based DDSS systems.

Methodology

- retrospective
- experimental study
- performed at one institution

Author details
[1]Department of Diagnostic and Interventional Radiology, University Hospital Bonn, Bonn, Germany. [2]Fraunhofer Institute for Intelligent Analysis and Information Systems IAIS, Sankt Augustin, Germany.

Received: 12 May 2023 Revised: 23 August 2023
Accepted: 5 September 2023 Published: 7 November 2023

References

1. Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts HJ (2018) Artificial intelligence in radiology. Nat Rev Cancer 18(8):500–510
2. Niehues SM, Adams LC, Gaudin RA et al (2021) Deep-learning-based diagnosis of bedside chest X-ray in intensive care and emergency medicine. Invest Radiol 56(8):525–534
3. Mango VL, Sun M, Wynn RT, Ha R (2020) Should we ignore, follow, or biopsy? Impact of artificial intelligence decision support on breast ultrasound lesion assessment. AJR Am J Roentgenol 214(6):1445–1452
4. Richter-Pechanski P, Amr A, Katus HA, Dieterich C (2019) Deep learning approaches outperform conventional strategies in de-identification of German medical reports. Stud Health Technol Inform 267:101–109. https://doi.org/10.3233/SHTI190813
5. Nobel JM, Kok EM, Robben SG (2020) Redefining the structure of structured reporting in radiology. Insights Imaging 11:1–5
6. Smit A, Jain S, Rajpurkar P, Pareek A, Ng AY, Lungren MP (2020) CheXbert: combining automatic labelers and expert annotations for accurate radiology report labeling using BERT. arXiv preprint arXiv:2004.09167
7. McDermott MB, Hsu TMH, Wenig WH, Ghassemi M, Szolovits P (2020) Chexpert++: Approximating the chexpert labeler for speed, differentiability, and probabilistic output. Proceedings of PMLR 126:913–927
8. Bressem KK, Adams LC, Gaudin RA et al (2020) Highly accurate classification of chest radiographic reports using a deep learning natural language model pre-trained on 3.8 million text reports. Bioinformatics 36(21):5255–5261
9. Nowak S, Biesner D, Layer YC et al (2023) Transformer-based structuring of free-text radiology report databases. Eur Radiol. https://doi.org/10.1007/s00330-023-09526-y
10. Jain S, Smit A, Ng AY, Rajpurkar P (2021) Effect of radiology report labeler quality on deep learning models for chest X-ray interpretation. arXiv preprint arXiv:2104.00793
11. Ronneberger O, Fischer P, Brox T (2015) U-net: convolutional networks for biomedical image segmentation. In proceedings of MICCAI 2015 18:234-241
12. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ (2017) Densely connected convolutional networks. In proceedings of CVPR 2017 4700-4708
13. Irvin J, Rajpurkar P, Ko M et al (2019) Chexpert: a large chest radiograph dataset with uncertainty labels and expert comparison. Proc AAAI Conf Artif Intell 33(1):590–597
14. Smith LN (2018) A disciplined approach to neural network hyper-parameters: Part 1 -- learning rate, batch size, momentum, and weight decay. arXiv preprint arXiv:1803.09820
15. 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(11):8807–8815
16. 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(1):1–8
17. Cumming G (2009) Inference by eye: Reading the overlap of independent confidence intervals. Stat Med 28(2):205–220
18. Spiritoso R, Padley S, Singh S (2015) Chest X-ray interpretation in UK intensive care units: A survey 2014. J Intensive Care Soc 16(4):339–344
19. Jain S, Smit A, Truong SQ et al (2021) VisualCheXbert: addressing the discrepancy between radiology report labels and image labels. In Proceedings of CHIL 2021 105-115
20. Thulasidasan S, Bhattacharya T, Bilmes J, Chennupati G, Mohd-Yusof J (2019) Combating label noise in deep learning using abstention. arXiv preprint arXiv:1905.10964

Publisher's Note

3.4 **Nowak S***, Bischoff LM*, Pennig L, Kaya K, Isaak A, Theis M, Block W, Pieper CC, Kuetting D, Zimmer S, Nickenig G, Attenberger UI, Sprinkart AM*, Luetkens JA*. Deep learning virtual contrast-enhanced T1 mapping for contrast-free myocardial extracellular volume assessment. Journal of the American Heart Association. 2024;13(19):e035599. DOI: 10.1161/JAHA.124.035599

Background

The acquisition of contrast-enhanced (CE) T1 maps to calculate extracellular volume (ECV) requires contrast agent administration and is time-consuming. This study investigates generative adversarial networks (GAN) for contrast-free, virtual extracellular volume (vECV) by generating virtual contrast-enhanced (vCE) T1 maps.

Material and Methods

This retrospective study includes 2518 registered native and CE T1 maps from 1000 patients who underwent cardiovascular magnetic resonance (CMR) at 1.5 Tesla. Recent hematocrit values of 123 patients (hold-out test) and 96 patients from a different institution (external evaluation) allowed for calculation of conventional ECV. A GAN was trained to generate vCE T1 maps from native T1 maps for vECV creation. Mean and standard deviation of the difference per patient (ΔECV) were calculated and compared by permutation of the two-sided t-test with 10000 resamples. For ECV and vECV, differences in area under the receiver operating characteristic curve (AUC) for discriminating hold-out test patients with normal CMR versus myocarditis or amyloidosis were tested with tested with Delong's test.

Results

ECV and vECV showed a high agreement in patients with myocarditis (ΔECV: hold-out test=2.0±1.5%, external evaluation=1.9±1.7%) and normal CMR (ΔECV: hold-out test=1.9±1.4%, external evaluation=1.5±1.2%), but variations in amyloidosis were higher (ΔECV: hold-out test=6.2±6.0%, external evaluation=15.5±6.4%). In the hold-out test, ECV and vECV had a comparable AUC for the diagnosis of myocarditis (ECV AUC=0.77 vs. vECV AUC=0.76, P=0.76) and amyloidosis (ECV AUC=0.99 vs. vECV AUC=0.96, P=0.52).

Conclusions

Generation of vECV based on native T1 maps is feasible. Multi-center training data are required to further enhance generalizability of vECV in amyloidosis.

Journal of the American Heart Association

ORIGINAL RESEARCH

Deep Learning Virtual Contrast-Enhanced T1 Mapping for Contrast-Free Myocardial Extracellular Volume Assessment

Sebastian Nowak, PhD*; Leon M. Bischoff, MD*; Lenhard Pennig, MD*; Kenan Kaya, MD; Alexander Isaak, MD; Maike Theis, MSc; Wolfgang Block, PhD; Claus C. Pieper, MD; Daniel Kuetting, MD; Sebastian Zimmer, MD; Georg Nickenig, MD; Ulrike I. Attenberger, MD; Alois M. Sprinkart, PhD*; Julian A. Luetkens, MD*

BACKGROUND: The acquisition of contrast-enhanced T1 maps to calculate extracellular volume (ECV) requires contrast agent administration and is time consuming. This study investigates generative adversarial networks for contrast-free, virtual extracellular volume (vECV) by generating virtual contrast-enhanced T1 maps.

METHODS AND RESULTS: This retrospective study includes 2518 registered native and contrast-enhanced T1 maps from 1000 patients who underwent cardiovascular magnetic resonance at 1.5 Tesla. Recent hematocrit values of 123 patients (hold-out test) and 96 patients from a different institution (external evaluation) allowed for calculation of conventional ECV. A generative adversarial network was trained to generate virtual contrast-enhanced T1 maps from native T1 maps for vECV creation. Mean and SD of the difference per patient (ΔECV) were calculated and compared by permutation of the 2-sided *t* test with 10 000 resamples. For ECV and vECV, differences in area under the receiver operating characteristic curve (AUC) for discriminating hold-out test patients with normal cardiovascular magnetic resonance versus myocarditis or amyloidosis were tested with Delong's test. ECV and vECV showed a high agreement in patients with myocarditis (ΔECV: hold-out test, 2.0%±1.5%; external evaluation, 1.9%±1.7%) and normal cardiovascular magnetic resonance (ΔECV: hold-out test, 1.9%±1.4%; external evaluation, 1.5%±1.2%), but variations in amyloidosis were higher (ΔECV: hold-out test, 6.2%±6.0%; external evaluation, 15.5%±6.4%). In the hold-out test, ECV and vECV had a comparable AUC for the diagnosis of myocarditis (ECV AUC, 0.77 versus vECV AUC, 0.76; *P*=0.76) and amyloidosis (ECV AUC, 0.99 versus vECV AUC, 0.96; *P*=0.52).

CONCLUSIONS: Generation of vECV on the basis of native T1 maps is feasible. Multicenter training data are required to further enhance generalizability of vECV in amyloidosis.

Key Words: cardiovascular magnetic resonance ■ deep learning ■ extracellular volume ■ generative adversarial networks ■ T1 mapping

Cardiovascular magnetic resonance (CMR) is an important imaging modality for diagnosis and follow-up of patients with various cardiomyopathies.[1,2] The clinical success of CMR is owed to its unique capabilities to apply different sequences for a detailed assessment of myocardial function, myocardial edema and inflammation, and myocardial fibrosis.[3–5] Furthermore, techniques like extracellular volume (ECV) fraction mapping provide a quantitative evaluation of underlying myocardial disease including the quantification of myocardial fibrosis and inflammation.[6]

Correspondence to: Sebastian Nowak, PhD and Julian A. Luetkens, MD, Department of Diagnostic and Interventional Radiology, University Hospital Bonn, Venusberg-Campus 1, Bonn 53127, Germany. Email: sebastian.nowak@ukbonn.de and julian.luetkens@ukbonn.de

*S. Nowak, L. M. Bischoff, A. M. Sprinkart, and J. A. Luetkens contributed equally.

This manuscript was sent to Sakima A. Smith, MD, MPH, Associate Editor, for review by expert referees, editorial decision, and final disposition.

Supplemental Material is available at https://www.ahajournals.org/doi/suppl/10.1161/JAHA.124.035599

For Sources of Funding and Disclosures, see page 13.

JAHA is available at: www.ahajournals.org/journal/jaha

CLINICAL PERSPECTIVE

What Is New?

- A generative adversarial network created virtual contrast-enhanced T1 maps for contrast-free, virtual extracellular volume (ECV) fraction assessment. The technique of ECV estimation is used in cardiovascular magnetic resonance, especially for better characterization of inflammatory, infiltrative, and fibrotic myocardial disease.
- Virtual ECV maps had a substantial agreement with conventional ECV maps. Virtual ECV allowed the disease-specific diagnosis of myocarditis and amyloidosis with a comparable diagnostic performance as conventional ECV.

What Are the Clinical Implications?

- Deep learning estimation of ECV by generating virtual contrast-enhanced T1 maps from native T1 maps might facilitate faster and more focused cardiovascular magnetic resonance examinations without the need for gadolinium-based contrast agents.

Nonstandard Abbreviations and Acronyms

CE	contrast-enhanced
ECV	extracellular volume
GAN	generative adversarial network
GBCA	gadolinium-based contrast agent
RFR	random forest regression
vCE	virtual contrast-enhanced
vECV	virtual extracellular volume
ΔECV	extracellular volume difference per patient

Myocardial ECV estimation exploits the nature of gadolinium-based contrast agents (GBCAs) to accumulate in the extracellular space. In a state of dynamic equilibrium with balanced GBCA concentrations between the blood and the extracellular space, ECV can dichotomize the myocardium into an interstitial and cellular compartment. ECV can be calculated using native and contrast-enhanced (CE) T1 mapping normalized for hematocrit.[7] In addition to myocardial fibrosis, ECV values are increased in myocardial edema like in acute myocarditis and are considered a quantitative biomarker for the detection of diffuse disease.[4,8] ECV also enables noninvasive quantification of myocardial amyloid deposition and can influence treatment decisions in cardiac amyloidosis.[9]

In clinical routine, however, the acquisition of ECV is time consuming, and ECV map generation complicates clinical workflow due to the need of acquiring a CE T1 map in the same slice position as the native scan, the need for a peripheral venous catheter, and the need for image registration. Furthermore, the use of GBCAs might not be completely risk free, especially for patients with impaired renal function.[10] Although the use of GBCAs is considered safe in people with preserved renal function, reports of gadolinium deposition in the brain and other organs with still not fully investigated long-term biological effects prompts reasonable usage of GBCAs in clinical routine.[11,12]

Recent studies investigated the utility of deep learning for the generation of virtual CE magnetic resonance images to increase patient safety and to decrease the economic and environmental effects of GBCA usage.[13–15] Research on deep learning–generated CE magnetic resonance images has been mostly applied to brain imaging, and only few studies reported their potential on CMR.[13] These studies employed a generative adversarial network (GAN) to estimate the delayed distribution of GBCAs on late gadolinium–enhanced images from native cine and T1 mapping.[16,17] In our study, we hypothesized that GANs could enable the contrast-free generation of virtual extracellular volume (vECV) maps. Therefore, as a proof of principle, we aimed to investigate the application of GANs to generate virtual contrast-enhanced (vCE) T1 maps from native T1 maps to assess ECV without the use of GBCA.

METHODS

Data Set

The study complies with the principles of the Helsinki Declaration and was approved by the local institutional review board that waived informed consent due to retrospective study design (reference number: 271/23). The data that support the findings of this study are available from the corresponding author upon reasonable request. This study included pairs of native and CE myocardial T1 maps from similar clinically 1.5 Tesla CMR scanners (1.5 T Ingenia, Philips Healthcare), which were acquired in the same short-axis orientation (basal, midventricular, and apical sections). All maps were acquired with a standard 3(3)3(3)5 modified Look-Locker inversion-recovery acquisition scheme before and 10 minutes after administration of a single bolus of 0.2 mmol/kg body weight of gadobutrol (Gadovist, Bayer Healthcare).[18]

Patients who underwent CMR with T1 mapping for various clinical indications between January 2019 and August 2021 at the University Hospital Bonn were consecutively identified by a radiology resident (L.M.B., 3 years of experience in CMR), who was instructed by a board-certified cardiovascular radiologist (J.A.L,

11 years of experience in CMR). To enable a thorough evaluation of vECV estimation in relation to myocarditis and amyloidosis, for which ECV estimation is of particular interest in clinical routine, additional patients were specifically identified applying a broader time range (July 2017 to July 2023). Acute myocarditis was diagnosed on the basis of diagnostic criteria for clinically suspected myocarditis as recommended by the European Society of Cardiology Working Group on myocardial and pericardial diseases.[4,19] Cardiac amyloidosis was diagnosed on the basis of myocardial biopsy, elevated light-chain immunoglobins, technetium-99m or 3,3-di phosphono-1,2-propanodicarboxylic acid scintigraphy or biopsy of another involved organ.[20]

CMR diagnosis and, if available, a hematocrit value within 48 hours before CMR were extracted from the clinical information system. Patients without available hematocrit were used to train and validate the GAN to generate vCE T1 maps using 5-fold cross-validation. Patients with available hematocrit enabling calculation of reference conventional ECV were included in a hold-out test set to investigate the utility of vCE for contrast-free vECV estimation.

Additionally, to create an external evaluation cohort, a radiology resident (K.K., 5 years of experience in CMR, supervised by L.P., a board-certified cardiovascular radiologist with 8 years of experience in CMR) specifically identified patients with CMR and recent hematocrit value of another institution (University Hospital of Cologne) between January 2021 and May 2024. External validation CMR was conducted at 1.5 T (Ingenia, Philips Healthcare) with administration of 0.2 mmol/kg body weight of gadobutrol but with a different 5(3)3 modified Look-Locker inversion-recovery acquisition scheme. Data S1 describes in detail the data preprocessed before training the deep learning method consisting of myocardial segmentation, rigid registration, cropping, and T1 value normalization. T1 map pairs with poor image quality (ie, due to severe motion or banding artifacts) or with failed rigid registration (due to obvious dissimilar orientation or cardiac phase) were excluded from analysis.

Generative Adversarial Network Training for vCE T1 Maps

As generator for vCE T1 maps, we employed a U-Net–like convolutional neural network, which is widely applied for medical image segmentation, but also for deep learning–based generation of CE magnetic resonance images.[14,21–25] We applied 3 common loss functions also used in previous studies to optimize the U-Net generator:

- Pixel-wise mean absolute distance loss[26]
- Structural similarity index loss[27]
- Adversarial loss via “PatchGAN” discriminator[26]

Figure 1 illustrates the investigated deep learning architecture and losses in detail. Detailed information on hyperparameters used during training can be found in Data S2. The methods for vCE T1 map generation were trained with 5-fold cross-validation on T1 maps from the University Hospital Bonn. We compared the GAN-based approach with a random forest regression (RFR). As with the GAN, we developed the RFR on the training set (without hematocrit) to predict mean vCE T1 values of myocardium and of the blood pool on the basis of the mean native T1 values of myocardium and the blood pool. In contrast to the GAN, we additionally included the age and sex of the patient as input variables for the RFR. RFR was implemented using Scikit-learn version 1.1.2 using default parameters (Data S3).

Evaluation of Contrast-Free, vECV Estimation

For the hold-out test set with recent hematocrit allowing for calculation of ECV, all 5 cross-validated GAN models were applied in an ensemble for generation of vCE T1 maps by pixel-wise aggregation of the maps from each model. The generated vCE T1 maps were evaluated for calculating contrast-free vECV. Myocardial ECV/vECV fraction was calculated using mean native and CE/vCE T1 map values within a region of interest in the myocardium and a region of interest within blood. Detailed information on region of interest creation can be found in the Data S4. Figure 1 illustrates the ECV calculation and the applied formula.[28] For visual assessment of vECV, patients with visible focal or diffuse lesions were identified on the conventional ECV and vECV map by a board-certified cardiovascular radiologist (J.A.L.). Lesions were further classified as follows: (A) lesions present in the same location in the conventional ECV and vECV map, (B) lesions present in different locations in vECV compared with conventional ECV, (C) no lesions present in vECV but in conventional ECV, and (D) lesion in the vECV map but no lesion in the conventional ECV map.

Statistical Analysis

Statistical analysis was conducted in Python version 3.9.12 (Python Software Foundation, Beaverton, OR) using SciPy version 1.9.0, Scikit-learn version 1.1.2, statsmodels version 0.13.2 and pROC version 1.18.5 for R, executed in Python by r2py version 3.5.14.[29–32] Separated into different subgroups on the basis of primary CMR diagnosis, differences between CE and vCE T1 on the validation data were tested by permutation of *t* test comparing mean of the difference per patient (ΔCE T1). Also, separated into different subgroups on the basis of CMR diagnosis, differences on hold-out test set and on the external validation cohort with recent hematocrit between conventional ECV

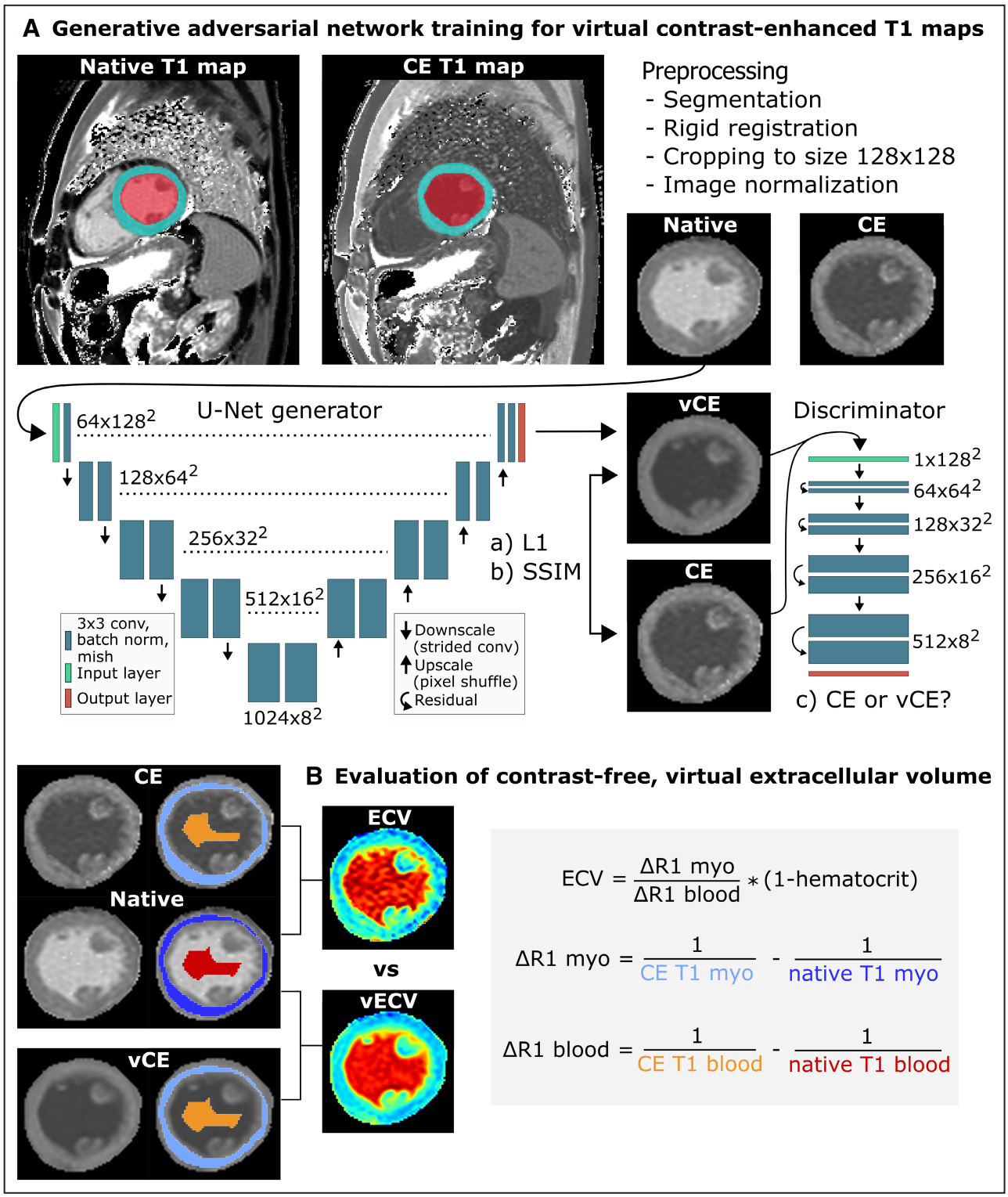


Figure 1. Overview of the presented approach for generating contrast-free, virtual extracellular volume (vECV).
A, Illustration of the generative adversarial network (GAN) used to generate virtual contrast-enhanced (vCE) T1 maps that were used to calculate vECV fraction. The GAN consists of a generator that learns to generate the vCE maps by comparing its predictions with the conventional contrast-enhanced (CE) T1 map and by deceiving a discriminator into recognizing the vCE maps as conventional CE. The GAN model was trained with 5-fold cross-validation on the developmental data set without current hematocrit. **B**, The trained model was applied to the hold-out test data set to generate vCE T1 maps. Myocardial (myo) and blood ΔR1 values were derived from native T1 maps in combination with conventional CE T1 and with vCE T1 maps to calculate conventional extracellular volume (ECV) and virtual extracellular volume fraction (vECV).

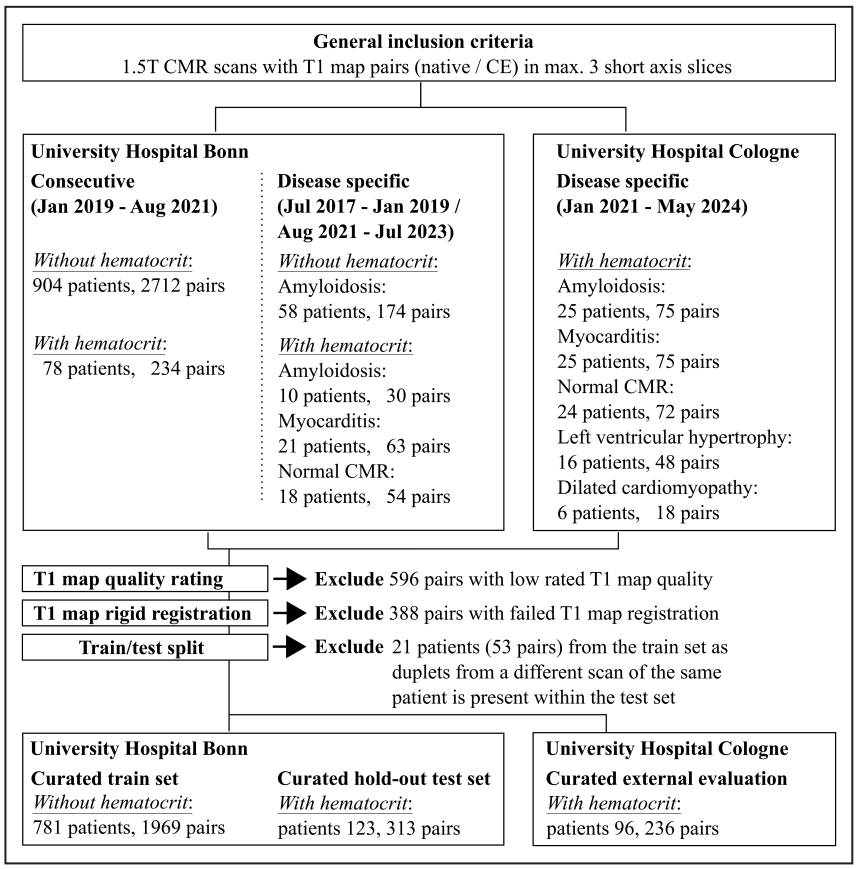


Figure 2. Flowchart illustrating the inclusion and exclusion of native and contrast-enhanced (CE) T1 map pairs for the training of the generative adversarial network to generate virtual CE T1 maps and for the testing of virtual extracellular volume (ECV) estimations.
Only hematocrit values were considered that were derived in a recent blood test within 48 hours before cardiovascular magnetic resonance (CMR).

and vECV was also tested using a permutation of the *t* test and by comparing mean of the difference per patient (ΔECV) for the GAN and RFR models. For ΔCE T1 and ΔECV, SDs are also reported. The *t* test was applied only if the group size was ≥5. For the hold-out test set, correlation between conventional ECV and GAN-based vECV was assessed by linear regression analysis with the Pearson correlation coefficient (*r*) and systematic differences were investigated using Bland–Altman analysis.

A previously published ECV cutoff of our group derived from ECV maps of myocarditis cases of the University Hospital Bonn of 28.8% was applied for both conventional ECV and vECV of the hold-out test set to differentiate between patients with disease and patients with normal CMR.[4,5] Diagnostic performance to differentiate between patients with normal CMR and patients with acute myocarditis or amyloidosis was assessed by area under the receiver operating characteristic curve (AUC). AUC differences between conventional ECV and vECV were tested with Delong's test. Sensitivity and specificity were determined. Differences between ECV and vECV were also tested by McNemar's test. Two-sided permutation tests with 10 000 resamples and the permutation type "samples" were applied for McNemar's and *t* tests using SciPy. For Delong's test, a 2-sided permutation version with 10 000 resamples using the method described by Venkatraman and Begg were conducted using pROC in R (R Foundation for Statistical Computing, Vienna, Austria).[33] Ninety-five percent CIs were determined by bootstrapping with 10 000 resamples for AUC, sensitivity and specificity, and for the difference between conventional ECV and vECV in these metrics. $P<0.05$ was considered indicative of a statistically significant difference.

RESULTS

Data Set

As the internal data set, 904 patients from the University Hospital Bonn were included (mean age, 49.8±19.6 years; 546 men [60%]; 2282 T1 map

pairs). Of those, 123 patients were used as the internal hold-out test set with recent hematocrit (mean age, 51.9±21.2 years; 81 men; 313 T1 map pairs). Additionally, 96 patients with recent hematocrit (mean age, 51.9±23.1; 65 men; 236 T1 map pairs) from the University Hospital Cologne served as external evaluation. A flowchart for patient selection is given in Figure 2. Patient characteristics for the training and test set are given in Table 1. Examples for T1 maps excluded due to bad T1 map quality or due to infeasible rigid registration due to, for example, mismatch in myocardial contraction can be found in Figures S1 to S4.

Table 1. Characteristics of Patients Included in the Hold-Out Test Set

Basic characteristics	Hold-out test set	Training set
No. of patients	123	781
Age, y	51.9±21.2	49.5±19.3
Sex, male/female	81/42	465/316
Body mass index, kg/m^2	25.4±5.0	26.2±6.1
Body surface area, m^2	1.95±0.25	1.96±0.30
Risk factors, n (%)		
Hypertension	42 (33)	306 (39)
Hypercholesterinemia	40 (31)	229 (29)
Smoking history	33 (26)	200 (25)
Family history of CVD	19 (15)	84 (11)
Diabetes	15 (12)	73 (9)
Obesity	21 (16)	168 (22)
Primary CMR diagnosis, n (%)		
Myocarditis	46 (37)	135 (17)
Normal CMR	25 (20)	328 (42)
Dilated cardiomyopathy	12 (10)	46 (6)
Amyloidosis	11 (9)	68 (9)
Ischemic cardiomyopathy	7 (6)	59 (8)
Pericarditis	6 (5)	40 (5)
Takotsubo/stress edema	5 (4)	2 (<1)
Myocardial fibrosis	4 (3)	33 (4)
Left ventricular hypertrophy	3 (2)	15 (2)
Other	4 (3)	55 (7)
MRI parameters		
LVEDV, mL	177±59	167±64
LVEDV index, mL/m^2	90±24	85±30
LVESV, mL	89±58	77±55
LVESV index, mL/m^2	45±25	40±27
LVEF, %	53±14	57±12
Late gadolinium enhancement, n (%)	85 (66)	360 (46)

Diagnoses that occurred in <3 patients of the hold-out test set were combined under "other." CMR indicates cardiovascular magnetic resonance; CVD, cardiovascular disease; LVEDV, left ventricular end-diastolic volume; LVEDV index, left ventricular end-diastolic volume/body surface area; LVEF, left ventricular ejection fraction; LVESV, left ventricular end-systolic volume; LVESV index, left ventricular end-systolic volume/body surface area; and MRI, magnetic resonance imaging.

GAN for vCE T1 Maps

Mean myocardial CE T1 and vCE T1 relaxation times showed similar mean T1 values (mean CE T1, 372±56 ms versus mean vCE T1, 373±29 ms; P=0.37), considering all patients (n=781) from the 5-fold cross-validated training set. For patients with ischemic cardiomyopathy (n=59), significant differences were noted (mean CE T1=350±45 ms vs. mean vCE T1=364±29 ms, P=0.03). Differences in mean myocardial CE T1 and vCE T1 relaxation times for various CMR findings are shown in Table 2.

Evaluation of Contrast-Free vECV

Considering all patients of the hold-out test set (n=123), mean myocardial ECV and GAN-based vECV showed no significant differences (mean ECV: 30.1%±8.0% versus mean vECV, 29.9%±6.4%; P=0.49). In patients with amyloidosis (n=11), significant differences were noted (mean ECV, 48.2%±11.6% versus mean vECV, 42.2±9.5%; P<0.01). Compared with GAN-based vECV, RFR-based vECV showed higher differences in amyloidosis (mean ECV, 48.2%±11.6% versus mean vECV, 34.9%±6.3%; P<0.01) and also in normal CMR (mean ECV, 25.5%±3.2% versus mean vECV, 26.9%±3.1%; P=0.01). Differences in mean myocardial ECV and vECV for various CMR findings and both machine-learning methods are given in Table 3.

Table 2. Five-Fold Cross-Validation Results for Predicting vCE T1 Maps

CMR diagnosis	CE T1 (ms)	vCE T1 (ms)	P value	ΔCE T1 (ms)
All (n=781)	372±56	373±29	0.37	39±33
Normal CMR (n=328)	381±50	382±23	0.57	36±32
Myocarditis (n=135)	371±52	372±22	0.77	40±33
Amyloidosis (n=68)	327±81	330±40	0.69	49±42
Ischemic cardiomyopathy (n=59)	350±45	364±29	0.03*	42±29
Dilated cardiomyopathy (n=46)	391±42	389±20	0.71	36±22
Pericarditis (n=40)	378±59	375±26	0.75	40±42
Myocardial fibrosis (n=33)	377±49	377±20	0.96	40±30
Left ventricular hypertrophy (n=15)	371±52	351±15	0.08	43±40
Takotsubo/stress edema (n=2)	352±20	367±6	N/A	16±15
Other (n=55)	373±46	376±20	0.58	32±27

Mean±SD, as well as ΔCE T1 are given for conventional myocardial contrast-enhanced T1 relaxation times and vCE T1 relaxation times separated for different subgroups on the basis of cardiovascular magnetic resonance diagnosis. P values were obtained using a permutation version of the 2-sided t test with 10000 resamples. Bold numbers indicate statistical significance. vCE indicates virtual contrast-enhanced.

*Statistical significance.

Table 3. Hold-Out Test Set Results for ECV and vECV Fraction

CMR diagnosis	ECV (%)	vECV (%)	*P* value	ΔECV (%)
Generative adversarial network				
All patients (n=123)	30.1±8.0	29.9±6.4	0.49	2.4±2.8
Myocarditis (n=46)	29.5±4.8	29.4±4.1	0.83	2.0±1.5
Normal CMR (n=25)	25.4±3.1	26.1±2.7	0.16	1.9±1.4
Dilated cardiomyopathy (n=12)	30.8±4.8	31.8±3.7	0.26	2.4±1.8
Amyloidosis (n=11)	48.2±11.6	42.2±9.5	<0.01*	6.2±6.0
Ischemic cardiomyopathy (n=7)	25.6±3.2	25.3±4.9	0.81	1.9±0.7
Pericarditis (n=6)	28.6±2.0	28.0±2.4	0.62	2.1±1.6
Takotsubo/stress edema (n=5)	31.7±5.1	32.0±4.7	0.75	1.1±0.8
Myocardial fibrosis (n=4)	27.5±2.3	25.8±1.4	N/A	2.6±0.7
Left ventricular hypertrophy (n=3)	28.9±1.2	36.3±3.6	N/A	7.4±4.5
Other (n=4)	25.6±2.3	25.6±2.4	N/A	0.1±0.1
Random forest regressor				
All patients (n=123)	30.1±8.0	29.7±5.4	0.50	3.7±5.2
Myocarditis (n=46)	29.5±4.8	30.3±5.6	0.13	2.5±2.2
Normal CMR (n=25)	25.5±3.2	26.9±3.1	0.01*	2.4±1.9
Dilated cardiomyopathy (n=12)	30.8±4.8	31.0±2.4	0.93	3.2±3.2
Amyloidosis (n=11)	48.1±11.8	34.9±6.3	<0.01*	14.7±10.8
Ischemic cardiomyopathy (n=7)	25.6±3.2	25.5±4.0	0.94	1.7±0.9
Pericarditis (n=6)	28.6±2.0	29.5±3.4	0.62	2.9±1.9
Takotsubo/stress edema (n=5)	31.7±5.1	34.6±6.2	0.12	3.3±2.0
Myocardial fibrosis (n=4)	27.5±2.3	26.8±3.4	N/A	1.7±1.0
Left ventricular hypertrophy (n=3)	28.9±1.2	31.2±5.6	N/A	5.5±3.7
Other (n=4)	25.6±2.3	25.6±3.4	N/A	1.2±0.9

ECV and vECV were calculated on the basis of contrast-enhanced T1 map and virtual contrast-enhanced T1 values generated by a generative adversarial network and by a random forest regressor. Mean±SD, as well as mean of the difference per patient are given for ECV and vECV separated into subgroups on the basis of CMR diagnosis. *P* values were obtained using a permutation version of the 2-sided *t* test with 10000 resamples. CMR indicates cardiovascular magnetic resonance; ECV, extracellular volume; ΔECV, mean of the difference in extracellular volume per patient; and vECV, virtual extracellular volume.

*Statistical significance.

We calculated GAN-based vECV values for 81 native T1 map pairs, which were excluded from comparison with conventional ECV because registration with the conventional CE T1 map failed due to dissimilar slice or myocardial contraction (normal CMR n=42, mean vECV, 27.7%±4.5%; myocarditis n=81, mean vECV, 29.9%±4.9%; amyloidosis n=11, mean vECV, 42.2%±8.4%; dilated cardiomyopathy n=1, vECV=39.3%).

Similar diagnostic performance of conventional ECV and vECV for the detection of patients with myocarditis

Table 4. Diagnostic Performance of Myocardial ECV and vECV Fraction

Variable	Sensitivity (%)	Specificity (%)	Accuracy (%)	McNemar *P* value	AUC	Delong *P* value
Myocarditis						
ECV	54 (42–69)	92 (80–100)	68 (56–78)		0.77 (0.65–0.87)	
vECV	52 (37–67)	80 (64–96)	62 (50–73)		0.76 (0.63–0.86)	
Difference	2 (0–16)	12 (0–31)	6 (0–15)	1.00	0.01 (0–0.10)	0.76
Amyloidosis						
ECV	100 (100–100)	92 (80–100)	94 (86–100)		0.99 (0.97–1)	
vECV	91 (70–100)	80 (62–95)	83 (69–94)		0.96 (0.86–1)	
Difference	9 (0–30)	12 (0–28)	11 (0–25)	1.00	0.03 (0–0.12)	0.52

Diagnostic performance was evaluated for patients with myocarditis or amyloidosis in the hold-out test set. Sensitivity, specificity, AUC, as well as their absolute mean differences are given. Data in parentheses are 95% CIs. Differences in ECV and vECV were determined by a McNemar's test and differences in AUC were determined by Delong's test. Both tests were conducted as permutation tests with 10000 resamples. AUC indicates area under the receiver operating curve; ECV, extracellular volume; and vECV, virtual extracellular volume.

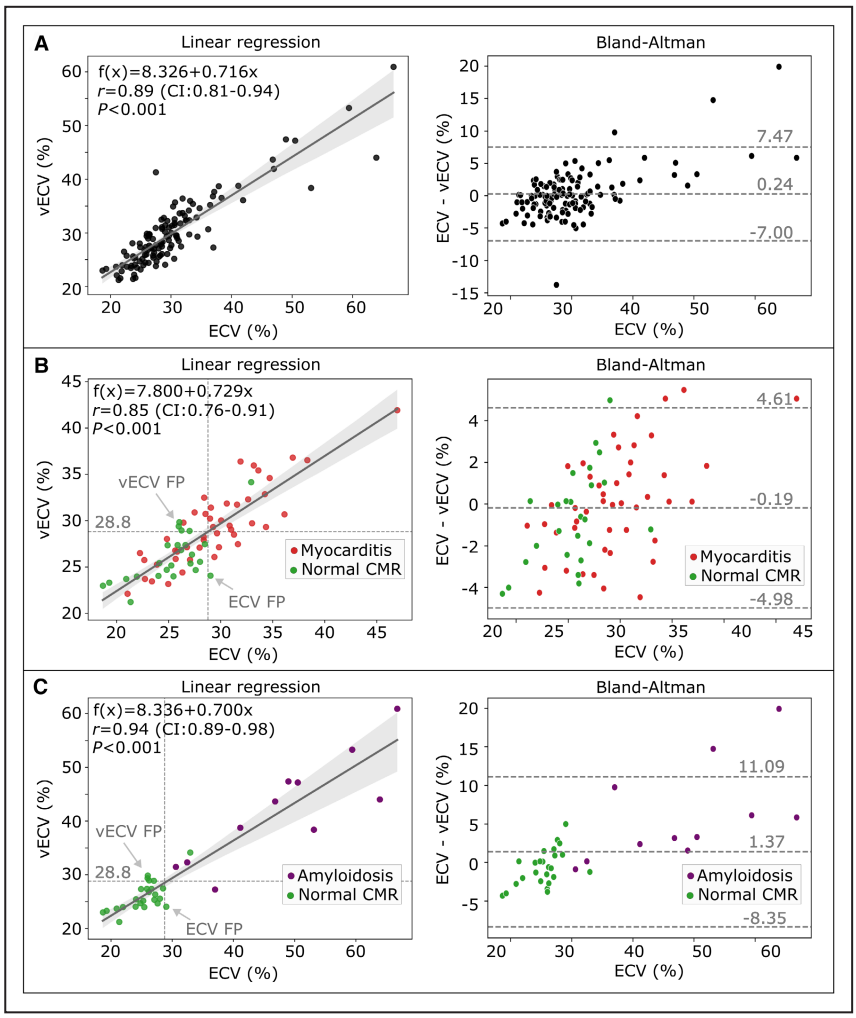


Figure 3. Linear regression and Bland–Altman analysis between conventional extracellular volume (ECV) and virtual extracellular volume (vECV) fraction for (A) using all test cases, (B) only using patients with normal cardiovascular magnetic resonance (CMR) results and patients with myocarditis and (C) only using patients with normal CMR and patients with amyloidosis.
In the linear regression plot, the applied ECV threshold of 28.8% for differentiating normal CMR to diseased is indicated by dotted lines. Cases that are detected false positive (FP) by vECV but true negative by ECV and vice versa are indicated by arrows. Conventional ECV was used as estimator of true ECV (*x* axis).

versus patients with normal CMR was observed (hold-out test: ECV AUC, 0.77 [95% CI, 0.65–0.87] versus vECV AUC, 0.76 [95% CI, 0.63–0.86]; *P*=0.76). Diagnostic performance results of vECV for the detection of cardiac amyloidosis are given in Table 4.

Correlation between conventional ECV and vECV was high in the hold-out test (all patients: *r*=0.89 [95% CI, 0.81–0.94], *P*<0.001; myocarditis subcohort: *r*=0.85 [95% CI, 0.76–0.91], *P*<0.001; amyloidosis subcohort: *r*=0.94 [95% CI, 0.89–0.98], *P*<0.001) (see Figure 3).

Examples of native, CE, and vCE T1 maps, as well as ECV and vECV maps are given in Figures 4 and 5. eventy-three of 123 (59%) patients in the hold-out test set had focal or diffuse lesions on the conventional ECV map. Of those, 38 of 73 (52%) patients featured also a lesion in the same location in the vECV. In 8 of 3 (11%) patients the vECV map showed also localized lesions, but at a different location than the lesion in the conventional ECV map. In 3 of 50 (4%) patients who had no lesion in the conventional ECV map, the vECV map falsely featured a focal lesion. Detailed results of visual evaluation of focal and diffuse lesions on the ECV maps separated for diagnosis are given in Table 5.

External Evaluation of Contrast-Free vECV

Differences in mean myocardial ECV and vECV are shown in Table 6 and linear regression and Bland-Altman analysis are shown in Figure 6. vECV based on

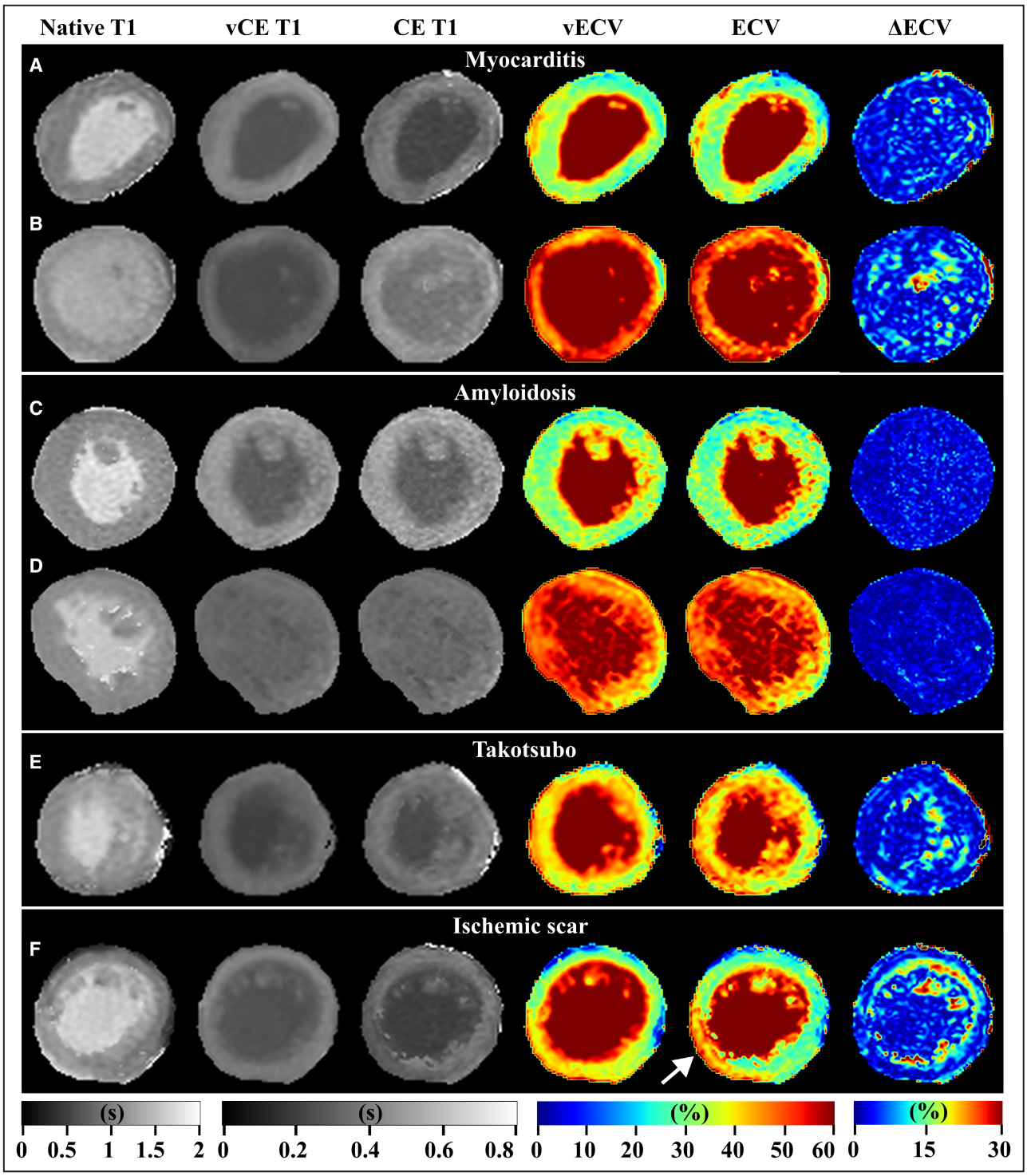


Figure 4. Clinical examples from the hold-out test set with global and focal agreement between conventional extracellular volume (ECV) and virtual extracellular volume (vECV).
Examples for patients with diagnosis of myocarditis (**A** and **B**), amyloidosis (**C** and **D**), Takotsubo syndrome (**E**) and ischemic scar (**F**, white arrow) are shown. Native, contrast-enhanced (CE) and virtual contrast-enhanced (vCE) T1 maps, as well as ECV, vECV, and pixel-wise ΔECV maps are shown.

the GAN developed on internal T1 maps generalized well to external T1 maps in patients with normal CMR (mean ECV, 23.0%±2.2% versus mean vECV, 23.6%±1.4%; P=0.12), with myocarditis (mean ECV, 26.8%±3.7% versus mean vECV, 26.8±3.3%; P=0.94) and with dilated cardiomyopathy (mean ECV, 35.3%±4.2% versus

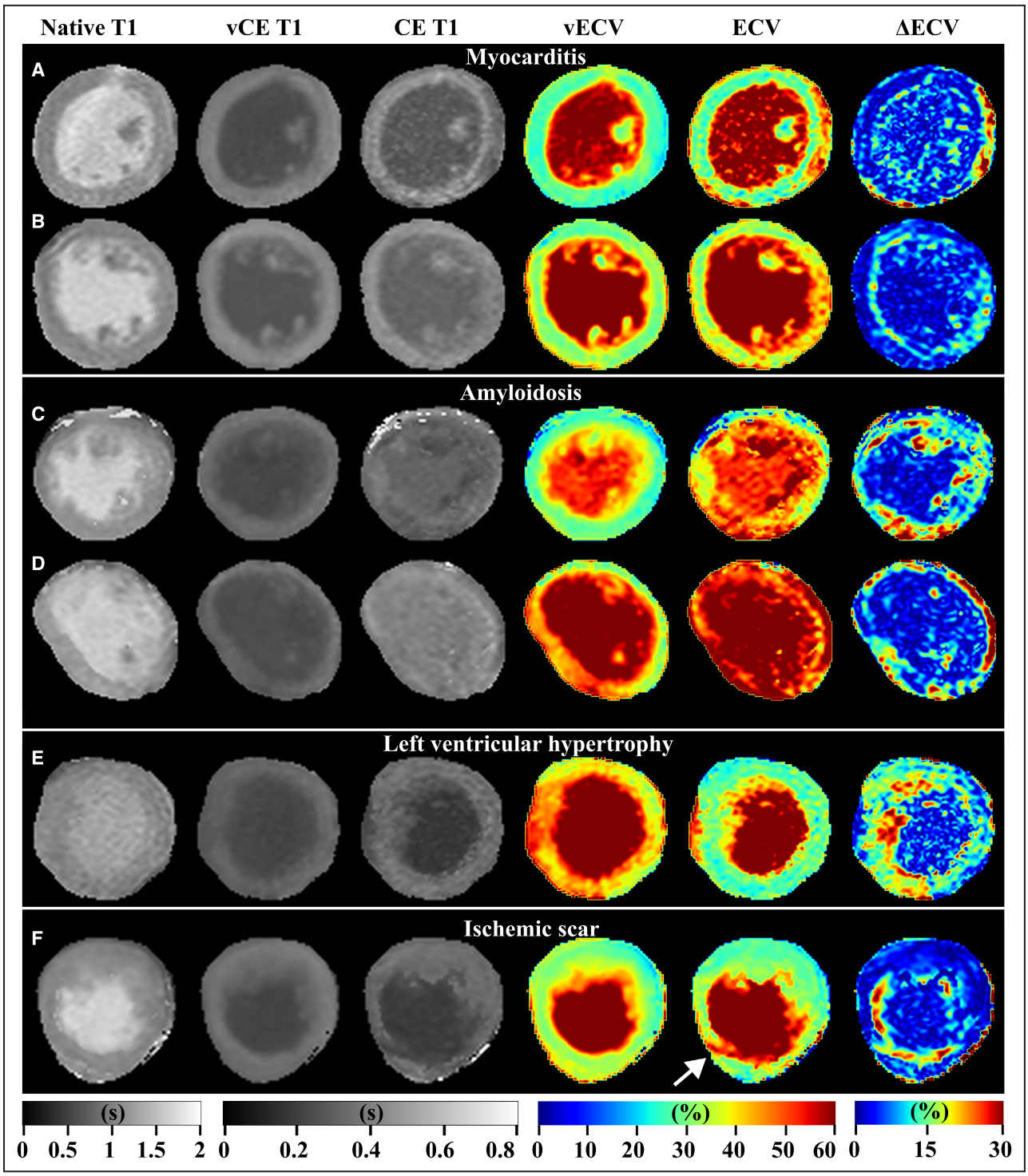


Figure 5. Clinical examples from the hold-out test set with low global and focal agreement between conventional extracellular volume (ECV) and virtual extracellular volume (vECV) maps.
Examples for patients with diagnosis of myocarditis (**A** and **B**), amyloidosis (**C** and **D**), Takotsubo syndrome (**E**) and ischemic scar (**F**, white arrow) are shown. Native, contrast-enhanced (CE) and virtual contrast-enhanced (vCE) T1 maps, as well as ECV, vECV, and pixel-wise ΔECV maps are shown.

mean vECV, 32.2%±4.2%; P=0.06). However, as in the internal hold-out test set, ECV and vECV differed for amyloidosis (mean ECV, 45.0%±7.2% versus mean vECV, 29.5%±3.4%; $P<0.01$).

DISCUSSION

In this explorative proof-of-principle study, we investigated the potential and limitations of deep learning for

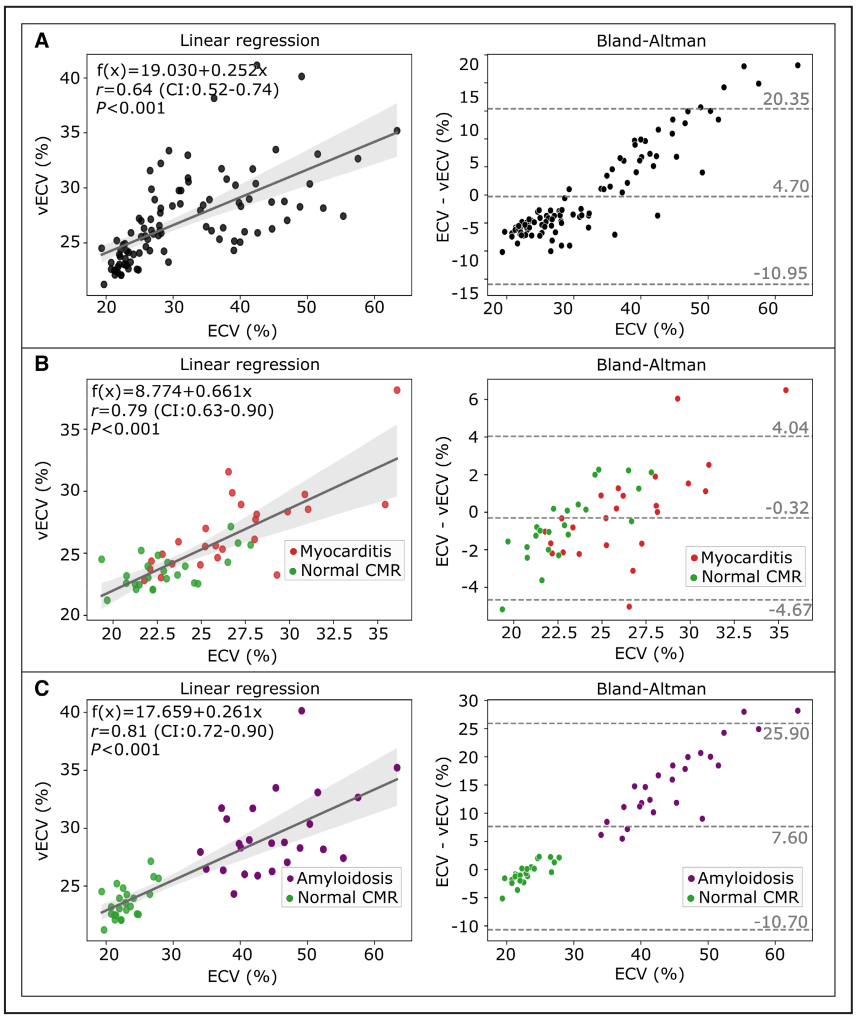


Figure 6. Linear regression and Bland–Altman analysis between conventional extracellular volume (ECV) and virtual extracellular volume (vECV) fraction for (A) using all cases of the external validation cohort, (B) only using patients with normal cardiovascular magnetic resonance (CMR) results and patients with myocarditis and (C) only using patients with normal CMR and patients with amyloidosis.
Conventional ECV was used as estimator of true ECV (*x* axis).

generating virtual CE T1 maps from native T1 maps to enable GBCA-free estimation of myocardial ECV. We could show that mean myocardial vECV values generated by a GAN had a substantial agreement with conventional ECV (hold-out test: *r*=0.89 [95% CI, 0.81–0.94]; ΔECV, 2.4%±2.8%) and that the application of vECV had similar diagnostic performance for the diagnosis of myocarditis and amyloidosis compared with conventional ECV (hold-out test: myocarditis ECV AUC, 0.77 [95% CI, 0.65–0.87] versus vECV AUC, 0.76 [95% CI, 0.63–0.86]; *P*=0.76; amyloidosis ECV AUC, 0.99 [95% CI, 0.97–1] versus vECV AUC, 0.96 [95% CI,0.86–1.00]; *P*=0.52). We could show that the GAN-based vECV had lower deviations to ECV compared with the application of random forest algorithms, indicating that the GAN-based T1 map generation does not simply represent a mapping of native to CE T1 values. Furthermore, we observed that the GAN model solely trained on internal T1 maps generalized well on an external data set in patients with normal CMR, myocarditis, and dilated cardiomyopathy. However, lower vECV values were observed in patients with amyloidosis in the internal hold-out test and in the external evaluation. One reason for this may be the relatively low proportion of patients with amyloidosis (8.7%) in the consecutively derived training data, which is lower than in patients with myocarditis (17.3%) or normal CMR (42.0%), for example. Nevertheless, the promising results of this proof-of-principle study for diseases with good representation motivate further exploring of deep learning for contrast-free vECV estimation in further studies with larger training data sets acquired from multiple centers, specifically targeting diseases with high ECV variance. Nowadays, myocardial mapping

Table 5. Results of Visual Assessment of Focal or Diffuse Lesions

Diagnosis	ECV with lesions, n (%)	Rating A, n (%)	Rating B, n (%)	Rating C, n (%)	ECV without lesion, n (%)	Rating D, n (%)
All patients (n=123)	73 (59)	38 (52)	8 (11)	34 (47)	50 (41)	3 (4)
Myocarditis (n=46)	34 (74)	13 (38)	4 (12)	20 (59)	12 (26)	1 (3)
Dilated cardiomyopathy (n=12)	9 (75)	6 (67)	1 (11)	3 (33)	3 (25)	2 (22)
Amyloidosis (n=11)	11 (100)	10 (91)	0 (0)	1 (9)	0 (0)	…
Ischemic cardiomyopathy (n=7)	5 (71)	2 (40)	0 (0)	3 (60)	2 (29)	0 (0)
Pericarditis (n=6)	3 (50)	1 (33)	0 (0)	2 (67)	3 (50)	0 (0)
Takotsubo/stress edema (n=5)	4 (80)	3 (75)	1 (25)	1 (25)	1 (20)	0 (0)
Myocardial fibrosis (n=4)	4 (100)	0 (0)	0 (0)	4 (100)	0 (0)	…
Left ventricular hypertrophy (n=3)	3 (100)	3 (100)	2 (67)	0 (0)	0 (0)	…
Other (n=4)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	…

Focal or diffuse lesions present in the conventional ECV and vECV maps of the patients were assessed in the hold-out test set. For each patient, the lesions that are present in the vECV maps were rated into the following categories: (A) lesions present in vECV in same location as in conventional ECV map; (B) lesions present in vECV in different location than in conventional ECV map; (C) no lesions present in vECV but in conventional ECV; and (D) lesions hallucinated, as none are present in conventional ECV map. Note that a patient can feature multiple lesions and thereby occur in the A and B rating simultaneously. For lesions in vECV, the proportion to the number of lesions in the conventional ECV is also given in brackets. ECV indicates extracellular volume; and vECV, virtual extracellular volume.

plays an important role in the diagnosis of patients with suspected cardiomyopathies as it increases diagnostic accuracy and reading confidence of CMR.[34] However, this might be especially true for native myocardial T1 and T2 mapping, as the clinical diagnostic value for an additional ECV assessment might be less impactful,[4,34] possibly due to a more cumbersome application of ECV in a "real world" setting. Even though ECV is listed as an independent T1 criterion in the 2018 Lake Louise criteria for the diagnosis of acute myocarditis, the diagnostic yield of ECV in acute myocarditis is lower than native T1 and T2 relaxation times.[4,35] So far, ECV has not clearly demonstrated an incremental diagnostic value beyond late gadolinium-enhanced and native mapping techniques in myocarditis CMR.[35] In the absence of significant myocardial inflammation, however, ECV is correlated with diffuse myocardial fibrosis.[36] This has prognostic implications in patients with left ventricular hypertrophy (eg, due to severe aortic stenosis).[37] In this regard, ECV is also important for the evaluation and detection of infiltrative and storage myocardial diseases including cardiac amyloidosis.[9] However, the assessment of ECV has some drawbacks and barriers in clinical routine, as additional CE T1 maps must be acquired (CE T1 maps are often acquired many minutes after the native maps, which limits registration in clinical routine), and timely hematocrit assessment is needed for accurate estimation.[38] Therefore, many centers might not routinely acquire CE T1 maps in clinical protocols and perform hematocrit blood test values before imaging due to additional costs.[38] In this regard, deep learning methods for vECV estimation, as proposed in our study, might facilitate a broader use of ECV in the future, as such algorithms can provide vECV maps immediately at the point of reading, like a point-of-care test, when the information is really needed. We demonstrated this by generating vECV for native T1 maps for which conventional

Table 6. External Evaluation of vECV Compared With ECV Fraction

CMR diagnosis	ECV (%)	vECV (%)	*P* value	ΔECV (%)
All patients (n=96)	31.7±9.9	27.0±3.9	<0.01*	6.0±7.1
Myocarditis (n=25)	26.8±3.7	26.8±3.3	0.94	1.9±1.7
Amyloidosis (n=25)	45.0±7.2	29.5±3.4	<0.01*	15.5±6.4
Normal (n=24)	23.0±2.2	23.6±1.4	0.12	1.5±1.2
LVH (n=16)	30.7±6.1	26.8±3.1	0.03*	5.3±5.0
Dilated cardiomyopathy (n=6)	35.3±4.2	32.2±4.2	0.06	3.3±3.1

ECV and vECV were calculated on the basis of contrast-enhanced T1 map and virtual contrast-enhanced T1 values generated by the generative adversarial network. Mean±SD, as well as ΔECV are given for ECV and vECV separated into subgroups on the basis of cardiovascular magnetic resonance diagnosis. *P* values were obtained using a permutation version of the 2-sided *t* test with 10000 resamples. ECV indicates extracellular volume; ΔECV, mean of the difference in extracellular volume per patient; and vECV, virtual extracellular volume.

*Statistical significance.

ECV could not be derived due to infeasible registration with the CE T1 map. Together with other algorithms, which can calculate virtual late gadolinium enhanced images,[16,17] faster and less expensive CMR without the use of GBCAs might become accessible in the future.

The not widespread acquisition of CE T1 maps in combination with missing hematocrit blood test values leads to a limited number of ground truth ECV in clinical databases compared with established imaging sequences such as late gadolinium enhanced. To avoid this limitation, we propose GBCA-free estimation of vECV by generating vCE T1 maps from native T1 maps. However, besides still requiring recent hematocrit blood values, this approach has some disadvantages: Imaging-specific factors, such as the time between contrast agent administration and CE T1 map acquisition, but also patient-dependent factors that influence how rapid the contrast agent is distributed and excreted, such as renal function, are subject to a certain degree of variability and influence CE myocardial T1.[3] To compensate for such factors, the CE T1 map in our study was acquired after a fixed interval of 10 minutes after contrast administration. In driven equilibrium ECV determination, the variability stemming from contrast agent distribution is normalized by bringing differences of native and CE R1 values of the myocardium and the blood pool into relation, which means that a deep learning model does not need additional information about possible influencing factors. This motivates future multicenter studies including more patients with appropriate CMR protocols and hematocrit to investigate a direct generation of vECV.

The proposed method has further limitations. On visual assessment of focal or diffuse lesions, we found out that only half of the lesions that were identified on the conventional ECV map were also present in the vECV map. Contrast agents provide additional information about the ECV and are especially important for identification of small lesions. Thus, the calculation of a perfect contrast-free vECV representing all small focal lesions is likely not possible solely on the basis of native T1 maps. Also, the myocardium of the vCE T1 and vECV maps have a “smoother” appearance, which is common for images generated by pixel-to-pixel GANs that optimize pixel-wise losses such as mean absolute distance or Euclidean distance.[39] In addition, in 8 patients with lesions in the conventional ECV, lesions were found in the vECV at other positions in the myocardium, and in 3 patients a lesion was falsely present in the vECV map as no lesions were found in the conventional ECV. The risk of hallucination of image content of GANs, which is facilitated by training with adversarial objectives by deceiving the discriminator, is particularly critical for applications in the medical field. One way to suppress the occurrence of hallucinated image content is to combine pixel-wise and adversarial losses, as applied in the current work.[40] Also, we did not investigate recently presented diffusion and vision transformer-based image-generation approaches.[41] However, vision transformers were shown to be only superior to convolutional neural networks, if vast amounts of training data were available, which was not the case in the current study.[42] Also, the standard U-Net already demonstrated capabilities for generation of contrast-agent CMR images in previous studies.[16,17] Therefore, we decided to evaluate U-Net–based GANs in this proof-of-principle study. Future work could compare transformer-based diffusion models with respect to quality and reliability of local tissue features. Finally, although the internal hold-out test and the external validation had different modified Look-Locker inversion-recovery acquisition schemes, the validation was performed with a single vendor and exclusively with 1.5T (1.5T Ingenia, Philips Healthcare).

In conclusion, deep learning–based contrast agent–free myocardial ECV estimation by generating vCE T1 maps is feasible and offers a clinically applicable approach for the determination of mean myocardial ECV values. This technique might ultimately facilitate faster CMR and increase patient safety, especially for those with impaired renal function. Further studies are warranted to ultimately determine the clinical implications of deep learning–based image generation in CMR.

ARTICLE INFORMATION

Received June 25, 2024; accepted August 19, 2024.

Affiliations

Department of Diagnostic and Interventional Radiology (S.N., L.M.B., A.I., M.T., W.B., C.C.P., D.K., U.I.A., A.M.S., J.A.L.), Quantitative Imaging Laboratory Bonn (QILaB) (S.N., L.M.B., A.I., M.T., W.B., D.K., A.M.S., J.A.L.), and Department of Internal Medicine II, Heart Center (S.Z., G.N.), University Hospital Bonn, Bonn, Germany; and Department of Diagnostic and Interventional Radiology, University Hospital Cologne, Cologne, Germany (L.P., K.K.).

Sources of Funding

This work was supported by the Open Access Publication Fund of the University of Bonn.

Disclosures

None.

Supplemental Material

Data S1–S4
Figures S1–S4

REFERENCES

1. Arbelo E, Protonotarios A, Gimeno JR, Arbustini E, Barriales-Villa R, Basso C, Bezzina CR, Biagini E, Blom NA, de Boer RA, et al. 2023 ESC guidelines for the management of cardiomyopathies: developed by the task force on the management of cardiomyopathies of the European Society of Cardiology (ESC). *Eur Heart J*. 2023;44:3503–3626. doi: 10.1093/eurheartj/ehad194
2. Luetkens JA, Homsi R, Dabir D, Kuetting DL, Marx C, Doerner J, Schlesinger-Irsch U, Andrié R, Sprinkart AM, Schmeel FC, et al.

Comprehensive cardiac magnetic resonance for short-term follow-up in acute myocarditis. *J Am Heart Assoc*. 2016;5:e003603. doi: 10.1161/JAHA.116.003603
3. Haaf P, Garg P, Messroghli DR, Broadbent DA, Greenwood JP, Plein S. Cardiac T1 mapping and extracellular volume (ECV) in clinical practice: a comprehensive review. *J Cardiovasc Magn Reson*. 2016;18:89. doi: 10.1186/s12968-016-0308-4
4. Luetkens JA, Faron A, Isaak A, Dabir D, Kuetting D, Feisst A, Schmeel FC, Sprinkart AM, Thomas D. Comparison of original and 2018 Lake Louise criteria for diagnosis of acute myocarditis: results of a validation cohort. *Radiol: Cardiothorac Imaging*. 2019;1:e190010. doi: 10.1148/ryct.2019190010
5. Luetkens JA, Homsi R, Sprinkart AM, Doerner J, Dabir D, Kuetting DL, Block W, Andrié R, Stehning C, Fimmers R, et al. Incremental value of quantitative CMR including parametric mapping for the diagnosis of acute myocarditis. *Eur Heart J Cardiovasc Imaging*. 2016;17:154–161. doi: 10.1093/ehjci/jev246
6. Cui Y, Cao Y, Song J, Dong N, Kong X, Wang J, Yuan Y, Zhu X, Yan X, Greiser A, et al. Association between myocardial extracellular volume and strain analysis through cardiovascular magnetic resonance with histological myocardial fibrosis in patients awaiting heart transplantation. *J Cardiovasc Magn Reson*. 2018;20:25. doi: 10.1186/s12968-018-0445-z
7. Luetkens JA, Klein S, Träber F, Schmeel FC, Sprinkart AM, Kuetting DLR, Block W, Uschner FE, Schierwagen R, Hittatiya K, et al. Quantification of liver fibrosis at T1 and T2 mapping with extracellular volume fraction MRI: preclinical results. *Radiology*. 2018;288:748–754. doi: 10.1148/radiol.2018180051
8. Isaak A, Praktiknjo M, Jansen C, Faron A, Sprinkart AM, Pieper CC, Chang J, Fimmers R, Meyer C, Dabir D, et al. Myocardial fibrosis and inflammation in liver cirrhosis: MRI study of the liver-heart axis. *Radiology*. 2020;297:51–61. doi: 10.1148/radiol.2020201057
9. Martinez-Naharro A, Patel R, Kotecha T, Karia N, Ioannou A, Petrie A, Chacko LA, Razvi Y, Ravichandran S, Brown J, et al. Cardiovascular magnetic resonance in light-chain amyloidosis to guide treatment. *Eur Heart J*. 2022;43:4722–4735. doi: 10.1093/eurheartj/ehac363
10. Grobner T. Gadolinium—a specific trigger for the development of nephrogenic fibrosing dermopathy and nephrogenic systemic fibrosis? *Nephrol Dial Transplant*. 2006;21:1104–1108. doi: 10.1093/ndt/gfk062
11. Kanda T, Fukusato T, Matsuda M, Toyoda K, Oba H, Kotoku J, Haruyama T, Kitajima K, Furui S. Gadolinium-based contrast agent accumulates in the brain even in subjects without severe renal dysfunction: evaluation of autopsy brain specimens with inductively coupled plasma mass spectroscopy. *Radiology*. 2015;276:228–232. doi: 10.1148/radiol.2015142690
12. Do C, DeAguero J, Brearley A, Trejo X, Howard T, Escobar GP, Wagner B. Gadolinium-based contrast agent use, their safety, and practice evolution. *Kidney360*. 2020;1:561–568. doi: 10.34067/KID.0000272019
13. Mallio CA, Radbruch A, Deike-Hofmann K, van der Molen AJ, Dekkers IA, Zaharchuk G, Parizel PM, Beomonte Zobel B, Quattrocchi CC. Artificial intelligence to reduce or eliminate the need for gadolinium-based contrast agents in brain and cardiac MRI: a literature review. *Investig Radiol*. 2023;58:746–753. doi: 10.1097/RLI.0000000000000983
14. Haase R, Pinetz T, Kobler E, Paech D, Effland A, Radbruch A, Deike-Hofmann K. Artificial contrast: deep learning for reducing gadolinium-based contrast agents in neuroradiology. *Investig Radiol*. 2023;58:539–547. doi: 10.1097/RLI.0000000000000963
15. Haase R, Pinetz T, Bendella Z, Kobler E, Paech D, Block W, Effland A, Radbruch A, Deike-Hofmann K. Reduction of gadolinium-based contrast agents in MRI using convolutional neural networks and different input protocols: limited interchangeability of synthesized sequences with original full-dose images despite excellent quantitative performance. *Investig Radiol*. 2023;58:420. doi: 10.1097/RLI.0000000000000955
16. Zhang Q, Burrage MK, Lukaschuk E, Shanmuganathan M, Popescu IA, Nikolaidou C, Mills R, Werys K, Hann E, Barutcu A, et al. Toward replacing late gadolinium enhancement with artificial intelligence virtual native enhancement for gadolinium-free cardiovascular magnetic resonance tissue characterization in hypertrophic cardiomyopathy. *Circulation*. 2021;144:589–599. doi: 10.1161/CIRCULATIONAHA.121.054432
17. Zhang Q, Burrage MK, Shanmuganathan M, Gonzales RA, Lukaschuk E, Thomas KE, Mills R, Leal Pelado J, Nikolaidou C, Popescu IA, et al. Artificial intelligence for contrast-free MRI: scar assessment in myocardial infarction using deep learning–based virtual native enhancement. *Circulation*. 2022;146:1492–1503. doi: 10.1161/CIRCULATIONAHA.122.060137
18. Messroghli DR, Radjenovic A, Kozerke S, Higgins DM, Sivananthan MU, Ridgway JP. Modified look-locker inversion recovery (MOLLI) for high-resolution T1 mapping of the heart. *Magn Reson Med*. 2004;52:141–146. doi: 10.1002/mrm.20110
19. Caforio ALP, Pankuweit S, Arbustini E, Basso C, Gimeno-Blanes J, Felix SB, Fu M, Heliö T, Heymans S, Jahns R, et al. Current state of knowledge on aetiology, diagnosis, management, and therapy of myocarditis: a position statement of the European Society of Cardiology Working Group on myocardial and pericardial diseases. *Eur Heart J*. 2013;34:2636–2648. doi: 10.1093/eurheartj/eht210
20. Garcia-Pavia P, Rapezzi C, Adler Y, Arad M, Basso C, Brucato A, Burazor I, Caforio ALP, Damy T, Eriksson U, et al. Diagnosis and treatment of cardiac amyloidosis: a position statement of the ESC working group on myocardial and pericardial diseases. *Eur Heart J*. 2021;42:1554–1568. doi: 10.1093/eurheartj/ehab072
21. Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation. *Medical Image Computing and Computer-Assisted Intervention–MICCAI*. In: Navab N, Hornegger J, Wells WM, Frangi A, eds. Springer International Publishing;2015:234–241. doi: 10.1007/978-3-319-24574-4_28
22. 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;31:8807–8815. doi: 10.1007/s00330-021-07858-1
23. Nowak S, Theis M, Wichtmann BD, Faron A, Froelich MF, Tollens F, Geißler HL, Block W, Luetkens JA, Attenberger UI, et al. End-to-end automated body composition analyses with integrated quality control for opportunistic assessment of sarcopenia in CT. *Eur Radiol*. 2022;32:3142–3151. doi: 10.1007/s00330-021-08313-x
24. Theis M, Tonguc T, Savchenko O, Nowak S, Block W, Recker F, Essler M, Mustea A, Attenberger U, Marinova M, et al. 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. doi: 10.1186/s13244-022-01342-0
25. Nowak S, Henkel A, Theis M, Luetkens J, Geiger S, Sprinkart AM, Pieper CC, Attenberger UI. Deep learning for standardized, MRI-based quantification of subcutaneous and subfascial tissue volume for patients with lipedema and lymphedema. *Eur Radiol*. 2023;33:884–892. doi: 10.1007/s00330-022-09047-0
26. Pasumarthi S, Tamir JI, Christensen S, Zaharchuk G, Zhang T, Gong E. A eneric deep learning model for reduced gadolinium dose in contrast-enhanced brain MRI. *Magn Reson Med*. 2021;86:1687–1700. doi: 10.1002/mrm.28808
27. Wang Z, Bovik AC, Sheikh HR, Simoncelli EP. Image quality assessment: from error visibility to structural similarity. *IEEE Trans Image Process*. 2004;13:600–612. doi: 10.1109/TIP.2003.819861
28. Flett AS, Hayward MP, Ashworth MT, Hansen MS, Taylor AM, Elliott PM, McGregor C, Moon JC. Equilibrium contrast cardiovascular magnetic resonance for the measurement of diffuse myocardial fibrosis. *Circulation*. 2010;122:138–144. doi: 10.1161/CIRCULATIONAHA.109.930636
29. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, Burovski E, Peterson P, Weckesser W, Bright J, et al. SciPy 1.0: fundamental algorithms for scientific computing in python. *Nat Methods*. 2020;17:261–272. doi: 10.1038/s41592-019-0686-2
30. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, et al. Scikit-learn: machine learning in python. *J Mach Learn Res*. 2011;12:2825–2830.
31. Seabold S, Perktold J. Statsmodels: econometric and statistical modeling with python. *Proceedings of the 9th Python in Science Conference*. In: van der Walt, S, Millman, J., eds. Vol 57. SciPy;2020:92–96.
32. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez J-C, Müller M. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*. 2011;12:77. doi: 10.1186/1471-2105-12-77
33. Venkatraman ES, Begg CB. A distribution-free procedure for comparing receiver operating characteristic curves from a paired experiment. *Biometrika*. 1996;83:835–848. doi: 10.1093/biomet/83.4.835
34. Warnica W, Al-Arnawoot A, Stanimirovic A, Thavendiranathan P, Wald RM, Pakkal M, Karur GR, Wintersperger BJ, Rac V, Hanneman K. Clinical impact of cardiac MRI T1 and T2 parametric mapping in patients with suspected cardiomyopathy. *Radiology*. 2022;305:319–326. doi: 10.1148/radiol.220067
35. Ferreira VM, Schulz-Menger J, Holmvang G, Kramer CM, Carbone I, Sechtem U, Kindermann I, Gutberlet M, Cooper LT, Liu P, et al.

Cardiovascular magnetic resonance in nonischemic myocardial inflammation: expert recommendations. *J Am Coll Cardiol*. 2018;72:3158–3176. doi: 10.1016/j.jacc.2018.09.072
36. Lurz JA, Luecke C, Lang D, Besler C, Rommel K-P, Klingel K, Kandolf R, Adams V, Schöne K, Hindricks G, et al. CMR-derived extracellular volume fraction as a marker for myocardial fibrosis: the importance of coexisting myocardial inflammation. *J Am Coll Cardiol Img*. 2018;11:38–45. doi: 10.1016/j.jcmg.2017.01.025
37. Everett RJ, Treibel TA, Fukui M, Lee H, Rigolli M, Singh A, Bijsterveld P, Tastet L, Musa TA, Dobson L, et al. Extracellular myocardial volume in patients with aortic stenosis. *J Am Coll Cardiol*. 2020;75:304–316. doi: 10.1016/j.jacc.2019.11.032
38. Shang Y, Zhang X, Zhou X, Wang J. Extracellular volume fraction measurements derived from the longitudinal relaxation of blood-based synthetic hematocrit may lead to clinical errors in 3 T cardiovascular magnetic resonance. *J Cardiovasc Magn Reson*. 2018;20:56. doi: 10.1186/s12968-018-0475-6
39. Sung TL, Lee HJ. Image-to-image translation using identical-pair adversarial networks. *Appl Sci*. 2019;9:2668. doi: 10.3390/app9132668
40. Lei K, Mardani M, Pauly JM, Vasanawala SS. Wasserstein GANs for MR imaging: from paired to unpaired training. *IEEE Trans Med Imaging*. 2021;40:105–115. doi: 10.1109/TMI.2020.3022968
41. Esser P, Kulal S, Blattmann A, Entezari R, Müller J, Saini H, Levi Y, Lorenz D, Sauer A, Boesel F, et al. Scaling Rectified Flow Transformers for High-Resolution Image Synthesis. arXiv. 2024:2403.03206.
42. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, et al. An Image Is Worth 16x16 Words: transformers for Image Recognition at Scale. arXiv. 2020:2010.11929.

3.5 Luetkens JA*, **Nowak S***, Mesropyan N, Block W, Praktiknjo M, Chang J, Bauckhage C, Sifa R, Sprinkart AM*, Faron A*, Attenberger UI*. Deep learning supports the differentiation of alcoholic and other-than-alcoholic cirrhosis based on MRI. Scientific reports. 2022;12(1):8297.–366. DOI: 10.1038/s41598-022-12410-2

Objective

Although CT and MRI are standard procedures in cirrhosis diagnosis, differentiation of etiology based on imaging is not established. This proof-of-concept study explores the potential of deep learning (DL) to support imaging-based differentiation of the etiology of liver cirrhosis.

Material and Methods

This retrospective, monocentric study included 465 patients with confirmed diagnosis of (a) alcoholic (n = 221) and (b) other-than-alcoholic (n = 244) cirrhosis. Standard T2-weighted single-slice images at the caudate lobe level were randomly split for training with fivefold cross-validation (85%) and testing (15%), balanced for (a) and (b). After automated upstream liver segmentation, two different ImageNet pre-trained convolutional neural network (CNN) architectures (ResNet50, DenseNet121) were evaluated for classification of alcohol-related versus non-alcohol-related cirrhosis.

Results

The highest classification performance on test data was observed for ResNet50 with unfrozen pre-trained parameters, yielding an area under the receiver operating characteristic curve of 0.82 (95% confidence interval (CI) 0.71–0.91) and an accuracy of 0.75 (95% CI 0.64–0.85). An ensemble of both models did not lead to significant improvement in classification performance.

Conclusion

This proof-of-principle study shows that deep-learning classifiers have the potential to aid in discriminating liver cirrhosis etiology based on standard MRI.

scientific reports

Check for updates

OPEN

Deep learning supports the differentiation of alcoholic and other-than-alcoholic cirrhosis based on MRI

Julian A. Luetkens[1,7], Sebastian Nowak[1,7], Narine Mesropyan[1], Wolfgang Block[1,2,3], Michael Praktiknjo[4], Johannes Chang[4], Christian Bauckhage[5,6], Rafet Sifa[6], Alois Martin Sprinkart[1,7]✉, Anton Faron[1,7] & Ulrike Attenberger[1,7]

Although CT and MRI are standard procedures in cirrhosis diagnosis, differentiation of etiology based on imaging is not established. This proof-of-concept study explores the potential of deep learning (DL) to support imaging-based differentiation of the etiology of liver cirrhosis. This retrospective, monocentric study included 465 patients with confirmed diagnosis of (a) alcoholic (n = 221) and (b) other-than-alcoholic (n = 244) cirrhosis. Standard T2-weighted single-slice images at the caudate lobe level were randomly split for training with fivefold cross-validation (85%) and testing (15%), balanced for (a) and (b). After automated upstream liver segmentation, two different ImageNet pre-trained convolutional neural network (CNN) architectures (ResNet50, DenseNet121) were evaluated for classification of alcohol-related versus non-alcohol-related cirrhosis. The highest classification performance on test data was observed for ResNet50 with unfrozen pre-trained parameters, yielding an area under the receiver operating characteristic curve of 0.82 (95% confidence interval (CI) 0.71–0.91) and an accuracy of 0.75 (95% CI 0.64–0.85). An ensemble of both models did not lead to significant improvement in classification performance. This proof-of-principle study shows that deep-learning classifiers have the potential to aid in discriminating liver cirrhosis etiology based on standard MRI.

As the end stage of chronic liver disease, liver cirrhosis is a major health issue. In particular, patients with liver cirrhosis have a concomitant risk for the development of hepatocellular carcinoma as well as complications arising from decompensation such as variceal bleeding or hepatic encephalopathy. Overall, the prevalence of chronic liver disease is internationally expected to grow within the next decades[1–3].

Many factors that contribute to the development of cirrhosis have been identified. The most common etiologies are alcohol consumption, obesity, and chronic viral infections[3]. Thereby, identification of the underlying cause of disease is crucial as appropriate treatment may not only stop disease from progression, but in certain cases also may facilitate regression of fibrosis[2,4,5]. In the majority of countries, alcohol consumption still represents the leading cause of liver disease and is directly related to liver mortality[3,6]. In these patients, alcohol abstinence was shown to be critical for long-term outcome, may improve various aspects of disease severity and is also fundamental with regard to potential liver transplantation[4,7,8]. However, until experiencing severe complications such as acute decompensation, many patients with cirrhosis are unaware of their underlying condition[2,3]. Liver cirrhosis results from chronic inflammation and hence leads to distinct changes in hepatic morphology, which in part can be detected by high-resolution imaging methods such as MRI[9,10].

[1]Department of Diagnostic and Interventional Radiology, Quantitative Imaging Lab Bonn (QILaB), University Hospital Bonn, Venusberg-Campus 1, 53127 Bonn, Germany. [2]Department of Radiotherapy and Radiation Oncology, University Hospital Bonn, Venusberg-Campus 1, 53127 Bonn, Germany. [3]Department of Neuroradiology, University Hospital Bonn, Venusberg-Campus 1, 53127 Bonn, Germany. [4]Department of Internal Medicine I, Center for Cirrhosis and Portal Hypertension Bonn (CCB), University Hospital Bonn, Venusberg-Campus 1, 53127 Bonn, Germany. [5]Institute for Computer Science, University of Bonn, Endenicher Allee 19C, 53113 Bonn, Germany. [6]Media Engineering Department, Fraunhofer IAIS, Schloss Birlinghoven 1, 53757 Sankt Augustin, Germany. [7]These authors contributed equally: Julian A. Luetkens, Sebastian Nowak, Alois Martin Sprinkart, Anton Faron and Ulrike Attenberger. ✉email: Sprinkart@uni-bonn.de

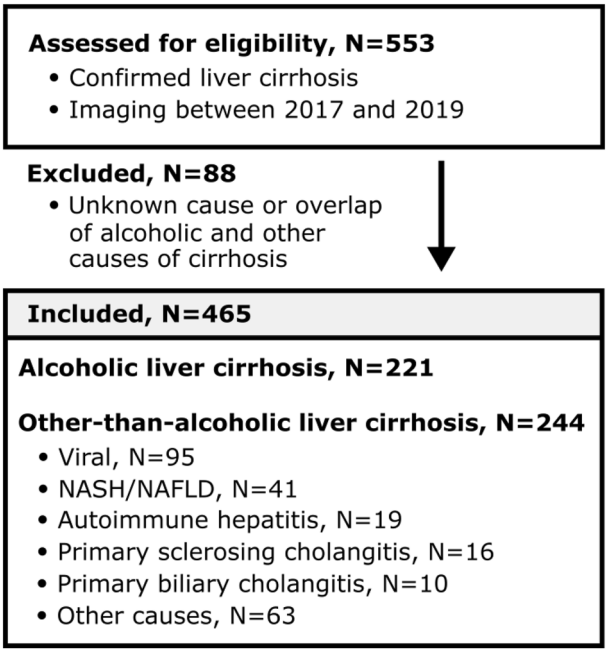


Figure 1. Study inclusion flow chart. Patients with confirmed diagnosis of liver cirrhosis who underwent liver MRI between 2017 and 2019 were evaluated for inclusion. Patients with unknown causes of liver cirrhosis and with documented overlap of alcoholic liver cirrhosis with other causes were excluded from the analysis. The final cohort consisted of 465 patients. Those patients were separated according to liver cirrhosis etiology into patients with (**a**) alcoholic liver cirrhosis (N = 221) and (**b**) other-than-alcoholic liver cirrhosis (N = 244). Abbreviations: *NAFLD/NASH* non-alcoholic fatty liver disease/non-alcoholic steatohepatitis.

Although it has been described that the micro- and macroscopic appearance of cirrhosis in medical imaging varies to some extent according to the underlying etiology, the use of imaging features as a means to determine the cause of the disease has not been established yet[9,11,12]. In a previous study, a convolutional neural network (CNN) was already shown to be able to detect liver cirrhosis based on standard clinical MRI sequences with expert-level accuracy irrespective of etiology[13]. Therefore, the aim of this proof-of-concept study was to investigate deep learning for standard MRI based characterization of disease etiology, differentiating alcoholic- versus other-than-alcoholic cirrhosis.

Materials and methods

Dataset. The study was approved by the Ethics Committee at the Medical Faculty of the Rheinische Friedrich-Wilhelms-Universität Bonn and the need for written informed consent was waived due to its retrospective, single-center nature. The research was performed in accordance with the Declaration of Helsinki. Patients with confirmed diagnosis of liver cirrhosis, defined by clinical manifestations of liver cirrhosis (e.g. presence of dermal features, ascites, splenomegaly or hyperestrogenism), laboratory parameters (e.g. presence of parameters of hepatocyte damage or impaired hepatic synthesis), and/or histopathological criteria, who underwent liver MRI for diagnostic purposes between 2017 and 2019 at the Department of Diagnostic and Interventional Radiology at the University Hospital of Bonn, were evaluated for inclusion. The clinical information management system of the relevant institution was used to derive clinical characteristics of the study population including the respective cause of liver cirrhosis. Patients with unknown causes of liver cirrhosis and with overlap of alcoholic and other causes were excluded. The final cohort was separated according to the underlying cause of liver cirrhosis into (a) patients with alcoholic liver cirrhosis and (b) other-than-alcoholic liver cirrhosis (Fig. 1).

Image segmentation and classification. All patients underwent a standardized imaging protocol including a standard clinical respiratory triggered multi-slice turbo spin echo sequence with non-cartesian k-space filling (T2 MultiVane XD) on a clinical 1.5 Tesla (Ingenia 1.5 T, Philips Healthcare, Best, the Netherlands) or 3.0 Tesla (Ingenia 3.0 T, Philips Healthcare, Best, the Netherlands) scanner. This sequence was shown to be suitable for deep learning-based detection of liver cirrhosis in a previous study[13]. Similar to the proposed approach for cirrhosis detection, a single cross-sectional image at the level of caudate lobe was exported, followed by liver segmentation performed by a U-net style convolutional neural network (CNN) with ResNet34 as backbone that was developed and validated on a dataset of 713 single slice T2- weighted MRI images[13]. The images were first normalized and image augmentation was applied during training. Supplementary information on imaging parameters and image preprocessing can be found in Supplement S1 and S2.

For imaging development of a classification CNN that differentiates patients with alcoholic liver cirrhosis and other-than-alcoholic liver cirrhosis, data were randomly split into a training (85%) and a hold-out test set

(15%). Training was performed with fivefold cross-validation. An ensemble of the cross-validated models was applied to the test set.

A CNN with residual connections (ResNet50) with ImageNet pre-trained parameters was used, as this established architecture was shown to be suitable for the detection of liver cirrhosis[13,14]. To investigate whether the use of a different pre-trained architecture than ResNet50 or an ensemble of two architectures is beneficial, a CNN with dense connections (DenseNet121) was additionally evaluated, which has fewer trainable parameters and is less computationally intensive compared to ResNet50[15].

Furthermore, two different training strategies were evaluated for ResNet50 and DenseNet121 in order to examine whether altering the pre-trained parameters of the CNN may impact classification performance. First, both networks were trained with frozen pre-trained parameters of the convolutional layers. In a second subsequent training run, the pre-trained convolutional layers of both networks were unfrozen with descending learning rates from the last to the first layer at several stages. Training was performed with Adam optimization, a cyclical learning rate scheme, and cross-entropy loss function. Supplementary information on the experimental design and hyper-parameters used for training are provided in Supplement S3.

Image regions that were particularly relevant to the classification task were highlighted by generating gradient-weighted class activation maps (Grad-CAM) for the test set[16].

Statistical analysis. Prism 8 (GraphPad software), SPSS Statistics 24 (IBM), MedCalc 20.014 (MedCalc Software Ltd) and Scikit-learn 0.23.2[17] were used for statistical analysis. Patient characteristics are expressed as frequencies or means with standard deviation, as appropriate. Classification accuracy (ACC), as well as receiver operating characteristic (ROC) analyses was performed for the cross-validation and the test sets for both studied CNN architectures (ResNet50, DenseNet121) and both training strategies (frozen, unfrozen). For the test set, 95% confidence intervals were determined for ACC and AUC values. ROC and precision-recall curves were generated[18]. Grad-CAM images were visually inspected by one experienced radiologist (A.F.) and highlighted regions were categorized according to their anatomical localization as being predominantly situated in the right liver lobe, the left liver lobe, the portal region, the caudate lobe, or in the image background. Resulting categorical data were compared using either Fisher`s exact test (for a cell count of ≤ 5) or χ^2 test (for a cell count > 5), as appropriate. The two-sided t-Test was used to compare differences between groups regarding continuous variables. $P < 0.05$ was set as the level of statistical significance.

Results

Baseline characteristics of the study population. A total of 465 patients (203 female; mean age, 60 ± 11 years) were included. Of those, 47.5% (221/465) of patients had alcoholic liver cirrhosis. 52.5% (244/465) of patients had other-than-alcoholic liver cirrhosis.

Liver biopsy was performed in 64.8% (301/465) of patients. The most common causes of non-alcohol related liver cirrhosis were viral hepatitis (39%, 95/244), non-alcoholic fatty liver disease or non-alcoholic steatohepatitis (17%, 41/244), and autoimmune hepatitis (8%, 19/244). In 5% (12/244) of patients with other-than-alcoholic liver cirrhosis, etiology of liver disease was multifactorial. Causes of liver cirrhosis of the entire study population are summarized in detail by Table 1. No significant differences regarding age (61 ± 9 years vs. 59 ± 13 years, $P = 0.110$) and gender distribution (48%, 105/221 female patients vs. 40%, 98/244 female patients, $P = 0.111$) were observed between patients with alcoholic and other-than-alcoholic liver cirrhosis. There was no difference in weight between patients with alcoholic and other-than-alcoholic cirrhosis (80.1 ± 20.4 kg vs. 80.1 ± 17.7 kg, $P = 0.972$). Values for γ-glutamyltransferase were higher in patients with alcoholic cirrhosis compared to patients with other-than-alcoholic cirrhosis (208.8 ± 264.1 U/l vs. 147.4 ± 166.1 U/l, $P = 0.003$).

Classification of liver cirrhosis etiology. Segmented images of the entire study population were randomly subdivided into a training (N = 396; 174 female; mean age, 60 ± 12 years), and a test set (N = 69; 29 female; mean age, 59 ± 10 years), with training sets being further split for fivefold cross-validation, balanced for patients with alcoholic and non-alcoholic cause of liver cirrhosis.

Trained with frozen parameters, a mean accuracy (ACC) and mean area under the curve (AUC) of 0.69 and 0.78 was observed for ResNet50 and a mean ACC of 0.66 and a mean AUC of 0.78 was observed for DenseNet121 for all 5 validation splits (Table 2). With unfrozen pre-trained parameters, mean ACC values of 0.74 and 0.71 and mean AUC values of 0.83 and 0.82 were obtained for ResNet50 and DenseNet121 on cross-validated training data, respectively.

On test data, the classification performance of ResNet50 was higher than DenseNet121 when training with unfrozen parameters, however the difference was not statistically significant (Value and 95% CI: $AUC_{ResNet50}$ 0.82 [0.71–0.91] vs. $AUC_{DenseNet121}$ 0.79 [0.67–0.88], $P = 0.40$; $ACC_{ResNet50}$ 0.75 [0.64–0.85] vs. $ACC_{DenseNet121}$ 0.70 [0.57–0.80], $P = 0.26$). Also, training with unfrozen parameters did not differ significantly compared to frozen parameters for both ResNet50 and DenseNet121 ($AUC_{ResNet50}$, 0.82 [0.71–0.91] vs. 0.82 [0.71–0.90], $P = 0.91$; $ACC_{ResNet50}$, 0.75 [0.64–0.85] vs. 0.74 [0.62–0.84], $P = 0.78$; $AUC_{DenseNet121}$, 0.79 [0.67–0.88] vs. 0.80 [0.69–0.89], $P = 0.69$; $ACC_{DenseNet121}$, 0.70 [0.57–0.80] vs. 0.74 [0.62–0.84], $P = 0.43$).

The ensemble of the two architectures did not lead to statistically significant improvement on the test set compared to ResNet50, neither for frozen ($AUC_{Ensemble}$ 0.84 [0.73–0.92] vs. $AUC_{ResNet50}$ 0.82 [0.71–0.90], $P = 0.54$; $ACC_{Ensemble}$ 0.75 [0.64–0.85] vs. $ACC_{ResNet50}$ 0.74 [0.62–0.84], $P = 0.78$), nor for unfrozen ($AUC_{Ensemble}$ 0.81 [0.70–0.90] vs. $AUC_{ResNet50}$ 0.82 [0.71–0.91], $P = 0.70$.; $ACC_{Ensemble}$ 0.71 [0.59–0.81], vs. $ACC_{ResNet50}$ 0.75 [0.64–0.85], $P = 0.40$) pre-trained parameters.

ROC and precision-recall curves of the models trained with unfrozen pre-trained parameters are given in Fig. 2.

Etiology of liver cirrhosis	Number of patients (%)
Alcoholic liver cirrhosis	221 (48%)
Other-than-alcoholic liver cirrhosis	244 (52%)
Hepatitis B virus	26 (6%)
Hepatitis C virus	69 (15%)
Fatty liver disease (NAFLD/NASH)	41 (9%)
Autoimmune hepatitis	19 (4%)
Primary sclerosing cholangitis	16 (3%)
Drug-induced	13 (3%)
Primary biliary cholangitis	10 (2%)
Portal vein thrombosis	9 (2%)
Nutritional	9 (2%)
Budd-Chiari syndrome	9 (2%)
Hemochromatosis	5 (1%)
Idiopathic	5 (1%)
Sinusoidal obstruction syndrome	3 (1%)
Secondary sclerosing cholangitis	3 (1%)
Alpha-1 Antitrypsin Deficiency	3 (1%)
Wilson disease	2 (< 1%)
Congestive hepatopathy	1 (< 1%)
Sarcoidosis	1 (< 1%)

Table 1. Liver cirrhosis etiology. Underlying causes of liver cirrhosis are reported for the entire study population (N = 465) as total numbers as well as percentages of the entire study cohort. *NAFLD/NASH* non-alcoholic fatty liver disease/non-alcoholic steatohepatitis.

	Frozen pre-trained parameters						Unfrozen pre-trained parameters					
	ResNet50		DenseNet121		Ensemble		ResNet50		DenseNet121		Ensemble	
	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC
Training + validation												
Split 1	0.737	0.684	0.764	0.671	0.776	0.658	0.798	0.696	0.780	0.671	0.807	0.709
Split 2	0.774	0.658	0.751	0.646	0.768	0.747	0.773	0.722	0.839	0.684	0.798	0.684
Split 3	0.800	0.722	0.797	0.646	0.815	0.722	0.864	0.785	0.817	0.772	0.876	0.797
Split 4	0.821	0.709	0.822	0.658	0.852	0.684	0.879	0.785	0.858	0.722	0.870	0.759
Split 5	0.742	0.671	0.756	0.684	0.770	0.722	0.822	0.722	0.805	0.696	0.813	0.709
Mean	0.775	0.689	0.778	0.661	0.796	0.707	0.827	0.742	0.820	0.709	0.833	0.732
Test												
	0.819	0.739	0.801	0.739	0.838	0.754	0.823	0.754	0.786	0.696	0.813	0.710

Table 2. Classification performance of the cross-validation and testing of the CNN architectures trained with frozen and unfrozen pre-trained parameters. Classification accuracy and AUC values for each validation split of the cross-validation and mean over all splits. The classification accuracy and AUC values of ensembles of the cross-validated models on the test set. *AUC* area under the curve, *ACC* accuracy.

Highlighted imaging regions according to Grad-CAM. The decision process to classify liver cirrhosis as being alcohol related or non-alcohol related was further visualized using Grad-CAM analysis for ResNet50 trained with unfrozen pre-trained parameters. According to Grad-CAM analysis, the right liver lobe (alcoholic liver cirrhosis 42%, 14/33; other-than-alcoholic liver cirrhosis 61%, 22/36) and the portal area (alcoholic liver cirrhosis 30%, 10/33; other-than-alcoholic liver cirrhosis 19%, 7/36) were the imaging regions that were most frequently decisive for the classification process in both groups. Thereby, no significant differences regarding distribution of decisive imaging regions between the two patient groups were observed (Table 3). Exemplary images of the Grad-CAM analysis are provided in Fig. 3.

Discussion

The purpose of this study was to investigate whether a deep learning-based analysis can aid in differentiating the etiology of liver cirrhosis based on routine clinical T2-weighted MRI. Acceptable to excellent discriminatory ability was found in distinguishing patients with alcoholic and other-than-alcoholic cirrhosis. In a previous

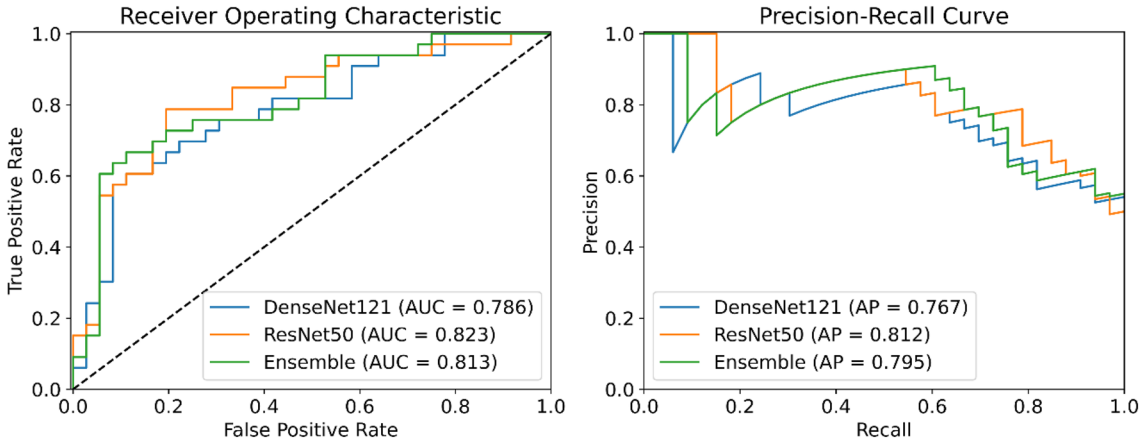


Figure 2. Receiver operating characteristic and precision-recall analysis for the classification performance of DenseNet121 and ResNet50, both trained with unfrozen pre-trained parameters. Abbreviations: *AUC* area under the curve, *AP* average precision.

	Alcoholic liver cirrhosis (N = 33)	Other-than-alcoholic liver cirrhosis (N = 36)	*P* value
Right lobe	14 (42%)	22 (61%)	0.12
Left lobe	3 (9%)	3 (8%)	1.00
Portal area	10 (30%)	7 (19%)	0.30
Caudate lobe	4 (12%)	1 (3%)	0.19
background	2 (6%)	3 (8%)	1.00

Table 3. Highlighted imaging regions according to gradient-weighted class activation maps (Grad-CAM). Results of the visual inspection of Grad-CAM images classified by ResNet50 are provided. Within each segmented image of the test set, highlighted regions were visually rated as being primary located within the right liver lobe, the left liver lobe, the portal area, the caudate lobe, or within image background by one radiologist experienced in abdominal imaging (A.F.).

study, a ResNet50 with frozen pre-trained ImageNet parameters was proposed for automatic detection of liver cirrhosis on T2-weighted MRI[13]. The results of our proof-of-concept study extend the findings of this previous report and show that deep learning not only enables the detection of cirrhosis, but can also help in identifying the underlying cause of the disease.

Although the ability of the ImageNet pre-trained ResNet50 to discriminate between alcoholic and other-than-alcoholic cirrhosis can be described as excellent[19], it is inferior to the differentiation of cirrhotic versus non-cirrhotic livers[13]. This may be due to less distinctiveness between imaging criteria indicating different causes of the disease compared to image criteria distinguishing a diseased organ from a non-diseased organ. For instance, it has been described that a hypertrophic appearance of the central hepatic parenchyma/caudate lobe is expected in alcohol-related cirrhosis, but also in primary sclerosing cholangitis and Budd-Chiari syndrome related cirrhosis[11].

Of both models investigated in the current study, ResNet50 showed higher classification performance on test data. However, the performance was not significantly higher compared to Densenet121. Interestingly, for both CNNs, subsequent training with unfrozen pre-trained parameters did not significantly increase classification performance on test data. This may suggest that the extraction capabilities of general imaging features of the convolutional kernels, learned during the pre-training with the ImageNet database, generalize well to T2-weighted MRI images. An ensemble of the two models trained with unfrozen parameters achieved equal accuracy and a slightly higher AUC compared to ResNet50, however, the difference was not statistically significant. Therefore, no clear advantage was observed by using an ensemble of the two different pre-trained ImageNet architectures.

Grad-CAM-analysis indicate that the imaging morphology of the right liver lobe and caudate lobe seem to comprise particularly relevant information for discrimination of alcoholic from other-than-alcoholic liver cirrhosis. This is in line with previous studies, which describe that the right posterior hepatic notch sign, defined as a sharp liver surface indentation at the posterior boundary of the right and caudate lobe, is considered to be particularly prevalent among patients with alcoholic liver cirrhosis[12,20]. As described above, hypertrophies of the caudate lobe and central hepatic areas are more frequently observed in patients with alcohol-related diseases, but are also seen in other etiologies. To the best of our knowledge, there are currently no studies presenting metrics for the diagnostic accuracy of cirrhosis etiology based on such imaging criteria[11,12]. However, a very recent work investigated a radiomics approach that relates imaging features to the etiology of liver cirrhosis, and also achieved promising results[21]. Unlike the deep learning method presented in the current study, the proposed radiomics approach requires manual definition of region of interests. To date, imaging features have not been used in routine clinical practice to identify alcohol as a cause of cirrhosis.

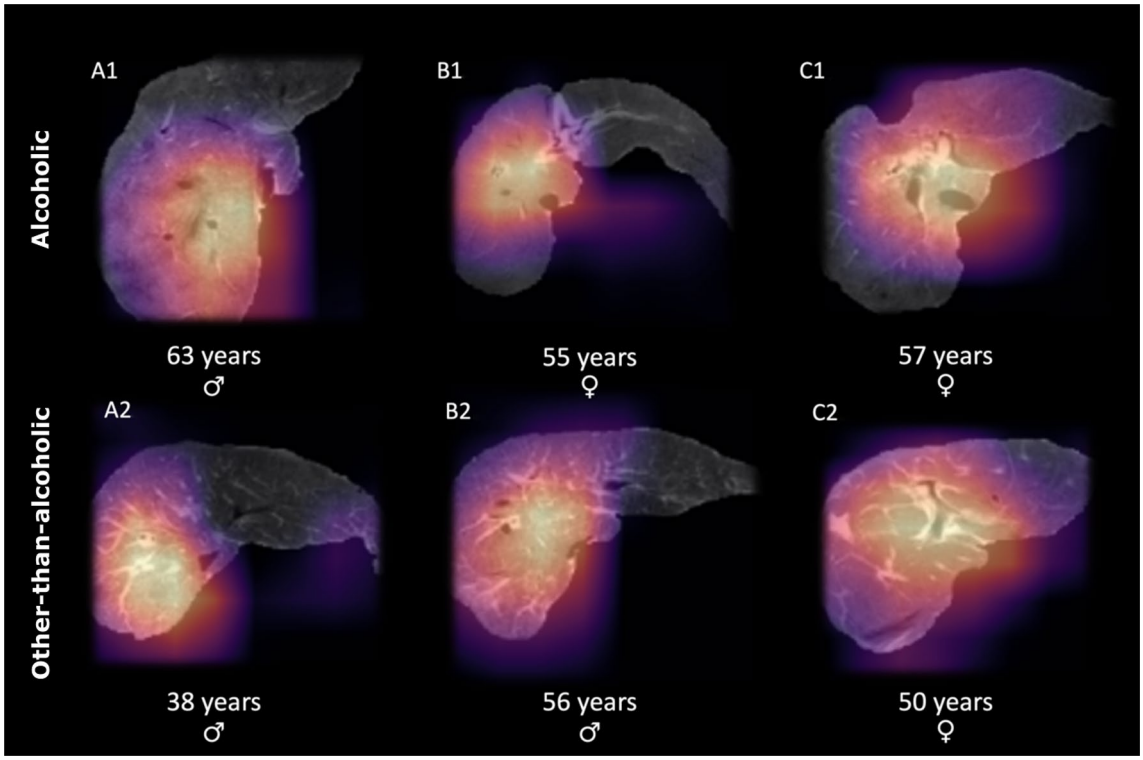


Figure 3. Exemplary images from the study population. ResNet50 trained with unfrozen pre-trained parameters was used for the classification task. Exemplary patients from the test set are provided and imaging regions that were particularly relevant to the classification task are highlighted using the gradient-weighted class activation maps (Grad-CAM) method. Panels A1, B1, C1 provide exemplary patients from the test set with alcoholic liver cirrhosis. In panels A2, B2, C2, images of exemplary patients from the test set with other-than-alcoholic liver cirrhosis are presented. In panels A1, B1, A2, B2, regions within the right liver lobe appeared to be particularly relevant for the classification task, as indicated by Grad-CAM images. In panels C1 and C2, the portal liver region appeared to be most decisive for classification.

In clinical routine, liver cirrhosis is typically diagnosed by a combination of characteristic clinical and imaging findings, corresponding laboratory testing and ancillary examinations such as abdominal sonography. Thereby, while this work-up is usually straightforward for virus-related cirrhosis, it may be much more effortful in patients with alcohol-related disease, which in many cases may be diagnosed only by exclusion since there are no specific laboratory findings[22]. Liver biopsy is recommended if cirrhosis etiology is uncertain, but is limited due to its invasive nature, inter-observer variability and potential sampling error[2,23]. Moreover, cirrhosis-related parenchymal changes may hamper or even preclude correct histological analysis[2].

Compensated liver cirrhosis is frequently asymptomatic; thus, it may be assumed that many patients who undergo routine clinical MRI for other indications may be unaware of a concomitant liver disease. In these patients, a pipeline that automatically identifies tissue alterations and can classify possible disease etiologies has the potential to better guide diagnostic pathways and thus initiate a specific therapy earlier. With the help of deep learning algorithms simple cross-sectional imaging modalities could serve as imaging-based biomarkers for the classification of liver disease in the future. Particularly in alcoholic liver cirrhosis, timely and correct identification of the underlying etiology is crucial, as early abstinence was demonstrated to be the key determinant of long-term outcome[8,24]. Sole clinical assessment of alcoholic liver disease alone might not be trivial in clinical practice because it mostly relies on patients' self-report. In this regard, deep learning applications have the potential to aid diagnosis by extracting also relevant information that may not be readily apparent to the human eye.

Our study has several limitations. The algorithm was developed for binary classification only and does currently not support differentiation of various non-alcohol-related cirrhosis etiologies. Due to the limited number of patients within the respective subclasses and to ensure collectives of comparable size for classification, we decided to pool patients with other-than-alcoholic cirrhosis. Future studies with larger samples of the respective subgroups are needed to substantiate the findings from this proof-of-concept study and to expand its application. The clinical benefit would also be significantly increased by an extension to other etiologies. Especially, NAFLD is becoming the main cause of chronic liver disease in many countries and the detection of metabolic related cirrhosis on cross-sectional imaging should be further explored in future studies. Also, we were not able to analyze possible coexisting etiologies of liver cirrhosis in our explorative analysis, as detailed data on additional risk factors were not available due to the retrospective study design. However, future studies should evaluate the ability of deep learning methods to differentiate overlapping liver disease, such as both alcoholic and non-alcoholic steatohepatitis (BASH). Moreover, we exclusively used single-slice T2-weighted images of segmented livers and ImageNet pre-trained models, as these have been shown to be suitable for the detection of liver cirrhosis in a

previous study[13]. Future studies may also address a three-dimensional approach accounting also for extrahepatic manifestations in cirrhotic patients or the use of other imaging sequences for differentiation of etiologies.

In summary, the results of this proof-of-principle study demonstrate that discrimination between alcoholic and other-than-alcoholic cirrhosis based on clinical T2-weighted single-slice images is feasible with acceptable to excellent discrimination ability. This indicates the potential of deep learning for a more comprehensive assessment of diffuse liver disease.

Data availability

The data sets analysed in this study are subject to data protection law and are therefore not publicly available.

Received: 11 November 2021; Accepted: 9 May 2022
Published online: 18 May 2022

References

1. Kamath, P. S. Acute on chronic liver failure. *Clin. Liver Dis.* **9**(4), 86–88 (2017).
2. Wiegand, J. & Berg, T. The etiology, diagnosis and prevention of liver cirrhosis: Part 1 of a series on liver cirrhosis. *Dtsch. Arztebl. Int.* **110**(6), 85–91 (2013).
3. Pimpin, L. *et al.* Burden of liver disease in Europe: Epidemiology and analysis of risk factors to identify prevention policies. *J. Hepatol.* **69**(3), 718–735 (2018).
4. Huang, Y. W., Yang, S. S. & Kao, J. H. Pathogenesis and management of alcoholic liver cirrhosis: A review. *Hepat. Med.* **3**, 1–11 (2011).
5. Sohrabpour, A. A., Mohamadnejad, M. & Malekzadeh, R. Review article: The reversibility of cirrhosis. *Aliment. Pharmacol. Ther.* **36**(9), 824–832 (2012).
6. Terris, M. Epidemiology of cirrhosis of the liver: National mortality data. *Am. J. Public Health Nations Health.* **57**(12), 2076–2088 (1967).
7. Marroni, C. A. *et al.* Liver transplantation and alcoholic liver disease: History, controversies, and considerations. *World J. Gastroenterol.* **24**(26), 2785–2805 (2018).
8. Altamirano, J. *et al.* Alcohol abstinence in patients surviving an episode of alcoholic hepatitis: Prediction and impact on long-term survival. *Hepatology* **66**(6), 1842–1853 (2017).
9. Yeom, S. K., Lee, C. H., Cha, S. H. & Park, C. M. Prediction of liver cirrhosis, using diagnostic imaging tools. *World J. Hepatol.* **7**(17), 2069–2079 (2015).
10. Watanabe, A. *et al.* Magnetic resonance imaging of the cirrhotic liver: An update. *World J. Hepatol.* **7**(3), 468–487 (2015).
11. Schwope, R. B., Katz, M., Russell, T., Reiter, M. J. & Lisanti, C. J. The many faces of cirrhosis. *Abdom. Radiol.* **45**(10), 3065–3080 (2020).
12. Okazaki, H. *et al.* Discrimination of alcoholic from virus-induced cirrhosis on MR imaging. *Am. J. Roentgenol.* **175**(6), 1677–1681 (2000).
13. Nowak, S. *et al.* Detection of liver cirrhosis in standard T2-weighted MRI using deep transfer learning. *Eur. Radiol.* https://doi.org/10.1007/s00330-021-07858-1 (2021).
14. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *CVPR 2016*, 770–778 (2016).
15. Huang, G., Liu, Z., van der Maaten, L. & Weinberger, K. Q. Densely connected convolutional networks. In *CVPR 2017*, 4700–4708 (2017).
16. Selvaraju R. R. *et al.* Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *CVPR 2017*, 618–626 (2017).
17. Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
18. Saito, T. & Rehmsmeier, M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS ONE* **10**(3), e0118432 (2015).
19. Mandrekar, J. N. Receiver operating characteristic curve in diagnostic test assessment. *J. Thorac. Oncol.* **5**(9), 1315–1316 (2010).
20. Tan, K. C. The right posterior hepatic notch sign. *Radiology* **248**(1), 317–318 (2008).
21. Elkilany, A. *et al.* A radiomics-based model to classify the etiology of liver cirrhosis using gadoxetic acid-enhanced MRI. *Sci. Rep.* **11**, 10778 (2021).
22. Sakhuja, P. Pathology of alcoholic liver disease, can it be differentiated from nonalcoholic steatohepatitis?. *World J. Gastroenterol.* **20**(44), 16474–16479 (2014).
23. Soresi, M., Giannitrapani, L., Cervello, M., Licata, A. & Montalto, G. Non invasive tools for the diagnosis of liver cirrhosis. *World J. Gastroenterol.* **20**(48), 18131–18150 (2014).
24. Verrill, C., Markham, H., Templeton, A., Carr, N. J. & Sheron, N. Alcohol-related cirrhosis: Early abstinence is a key factor in prognosis, even in the most severe cases. *Addiction* **104**(5), 768–774 (2009).

Author contributions

J.A.L., A.M.S.: study concept; S.N.: Programming of the Machine Learning application; N.M., M.P., J.C.: Workup of clinical cohort; S.N., A.F., A.M.S., C.B., R.S., J.A.L., U.A.: drafting the manuscript; W.B.: data management; A.F., S.N.: statistics; All authors reviewed the manuscript. J.L., S.N., A.M.S., A.F., and U.A. contributed equally.

Funding

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

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-022-12410-2.

Correspondence and requests for materials should be addressed to A.M.S.

3.6 **Nowak S***, Kloth C*, Theis M, Marinova M, Attenberger UI, Sprinkart AM*, Luetkens JA*. Deep learning–based assessment of CT markers of sarcopenia and myosteatosis for outcome assessment in patients with advanced pancreatic cancer after high-intensity focused ultrasound treatment. European Radiology. 2024;34(1):279–286. DOI: 10.1007/s00330-023-09974-6

Objectives

To evaluate the prognostic value of CT-based markers of sarcopenia and myosteatosis in comparison to the Eastern Cooperative Oncology Group (ECOG) score for survival of patients with advanced pancreatic cancer treated with high-intensity focused ultrasound (HIFU).

Materials and methods

For 142 retrospective patients, the skeletal muscle index (SMI), skeletal muscle radiodensity (SMRD), fatty muscle fraction (FMF), and intermuscular fat fraction (IMFF) were determined on superior mesenteric artery level in pre-interventional CT. Each marker was tested for associations with sex, age, body mass index (BMI), and ECOG. The prognostic value of the markers was examined in Kaplan-Meier analyses with the log-rank test and in uni- and multivariable Cox proportional hazards (CPH) models.

Results

The following significant associations were observed: Male patients had higher BMI and SMI. Patients with lower ECOG had lower BMI and SMI. Patients with BMI lower than 21.8 kg/m2 (median) also showed lower SMI and IMFF. Patients younger than 63.3 years (median) were found to have higher SMRD, lower FMF, and lower IMFF. In the Kaplan-Meier analysis, significantly lower survival times were observed in patients with higher ECOG or lower SMI. Increased patient risk was observed for higher ECOG, lower BMI, and lower SMI in univariable CPH analyses for 1-, 2-, and 3-year survival. Multivariable CPH analysis for 1-year survival revealed increased patient risk for higher ECOG, lower SMI, lower IMFF, and higher FMF. In multivariable analysis for 2- and 3-year survival, only ECOG and FMF remained significant.

Conclusion

CT-based markers of sarcopenia and myosteatosis show a prognostic value for assessment of survival in advanced pancreatic cancer patients undergoing HIFU therapy.

Clinical relevance statement

The results indicate a greater role of myosteatosis for additional risk assessment beyond clinical scores, as only FMF was associated with long-term survival in multivariable CPH analyses along ECOG and also showed independence to ECOG in group analysis.

European Radiology (2024) 34:279–286
https://doi.org/10.1007/s00330-023-09974-6

European Radiology
EUROPEAN SOCIETY OF RADIOLOGY

MUSCULOSKELETAL

Deep learning–based assessment of CT markers of sarcopenia and myosteatosis for outcome assessment in patients with advanced pancreatic cancer after high-intensity focused ultrasound treatment

Sebastian Nowak[1] · Christoph Kloth[1] · Maike Theis[1] · Milka Marinova[1,2] · Ulrike I. Attenberger[1] · Alois M. Sprinkart[1] · Julian A. Luetkens[1]

Received: 21 February 2023 / Revised: 21 April 2023 / Accepted: 28 May 2023 / Published online: 12 August 2023

Abstract

Objectives To evaluate the prognostic value of CT-based markers of sarcopenia and myosteatosis in comparison to the Eastern Cooperative Oncology Group (ECOG) score for survival of patients with advanced pancreatic cancer treated with high-intensity focused ultrasound (HIFU).

Materials and methods For 142 retrospective patients, the skeletal muscle index (SMI), skeletal muscle radiodensity (SMRD), fatty muscle fraction (FMF), and intermuscular fat fraction (IMFF) were determined on superior mesenteric artery level in pre-interventional CT. Each marker was tested for associations with sex, age, body mass index (BMI), and ECOG. The prognostic value of the markers was examined in Kaplan-Meier analyses with the log-rank test and in uni- and multivariable Cox proportional hazards (CPH) models.

Results The following significant associations were observed: Male patients had higher BMI and SMI. Patients with lower ECOG had lower BMI and SMI. Patients with BMI lower than 21.8 kg/m^2 (median) also showed lower SMI and IMFF. Patients younger than 63.3 years (median) were found to have higher SMRD, lower FMF, and lower IMFF. In the Kaplan-Meier analysis, significantly lower survival times were observed in patients with higher ECOG or lower SMI. Increased patient risk was observed for higher ECOG, lower BMI, and lower SMI in univariable CPH analyses for 1-, 2-, and 3-year survival. Multivariable CPH analysis for 1-year survival revealed increased patient risk for higher ECOG, lower SMI, lower IMFF, and higher FMF. In multivariable analysis for 2- and 3-year survival, only ECOG and FMF remained significant.

Conclusion CT-based markers of sarcopenia and myosteatosis show a prognostic value for assessment of survival in advanced pancreatic cancer patients undergoing HIFU therapy.

Clinical relevance statement The results indicate a greater role of myosteatosis for additional risk assessment beyond clinical scores, as only FMF was associated with long-term survival in multivariable CPH analyses along ECOG and also showed independence to ECOG in group analysis.

Key Points

- *This study investigates the prognostic value of CT-based markers of sarcopenia and myosteatosis for patients with pancreatic cancer treated with high-intensity focused ultrasound.*
- *Ma kers for sarcopenia and myosteatosis showed a prognostic value besides clinical assessment of the physical status by the Eastern Cooperative Oncology Group score. In contrast to muscle size measurements, the myosteatosis marker fatty muscle fraction demonstrated independence to the clinical score.*
- *The esults indicate that myosteatosis might play a greater role for additional patient risk assessments beyond clinical assessments of physical status.*

Keywords Tomography, X-ray computed · Pancreatic carcinoma · Sarcopenia · Survival analysis

Sebastian Nowak, Christoph Kloth, Alois M. Sprinkart, and Julian A. Luetkens contributed equally.

✉ Sebastian Nowak
Sebastian.Nowak@ukbonn.de

[1] Department of Diagnostic and Interventional Radiology and Quantitative Imaging Lab Bonn (QILaB), University Hospital Bonn, Venusberg-Campus 1, 53127 Bonn, Germany

[2] Department of Nuclear Medicine, University Hospital Bonn, Venusberg-Campus 1, 53127 Bonn, Germany

Abbreviations

BMI	Body mass index
CPH	Cox proportional hazards
ECOG	Eastern Cooperative Oncology Group
FMF	Fatty muscle fraction
HIFU	High-intensity focused ultrasound
IMFF	Intermuscular fat fraction
IQR	Interquartile range
SMI	Skeletal muscle index
SMRD	Skeletal muscle radiodensity
US	Ultrasound

Introduction

Pancreatic cancer is an oncologic disease with a very poor prognosis and an estimated 5-year survival rate of below 10%. Surgical resection can cure pancreatic cancer in early stage; however, the majority of patients are already unresectable at initial diagnosis [1, 2]. Advanced pancreatic cancer is often associated with a very poor quality of life due to cancer pain and a very short life expectancy despite current oncological treatment with chemotherapy or chemoradiotherapy [1]. Local ablation with minimal invasive high-intensity focused ultrasound (HIFU) is an additional treatment option that is often combined with palliative standard treatment, e.g., systemic chemotherapy. With this technique, therapeutic ultrasound (US) waves are focused on the pancreatic lesion to induce coagulative necrosis, leaving healthy tissue outside the focus unharmed. HIFU treatment has been shown to reduce disease-associated symptoms, e.g., cancer pain or tumor mass, and to prolong the survival of patients compared to patients undergoing chemotherapy alone [1, 3, 4].

In patients with pancreatic cancer, a multifactorial syndrome, termed cancer cachexia, is particularly common [5]. This syndrome is induced by reduction of nutritional intake (e.g., due to cancer pain, fatigue, depression, insufficiency of pancreatic enzymes, or side effects of chemotherapy, e.g., nausea and vomiting) and an elevated energy metabolism (e.g., due to increased glucose and protein turnover because of advanced cancer) [2, 6]. Cancer cachexia is associated with ongoing loss of weight, skeletal muscle mass (termed sarcopenia), and physical performance leading to reduced quality of life and life expectancy [2, 5]. An established clinical score for evaluating the physical performance and general physical condition of patients is the Eastern Cooperative Oncology Group (ECOG) performance status that describes how limited the patient is in work activity, self-care, and his walking ability [7]. Evaluating the general physical condition by ECOG showed a strong prognostic value for outcome assessments in pancreatic and other cancer patients [8, 9].

Patients with advanced pancreatic cancer usually receive CT examinations for staging purposes prior to initiation of treatment and during therapy. Besides the diagnostic intention, these CT scans can also be utilized opportunistically to assess the patient's constitution. Here, muscle size measurements are typically used as a surrogate marker for assessment of muscle wasting in sarcopenia. Additionally, muscle radiodensity evaluations are used to measure infiltration of lipids into the intra- and intermyocellular compartments, termed myosteatosis [10–13]. This opportunistic approach is also driven by recent successes of deep learning in medical imaging, which increases the clinical applicability of image-based analysis by automating otherwise time-consuming tissue segmentations [14–16]. To date, the prognostic value of imaging-based assessment of sarcopenia and myosteatosis has been demonstrated for patients with pancreatic cancer receiving chemotherapy or surgical resection, but not for patients undergoing HIFU therapy [2, 10, 17–19]. For instance, a meta-analysis of multiple studies on body composition and sarcopenia observed significant overall effects for sarcopenia evaluations based on muscle size measurements in CT imaging of patients with resectable and unresectable pancreatic cancer [17]. Another study additionally observed a prognostic value of radiodensity measurements of the muscles in CT for the evaluation of myosteatosis in patients with pancreatic cancer treated with palliative chemotherapy [10].

Therefore, the aim of this study was to evaluate the prognostic value of CT-based assessment of sarcopenia and myosteatosis in comparison to clinical assessment of physical status by ECOG for survival prediction in patients with advanced pancreatic cancer undergoing local US-guided HIFU ablation. Additionally, this study aims to investigate associations between the image-based markers, ECOG, and basic clinical parameters.

Material and methods

With the approval of institutional review board of the Medical Faculty of the Rheinische Friedrich-Wilhelms-Universitat Bonn, written informed consent was waived due to the retrospective, single-center nature of the study. The study was carried out in compliance with the ethical standards set in the 1964 Declaration of Helsinki as well as its later amendments. Consecutive patients with advanced pancreatic adenocarcinoma undergoing local US-guided HIFU treatment at our center between May 2014 and April 2020 and available CT within 14 days prior to intervention were included. Sex, age, body mass index (BMI), and ECOG were assessed from the clinical data system. The musculus erector spinae was segmented on an axial CT slice at the level of the superior mesenteric artery by a deep learning method [15]. The automatic segmentation was then manually optimized by a medical resident and finally approved by a board-certified radiologist.

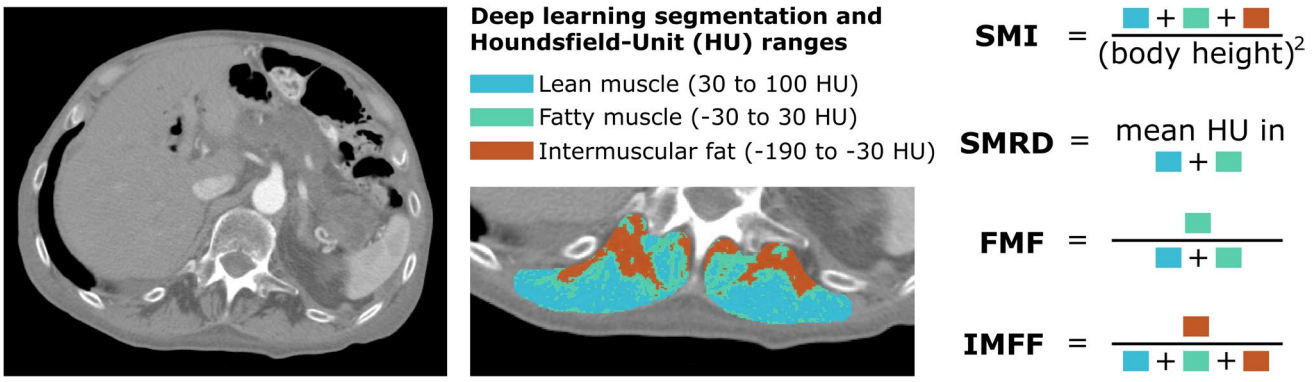


Fig. 1 Overview of image-based markers. On the left, a transversal CT scan slice at the level of superior mesenteric artery is shown. The musculus erector spinae was segmented applying a deep learning model. If necessary, automatic segmentations were manually optimized. The segmented area was subdivided into different tissue classes using different ranges of Hounsfield units (HU). The skeletal muscle index (SMI), skeletal muscle radiodensity (SMRD), fatty muscle fraction (FMF), and intermuscular fat fraction (IMFF) were calculated according to the definitions shown on the right

CT-based markers of sarcopenia and myosteatosis

Figure 1 illustrates the computation of image-based markers in detail. To assess muscle size, the established "skeletal muscle index" (SMI) was calculated from the total area of muscle compartment and body height, as frequently applied in muscle-based body composition analysis [2, 10, 12, 18, 20]. For myosteatosis assessment, the two previously proposed markers "skeletal muscle radiodensity" (SMRD) and "fatty muscle fraction" (FMF) were determined [10, 13]. FMF aims to quantify the extent of intramuscular fat infiltration by relating the area of fatty degenerated muscle to the combined area of lean muscle and fatty degenerated muscle. To additionally assess intermuscular fat not captured by SMRD and FMF, the percentage of pure intermuscular fat tissue was extracted as "intermuscular fat fraction" (IMFF).

Statistical analysis

First, each CT marker was tested for associations with clinical attributes. Therefore, the patients were divided into subgroups for sex, age and BMI (split by the respective median value), ECOG (score 0, 1, and ≥ 2), and the survival status after 1 and 2 years. For each subgroup, median and 25th and 75th interquartile ranges (IQR) are provided. Differences between the subgroups were assessed by the Mann-Whitney *U* test for sex, age, BMI, and survival status, and the Kruskal-Wallis *H* test for ECOG (SciPy 1.8.0) [21]. Differences with p value < 0.05 were considered significant.

Then, differences in survival time between subgroups split by clinical and image-based markers were examined in the Kaplan-Meier analysis with the log-rank test. For this, all continuous parameters, i.e., all parameters except sex and ECOG, were divided into subgroups according to sex-specific median values.

Finally, all clinical attributes and imaging markers were examined in univariable CPH models as well as in multivariable CPH models including all parameters. Kaplan-Meier and CPH analyses were conducted with SPSS (27.0.0, IBM).

Results

Prior to analysis, 153 eligible patients treated with HIFU were identified. Eleven patients were excluded because no CT scan was acquired within 14 days prior to intervention, or due to missing body height or weight records at time of CT imaging. Therefore, a total of 142 patients (73 females, mean age 64.1 ± 10.5 years, range 38–87.5) were included for analysis. Table 1 shows detailed clinical characteristics of the patients included.

Median values and interquartile ranges for age, BMI, SMI, SMRD, FMF, and IMFF split into subgroups by sex, median age, median BMI, survival status after 1 and 2 years, and ECOG, are presented in Table 2. The following significant associations were observed: Male patients showed higher BMI and SMI. Patients with lower ECOG score had higher BMI and higher SMI. Patients with BMI higher than 21.8 kg/m^2 (median) were observed to have higher SMI and higher IMFF. Patients older than 63.3 years (median) showed lower SMRD, higher FMF, and higher IMFF. Patients who survived 1 year had higher BMI and SMI, and patients who survived 2 years had higher BMI compared to patients who died earlier. Figure 2 illustrates violin and boxplots for SMI, BMI, and IMFF split by sex, ECOG score, median age, and median BMI.

Table 3 shows 1-, 2-, and 3-year survival and the results of the Kaplan-Meier analysis with log-rank test for patients' subgroups split by clinical and imaging-based parameters. Overall, the median survival time of all patients was 185 (IQR: 99–404) days. Only 10 patients survived longer than 2 years. Patients with low SMI and patients with higher ECOG score

Table 1 Clinical characteristics of the patients with advanced pancreatic cancer treated by high-intensity focused ultrasound at our center. ECOG:, Eastern Cooperative Oncology Group performance status

	All	Females	Males		All	Females	Males
Site of disease				Biliary drainage			
Body and/or tail	48 (34%)	25 (34%)	23 (33%)	Metallic stent	20 (14%)	11 (16%)	9 (13%)
Head	60 (42%)	28 (38%)	32 (47%)	Plastic stent	14 (10%)	9 (12%)	5 (7%)
Head and body	34 (24%)	20 (28%)	14 (20%)	PTCD	1 (1%)	0 (0%)	1 (1%)
UICC stage				Metastases			
Stage II	2 (1%)	2 (3%)	0 (0%)	Hepatic	65 (46%)	32 (44%)	33 (48%)
Stage III	52 (37%)	26 (36%)	26 (38%)	Pulmonary	12 (8%)	7 (10%)	5 (7%)
Stage IV	83 (58%)	44 (60%)	39 (56%)	Lymph nodes	36 (25%)	18 (25%)	18 (26%)
Recurrence (after Whipple)	5 (4%)	1 (1%)	4 (6%)	Peritoneal	31 (22%)	16 (22%)	15 (22%)
ECOG				Previous treatment			
Status = 0	42 (30%)	15 (21%)	27 (39%)	Chemotherapy	116 (82%)	60 (82%)	56 (81%)
Status = 1	76 (53%)	41 (56%)	35 (51%)	Radiotherapy	9 (6%)	2 (3%)	7 (10%)
Status ≥ 2	24 (17%)	17 (23%)	7 (10%)	Surgery (Whipple)	5 (4%)	1 (1%)	4 (6%)

had lower survival times. Figure 3 shows the corresponding 3-year Kaplan-Meier survival curves. Hazard ratios and *p* values of the univariable and multivariable CPH models are shown in Table 4. Univariable CPH analyses for 1-, 2-, and 3-year survival showed that higher ECOG score, lower BMI, and lower SMI were associated with increased patient risk. Combining all parameters in multivariable CPH analyses for 1-year survival revealed that higher ECOG score, lower SMI, lower IMFF, and higher FMF were associated with increased patient risk. When parameters are examined in multivariable CPH models for 2- and 3-year survival, only ECOG and FMF remained significant.

Discussion

This study investigates the prognostic value of CT imaging markers for sarcopenia and myosteatosis in comparison to clinical assessment of physical status by the ECOG score

Table 2 Median values and interquartile ranges of subgroups split by clinical parameters sex, age, body mass index (BMI), and Eastern Cooperative Oncology Group performance status (ECOG) and survival status after 1 and 2 years. Between the investigated subgroups, associations to age, BMI, survival status, skeletal muscle index (SMI), skeletal muscle radiodensity (SMRD), fatty muscle fraction (FMF), and intermuscular fat fraction (IMFF) were tested using the Mann-Whitney *U* test and for ECOG using the Kruskal-Wallis *H* test. Significant differences with *p* value ≤ 0.05 are indicated in bold

Clinical param.	Subgroup	Age	BMI	SMI	SMRD	FMF	IMFF
Sex	Male	63.9 [55.8–72.5]	22.6 [21.3–24.4]	13.5 [11.7–15.8]	45.1 [40.7–50.2]	17.3 [10.5–22.6]	4.3 [2.2–7.9]
	Female	62.1 [56.8–73.2]	20.5 [19.4–22.7]	11.7 [10.1–12.9]	44.1 [37.6–47.8]	17.1 [13.6–27.1]	5.2 [3.3–8.2]
	p value	0.64	**< 0.01**	**< 0.01**	0.24	0.4	0.08
Age	> 63.	-	22.0 [20.0–24.2]	12.3 [10.8–15.0]	41.0 [35.5–45.2]	21.9 [16.8–32.2]	5.8 [3.9–9.1]
	≤ 63.3	-	21.8 [19.8–23.0]	12.5 [10.8–14.2]	48.4 [43.9–51.7]	13.2 [9.7–18.6]	3.3 [1.9–6.2]
	p value	-	0.43	0.54	**< 0.01**	**< 0.01**	**< 0.01**
BMI	> 21.	64.3 [56.6–73.3]	-	13.5 [12.1–15.7]	43.9 [38.7–49.1]	19.1 [12.5–27.9]	5.3 [3.2–9.0]
	≤ 21.8	62.0 [56.6–71.7]	-	11.2 [9.4–12.9]	46.5 [40.6–49.8]	15.8 [11.8–22.5]	3.9 [2.0–7.2]
	p value	0.63	-	**< 0.01**	0.17	0.14	**0.01**
ECOG	= 0	61.3 [54.6–67.1]	22.6 [20.8–26.1]	13.7 [11.7–16.3]	45.1 [40.6–48.9]	16.5 [11.8–23.9]	4.7 [2.9–8.9]
	= 1	64.9 [56.7–73.4]	21.6 [19.6–23.2]	12.4 [10.7–14.2]	44.9 [39.5–50.0]	17.0 [12.0–26.0]	5.0 [2.9–6.9]
	≥ 2	62.0 [59.2–74.6]	21.1 [19.4–22.9]	11.9 [9.8–12.5]	43.2 [36.2–48.0]	18.2 [12.0–28.6]	4.0 [2.3–8.7]
	p value	0.17	**0.03**	**0.01**	0.47	0.74	0.94
Survival status after 1 year	Died	61.8 [56.6–73.2]	21.5 [19.6–23.4]	12.2 [10.5–13.7]	45.1 [38.8–50.1]	16.8 [11.9–26.7]	4.6 [2.3–7.8]
	Survived	64.3 [55.6–71.6]	22.5 [20.8–25.0]	13.5 [10.9–16.6]	44.1 [39.9–48.8]	18.8 [11.6–25.6]	4.9 [3.0–7.8]
	p value	0.96	**0.02**	**0.01**	0.97	0.95	0.65
Survival status after 2 years	Died	62.1 [56.6–73.2]	21.8 [19.8–23.6]	12.3 [10.7–14.2]	45.1 [38.5–50.0]	17.1 [11.8–27.1]	4.9 [2.4–7.9]
	Survived	61.9 [56.1–65.3]	24.4 [22.3–27.9]	14.3 [12.6–15.6]	44.1 [42.7–47.6]	19.5 [14.3–21.2]	4.7 [4.2–5.1]
	p value	0.47	**0.02**	0.10	0.83	0.77	0.74

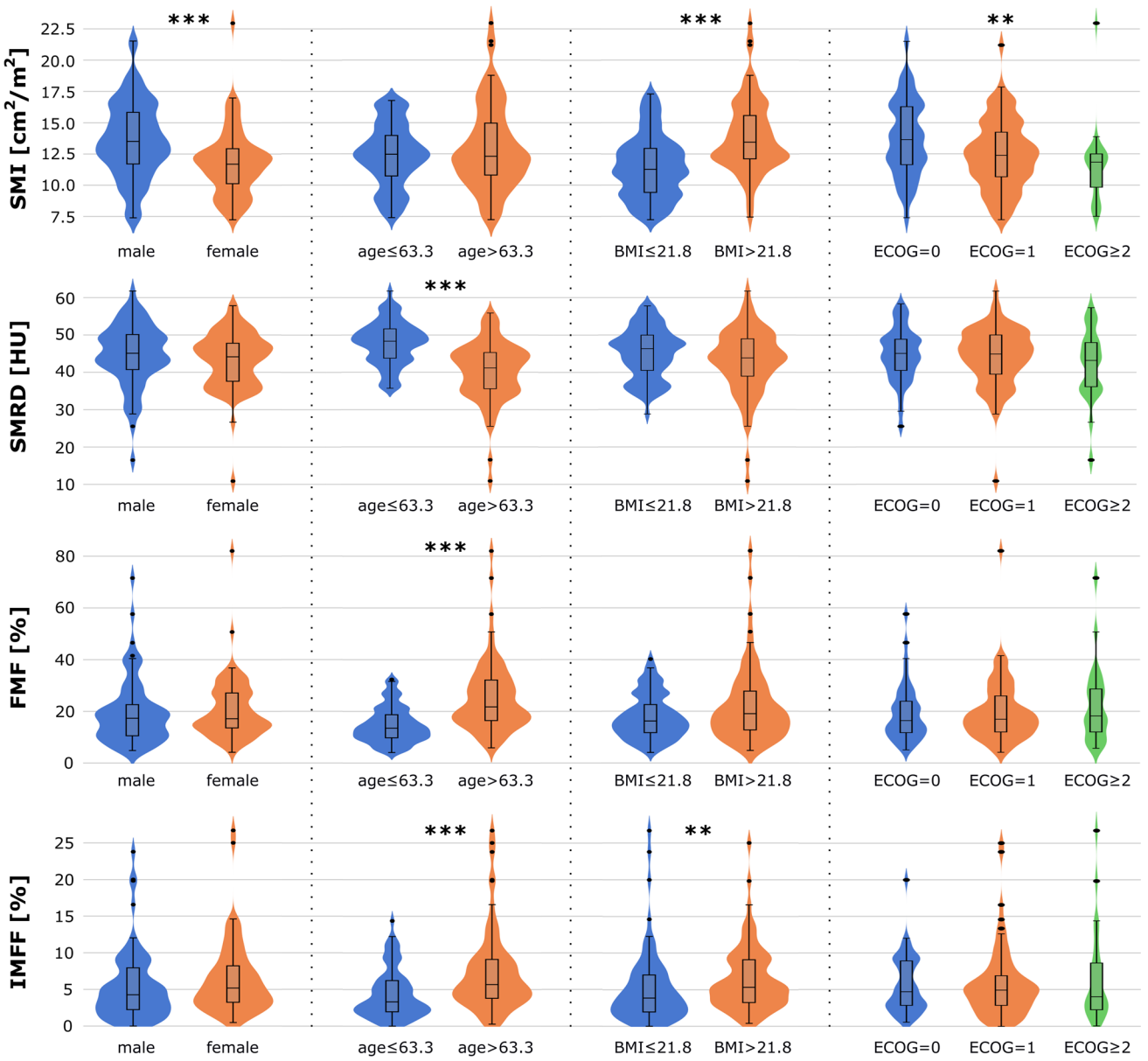


Fig. 2 Violin and boxplots for the muscle size assessing the skeletal muscle index (SMI), as well as for myosteatosis assessing markers skeletal muscle radio density (SMRD), fatty muscle fraction (FMF), and intermuscular fat fraction (IMFF) separated for females and males, separated at median age, separated at median body mass index (BMI), and separated for Eastern Cooperative Oncology Group performance status (ECOG). Differences between sexes, age, and BMI were tested using a two-tailed *t* test and between ECOG with Kruskal-Wallis *H* test, with significance indicated by an asterisk (***p values ≤ 0.001, **p values ≤ 0.01, *p values ≤ 0.05)

in patients with advanced pancreatic cancer treated with local US-guided HIFU ablation in addition to other palliative oncological treatments.

The results demonstrate that the ECOG performance status is a strong predictor of patient survival following HIFU therapy in both, Kaplan-Meier and CPH analyses. Previous studies have already demonstrated that the assessment of the patient's general physical condition by the ECOG score has a strong prognostic value for pancreatic cancer patients treated with chemo- or chemoradiotherapy and for other cancer patients [8, 9]. Interestingly, other previous studies that evaluated ECOG for pancreatic cancer patients treated with chemotherapy and/or surgical resection did not observe such a strong prognostic value [18, 19]. However, only two ECOG groups were considered in these studies (ECOG = 0 and ≥ 1). This indicates that particularly a score of ECOG ≥ 2 may be associated with increased mortality.

We investigated SMI, as this marker already demonstrated a prognostic value for patients with pancreatic cancer treated by surgical resection and for various other oncological

Table 3 Evaluation of predictors of 1-, 2-, and 3-year survival in patients with advanced pancreatic cancer undergoing ultrasound-guided HIFU treatment using the Kaplan-Meier analysis. Differences in survival times were tested by the log-rank test. For each variable, patients were split into subgroups. For age, body mass index (BMI), skeletal muscle index (SMI), skeletal muscle radiodensity (SMRD), fatty muscle fraction (FMF), and intermuscular fat fraction (IMFF), patients were split according to the sex-specific median (SSM). Median survival times for each subgroup are given with 95% confidence interval. p values ≤ 0.05, that indicate significance, are highlighted in bold

			Number of events in…			Median survival	p value of the log-rank test		
Variable	Subgroup	N	1 year	2 years	3 years	Time [days]	1 year	2 years	3 years
Sex	Male	69	40	60	62	222 [122–322]	0.10	0.23	0.21
	Female	73	53	63	65	185 [161–208]			
Age	≤ SSM	72	51	63	67	188 [138–238]	0.47	0.87	0.69
	> SS	70	42	60	60	196 [152–240]			
BMI	≤ SSM	73	51	66	68	171 [138–204]	0.18	0.21	0.17
	> SS	69	42	57	59	241 [120–362]			
ECOG	= 0	42	21	31	33	353 [283–423]	**< 0.0**	**< 0.0**	**< 0.0**
	= 1	76	51	70	72	187 [118–256]			
	≥ 2	24	21	22	22	75 [0–154]			
SMI	≤ SSM	72	50	66	68	161 [116–205]	**0.04**	**0.02**	**0.01**
	> SS	70	43	57	59	265 [174–356]			
FMF	≤ SSM	72	51	64	66	173 [138–207]	0.26	0.22	0.16
	> SS	70	42	59	61	206 [152–260]			
SMRD	≤ SSM	72	43	60	62	213 [161–265]	0.23	0.13	0.09
	> SS	70	50	63	65	175 [140–210]			
IMFF	≤ SSM	72	48	63	64	173 [136–210]	0.49	0.52	0.72
	> SS	70	45	60	63	222 [132–312]			

diseases in previous studies [17, 19, 20, 22, 23]. In the subgroup analysis, SMI showed a significant correlation to BMI. The significant hazard ratios of BMI and SMI in the univariable CPH analysis, their lower prognostic value in the multivariable CPH analyses, and their observed associations with ECOG suggest that part of the information of BMI and SMI is already well reflected by ECOG. However, the significant differences in SMI between patients who survived and patients who died within 1 year, along with the prognostic value of SMI in Kaplan-Meier and in the multivariable CPH analysis for 1-year survival, indicate that image-based assessment of muscle size on CT imaging prior to HIFU treatment provides additional information particularly for short-term survival.

In addition to quantification of muscle size by SMI, which is typically used for image-based assessment of sarcopenia, we also investigated markers assessing myosteatosis. Of note, in contrast to SMI, all myosteatosis markers were associated with age. The increase of myosteatosis with age was also described in previous studies [11]. The markers for myosteatosis showed no associations with the ECOG performance score, sex nor BMI, while SMI and BMI were associated with sex and ECOG.

Mean radiodensity of the musculature is a myosteatosis marker that showed prognostic relevance within univariable CPH analyses in a previous study of patients with unresectable pancreatic cancer treated with palliative chemotherapy [10]. However, for the patient cohort

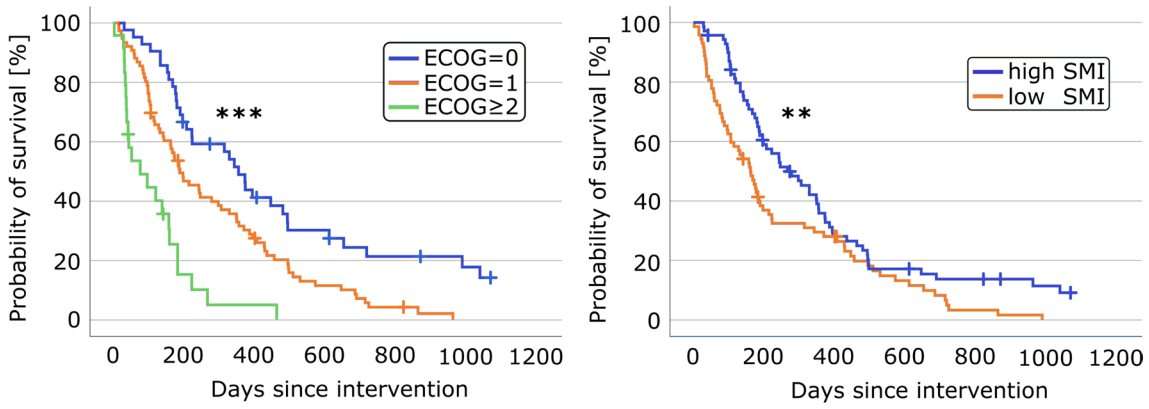


Fig. 3 The Kaplan-Meier curves for the 3-year survival of patients separated by the Eastern Cooperative Oncology Group performance status (ECOG) and by sex-specific median of the skeletal muscle index (SMI). Differences between groups were tested by the log-rank test and significance is indicated by an asterisk (***p values ≤ 0.001, **p values ≤ 0.01)

Table 4 Evaluation of predictors of 1-, 2-, and 3-year mortality in patients with pancreatic cancer undergoing high-intensity focused ultrasound therapy using Cox proportional hazards models. First univariable analysis was performed with the imaging-based markers (skeletal muscle index (SMI), skeletal muscle radiodensity (SMRD), fatty muscle fraction (FMF), and intermuscular fat fraction (IMFF)) and clinical attributes (sex, age, body mass index (BMI), Eastern Cooperative Oncology Group performance status (ECOG)). Then, multivariable models with inclusion of all parameters were tested. Significant hazard ratios with p values ≤ 0.05 are shown in bold

	1-year survival				2-year survival				3-year survival			
	Univariable analysis		Multivariable analysis		Univariable analysis		Multivariable analysis		Univariable analysis		Multivariable analysis	
Variables	Hazard ratio	p	Hazard ratio	p	Hazard ratio	p	Hazard ratio	p	Hazard ratio	p	Hazard ratio	p
Sex	0.71 [0.47–1.08]	0.11	1.05 [0.66–1.67]	0.85	0.80 [0.56–1.15]	0.23	1.11 [0.75–1.65]	0.61	0.80 [0.56–1.14]	0.22	1.13 [0.76–1.66]	0.55
Age	1.00 [0.98–1.02]	0.88	0.99 [0.97–1.01]	0.37	1.00 [0.99–1.02]	0.73	0.99 [0.97–1.01]	0.46	1.00 [0.98–1.02]	0.88	0.99 [0.97–1.01]	0.34
BMI	0.93 [0.87–0.99]	**0.02**	0.99 [0.92–1.06]	0.70	0.94 [0.89–0.99]	**0.02**	0.97 [0.92–1.03]	0.36	0.94 [0.89–0.99]	**0.01**	0.97 [0.92–1.03]	0.35
ECOG	2.18 [1.57–3.04]	**< 0.01**	2.23 [1.54–3.24]	**< 0.01**	2.15 [1.60–2.90]	**< 0.01**	2.29 [1.64–3.19]	**< 0.01**	2.22 [1.65–2.98]	**< 0.01**	2.36 [1.70–3.28]	**< 0.01**
SMI	0.88 [0.82–0.95]	**< 0.01**	0.90 [0.82–0.98]	**0.02**	0.92 [0.86–0.98]	**0.01**	0.94 [0.87–1.01]	0.09	0.91 [0.86–0.97]	**< 0.01**	0.94 [0.87–1.01]	0.07
SMRD	0.99 [0.97–1.02]	0.67	1.06 [0.97–1.15]	0.24	1.00 [0.97–1.02]	0.68	1.06 [0.98–1.15]	0.13	1.00 [0.97–1.02]	0.76	1.06 [0.98–1.15]	0.14
FMF	1.00 [0.99–1.02]	0.67	1.07 [1.01–1.14]	**0.03**	1.00 [0.99–1.02]	0.56	1.07 [1.01–1.13]	**0.02**	1.00 [0.99–1.02]	0.66	1.07 [1.01–1.13]	**0.02**
IMFF	0.99 [0.94–1.03]	0.57	0.94 [0.89–1.00]	**0.05**	0.99 [0.95–1.03]	0.72	0.96 [0.91–1.01]	0.09	0.99 [0.95–1.03]	0.76	0.96 [0.91–1.01]	0.13

of our study, no prognostic value of SMRD was observed in Kaplan-Meier nor CPH analyses.

Besides SMI and SMRD, we also investigated the two markers FMF and IMFF explicitly aimed at assessing inter- and intramuscular fat infiltration in myosteatosis. These two markers did not show any prognostic value when considered in Kaplan-Meier or univariable CPH analysis alone. However, both markers were significant predictors along with SMI when combined with clinical parameters in multivariable CPH analysis for 1-year survival. Furthermore, FMF was the only image-based marker that retained predictive value along with ECOG in the multivariable Cox models for 2- and 3-year survival.

Interestingly, IMFF was observed as a protective predictor with hazard ratios below one, in contrast to FMF, for which patient risk increases with higher values. Due to the observed association of the protective predictor IMFF with BMI, it may be assumed that larger intermuscular fat depots represented a better nutritional status that prolongs short-term survival in the current cohort.

As described in other studies, the results of our study also underscore that sarcopenia and myosteatosis are not synonymous and that assessment of myosteatosis has the potential to provide important additional information [11].

Conclusion

In conclusion, this study demonstrates that image-based markers of sarcopenia and myosteatosis derived from pre-therapeutic CT scans have a prognostic value for patients with advanced pancreatic cancer after palliative HIFU therapy. Image-based assessment of myosteatosis might play a greater role in the evaluation of a patient's physical status along with the established ECOG score than simple muscle size measurements.

Funding Open Access funding enabled and organized by Projekt DEAL. 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 01KX2021. The funders had no influence on the conceptualization and design of the study, data analysis and data collection, preparation of the manuscript, and the decision to publish.

Declarations

Guarantor The scientific guarantor of this publication is PD Dr. med. Julian Luetkens.

Conflict of interest The authors of this manuscript declare no relationships with any companies whose products or services may be related to the subject matter of the article.

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

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

Ethical approval Due to the single-center and retrospective nature of the study, it was approved by the institutional review board of the Medical Faculty of the Rheinische Friedrich-Wilhelms-Universitat Bonn with waiver of written informed consent.

Study subjects or cohorts overlap Some study subjects were included in a previous study. The aim of this previous study was to evaluate if a HIFU-induced early sterile inflammatory reaction is initiated after ablation of uterine fibroids and pancreatic carcinoma. The research question of the current submitted study has no overlap to this previous investigation (DOI: 10.1080/02656736.2021.1900926).

Methodology
- retrospective
- prognostic study
- performed at one institution

References

1. Sofuni A, Asai Y, Mukai S, Yamamoto K, Itoi T (2022) High-intensity focused ultrasound therapy for pancreatic cancer. J Med Ultrason. https://doi.org/10.1007/s10396-022-01208-4
2. Naumann P, Eberlein J, Farnia B et al (2019) Cachectic body composition and inflammatory markers portend a poor prognosis in patients with locally advanced pancreatic cancer treated with chemoradiation. Cancers 11(11):1655
3. 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(05):625–637
4. Marinova M, Feradova H, Gonzalez-Carmona MA et al (2021) Improving quality of life in pancreatic cancer patients following high-intensity focused ultrasound (HIFU) in two European centers. Eur Radiol 31(8):5818–5829
5. Dhanapal R, Saraswathi TR, Govind RN (2011) Cancer cachexia. J Oral Maxillofac Pathol 15(3):257–260
6. Fearon KC, Moses AG (2002) Cancer cachexia. Int J Cardiol 85(1):73–81
7. Oken MM, Creech RH, Tormey DC et al (1982) Toxicity and response criteria of the Eastern Cooperative Oncology Group. Am J Clin Oncol 5(6):649–656
8. Kalser MH, Barkin J, Macintyre JM (1985) Pancreatic cancer. Assessment of prognosis by clinical presentation. Cancer 56(2):397–402
9. Demirelli B, Babacan NA, Ercelep Ö et al (2021) Modified Glasgow prognostic score, prognostic nutritional index and ECOG performance score predicts survival better than sarcopenia, cachexia and some inflammatory indices in metastatic gastric cancer. Nutr Cancer 73(2):230–238
10. Rollins KE, Tewari N, Ackner A et al (2016) The impact of sarcopenia and myosteatosis on outcomes of unresectable pancreatic cancer or distal cholangiocarcinoma. Clin Nutr 35(5):1103–1109
11. Correa-de-Araujo R, Addison O, Miljkovic I et al (2020) Myosteatosis in the context of skeletal muscle function deficit: an interdisciplinary workshop at the National Institute on Aging. Front Physiol 11:963
12. Murray TE, Williams D, Lee MJ (2017) Osteoporosis, obesity, and sarcopenia on abdominal CT: a review of epidemiology, diagnostic criteria, and management strategies for the reporting radiologist. Abdom Radiol (NY) 42(9):2376–2238
13. 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(3):234–236
14. Magudia K, Bridge CP, Bay CP et al (2021) 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(2):319–329
15. 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(6):357–366
16. Nowak S, Theis M, Wichtmann BD et al (2022) End-to-end automated body composition analyses with integrated quality control for opportunistic assessment of sarcopenia in CT. Eur Radiol 32(5):3142–3151
17. Bundred J, Kamarajah SK, Roberts KJ (2019) Body composition assessment and sarcopenia in patients with pancreatic cancer: a systematic review and meta-analysis. HPB (Oxford) 21(12):1603–1612
18. Basile D, Parnofiello A, Vitale MG et al (2019) The IMPACT study: early loss of skeletal muscle mass in advanced pancreatic cancer patients. J Cachexia Sarcopenia Muscle 10(2):368–377
19. Sugimoto M, Farnell MB, Nagorney DM et al (2018) Decreased skeletal muscle volume is a predictive factor for poorer survival in patients undergoing surgical resection for pancreatic ductal adenocarcinoma. J Gastrointest Surg 22(5):831–839
20. Faron A, Opheys NS, Nowak S et al (2021) Deep learning-based body composition analysis predicts outcome in melanoma patients treated with immune checkpoint inhibitors. Diagnostics 11(12):2314
21. Virtanen P, Gommers R, Oliphant TE et al (2020) SciPy 10: fundamental algorithms for scientific computing in Python. Nat Methods 17(3):261–272
22. Prado CM, 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(7):629–635
23. Shachar SS, Williams GR, Muss HB, Nishijima TF (2016) Prognostic value of sarcopenia in adults with solid tumours: a meta-analysis and systematic review. Eur J Cancer 57:58–67

3.7 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. European Journal of Radiology. 2023;168:111150. DOI: 10.1016/j.ejrad.2023.111150

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 ± 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.

European Journal of Radiology 168 (2023) 111150

Contents lists available at ScienceDirect

European Journal of Radiology

journal homepage: www.elsevier.com/locate/ejrad

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

Maike Theis [a,*], Wolfgang Block [a,b,c], Julian A. Luetkens [a], Ulrike I. Attenberger [a], Sebastian Nowak [a,1], Alois M. Sprinkart [a,1]

[a] Department of Diagnostic and Interventional Radiology, Quantitative Imaging Lab Bonn (QILaB), University Hospital Bonn, Venusberg-Campus 1, 53127 Bonn, Germany
[b] Department of Radiotherapy and Radiation Oncology, University Hospital Bonn, Venusberg-Campus 1, 53127 Bonn, Germany
[c] 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 ± 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 (M. Theis), Wolfgang.Block@ukbonn.de (W. Block), Julian.Luetkens@ukbonn.de (J.A. Luetkens), Ulrike.Attenberger@ukbonn.de (U.I. Attenberger), Sebastian.Nowak@ukbonn.de (S. Nowak), sprinkart@uni-bonn.de (A.M. Sprinkart).

[1] 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

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 ± 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 makers 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 ± Std	**Range**
Age (years)	81.21 ± 6.05	[57.00, 96.00]
FMF (%)	62.51 ± 20.10	[9.97, 97.22]
Mean SMRD (HU)	18.98 ± 10.80	[-11.27, 49.42]
SMA (cm^2)	55.89 ± 10.91	[29.37, 107.03]

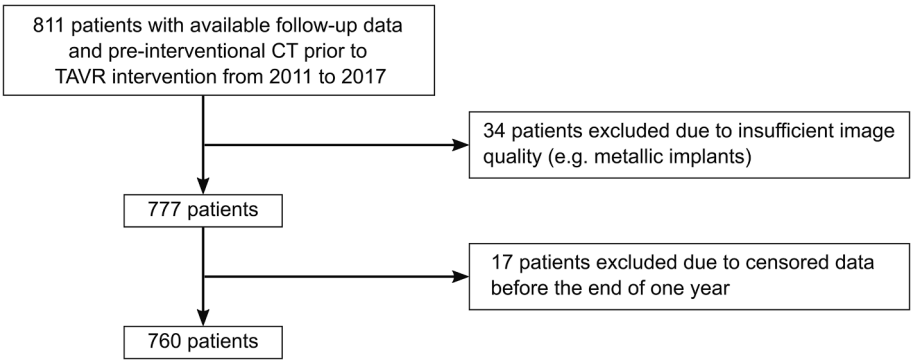


Fig. 1. Flow chart to illustrate the inclusion and exclusion criteria.

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) ($CPHM_{Multivar,sign}$). Lastly, a second multivariable CPHM including all predictors (FMF, SMRD, SMA, sex, age) was investigated, independent from its significance in univariable analysis ($CPHM_{Multivar,all}$).

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 ($DL_{Scratch}$). We refer to the survival prediction CNN with and without focus on the paraspinal musculature in pre-training as DL_{Masked} and $DL_{Unmasked}$. For training of DL_{Masked} and $DL_{Unmasked}$, the weights of the pre-trained encoder are kept frozen ($DL_{Masked,frozen}$ and $DL_{Unmasked,frozen}$) [21–23]. The best approach of $DL_{Masked,frozen}$, $DL_{Unmasked,frozen}$ and $DL_{Scratch}$ 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 ($DL_{Unfrozen}$).

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 ($CPHM_{DL+Sex+Age}$) [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 ($CPHM_{EuroSCORE+Sex+Age}$) and EuroSCORE II ($CPHM_{EuroSCOREII+Sex+Age}$) 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 ($CPHM_{Multivar,sign}$). SMA and sex showed significant hazard ratios in the CPHM including all investigated variables ($CPHM_{Multivar,all}$). 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 $CPHM_{Multivar,sign}$ to the hold-out test set. For $CPHM_{Multivar,all}$ 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 ($DL_{Scratch}$, $DL_{Masked,frozen}$, $DL_{Unmasked,frozen}$) are listed in Table 3. The $DL_{Masked,frozen}$ 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 $DL_{Masked,frozen}$ 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 ($DL_{Unfrozen}$) (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 $DL_{Unfrozen}$ model compared to the $CPHM_{Multivar,sign}$, which only includes the three scalar

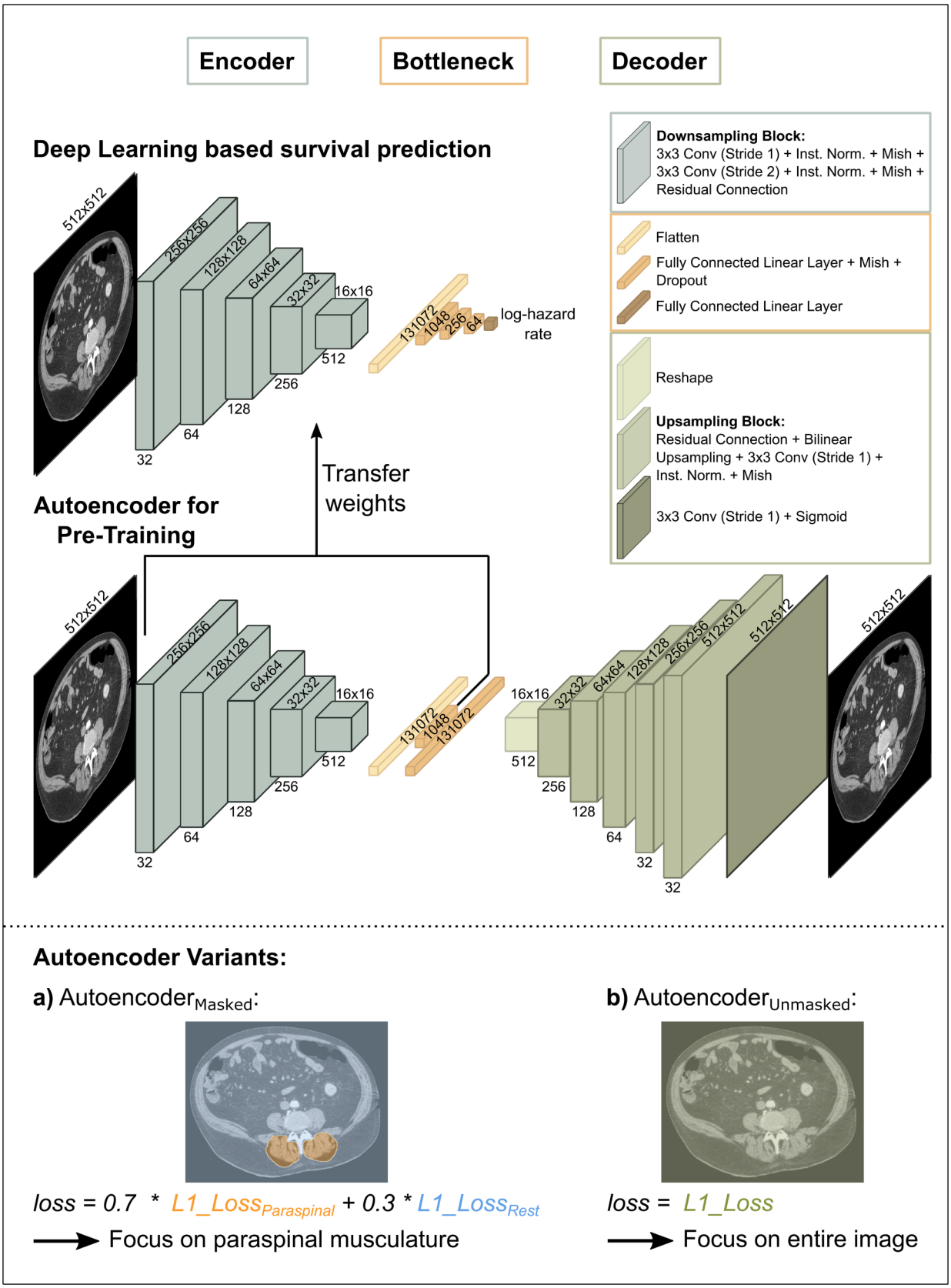


Fig. 2. Overview of the investigated pre-training strategies for the development of an image-based survival prediction. Two autoencoders were trained, one with and one without focusing on paraspinal musculature. Pre-trained weights were afterwards transferred to the deep learning model that predicts 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 ($CPHM_{Multivar,sign}$) and another including all variables ($CPHM_{Multivar,all}$).

	Univariable analysis		Multivariable analysis			
			$CPHM_{Multivar,sign}$		$CPHM_{Multivar,all}$	
Variables	HR	p-value	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^2)	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 ($DL_{Scratch}$), pre-trained on the masked autoencoder ($DL_{Masked,frozen}$) and pre-trained on the standard autoencoder ($DL_{Unmasked,frozen}$). 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
$DL_{Scratch}$	4.09	31	0.609
$DL_{Masked,frozen}$	**4.05**	**43**	**0.636**
$DL_{Unmasked,frozen}$	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**
$DL_{Masked,frozen}$	0.637 [0.570, 0.701]	0.687 [0.567, 0.792]	0.683 [0.583, 0.773]
$DL_{Unfrozen}$	**0.645 [0.580, 0.706]**	**0.687 [0.564, 0.792]**	**0.692 [0.594, 0.777]**
$CPHM_{Multivar,sign}$	0.508 [0.439, 0.578]	0.496 [0.389, 0.614]	0.457 [0.349, 0.567]
$CPHM_{Multivar,all}$	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 $CPHM_{Multivar,all}$, which additionally included sex and age (see Table 4). Fig. 3 shows the Kaplan-Meier analysis of the log-hazard rate predicted by the $DL_{Unfrozen}$ 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 $DL_{Unfrozen}$ in a CPHM together with age and sex ($CPHM_{DL+Sex+Age}$) 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.

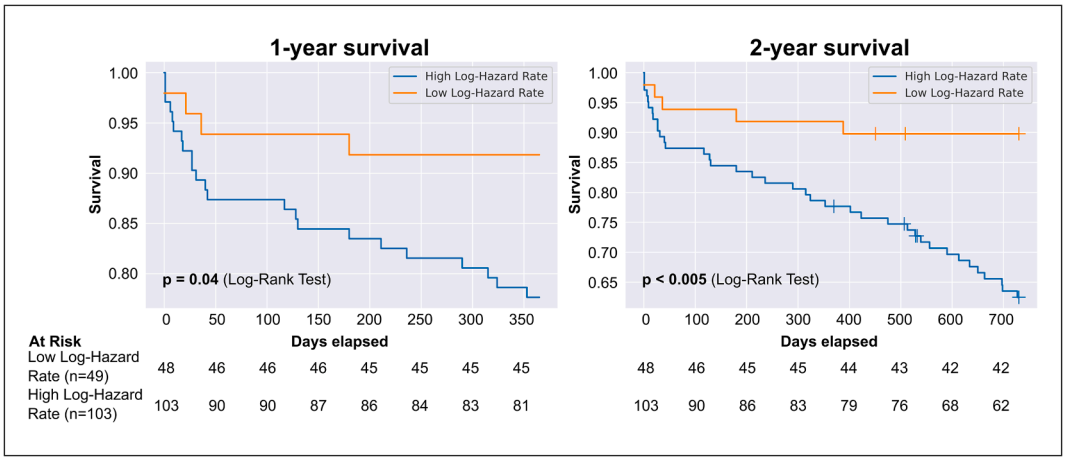


Fig. 3. Kaplan-Meier curves for 1- and 2-year survival. The figure illustrates Kaplan-Meier curves for patients in the hold-out test group (n = 152) stratified by low (orange) and high (blue) predicted log-hazard rate from the $DL_{Unfrozen}$ model, whose weights were unfrozen after previous training 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 $DL_{Unfrozen}$ model together with 95%-confidence intervals in brackets.

Combination of CPHM and $DL_{Unfrozen}$			
	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 TotalSegmentator 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

Maike 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 ≥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. Lefèvre, N. Piazza, K. Lachapelle, G. Martucci, A. Lamy, M. Labinaz, M.D. Peterson, R.C. Arora, N. Noiseux, A. Rassi, I.F. Palacios, P. Généreux, 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 Gonçalves, 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. Endler, 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. Geißler, 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. Geißler, 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. Kalinauskaite, I. Tinhofer, N. Guberina, M. Guberina, P. Balermpas, J. von der Grün, 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. Löck, 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. Endler, 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, 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. Bauckhage, 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.
[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/japplphysiol.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.

3.8 Ginzburg D*, **Nowak S***, Attenberger U, Luetkens J, Sprinkart AM*, Kuetting D*. Computer tomography-based assessment of perivascular adipose tissue in patients with abdominal aortic aneurysms. Scientific Reports 2024;14(1):20512. DOI: 10.1038/s41598-024-71283-9

Objectives

This retrospective study investigates perivascular adipose tissue (PVAT) alterations in CT as a marker of inflammation in patients with abdominal aortic aneurysms (AAA).

Materials and methods

100 abdominal CT scans of patients with abdominal aortic aneurysms and 100 age and sex matched controls without underlying aortic disease were included. Artificial Intelligence (AI) assisted segmentation of the aorta and the surrounding adipose tissue was performed. Adipose tissue density was measured in Hounsfield units (HU) close (2-5mm, HU_{close}) and distant (10-12mm, $HU_{distant}$) to the aortic wall. To investigate alterations in adipose tissue density close to the aorta (HU_{close}) as a potential marker of inflammation, we calculated the difference $HU\Delta = HU_{close}-HU_{distant}$ and the fat attenuation ratio $HU_{ratio} = HU_{close}/ HU_{distant}$ as normalized attenuation measures. These two markers were compared i) inter-individually between AAA patients and controls and ii) intra-individually between the aneurysmal and non-aneurysmal segments in AAA patients. Since most AAAs are generally observed infrarenal, the aneurysmal section of the AAA patients was compared with the infrarenal section of the aorta of the control patients.

Results

In inter-individual comparisons, higher HUΔ and a lower HU_{ratio} were observed (aneurysmal: 8.9 ± 5.1 HU vs. control: 6.9 ± 4.8 HU, p-value = 0.006; aneurysmal: 89.8 ± 5.7% vs. control: 92.1 ± 5.5% p-value = 0.004). In intra-individual comparisons, higher HUΔ and lower HU_{ratio} were observed (aneurysmal: 8.9 ± 5.1 HU vs. non-aneurysmal: 5.5 ± 4.1 HU, p-value < 0.001; aneurysmal: 89.8 ± 5.7% vs. non-aneurysmal 93.3 ± 4.9%, p-value < 0.001).

Conclusion

The results indicate PVAT density alterations in AAA patients. This motivates further research to establish non-invasive imaging markers for vascular and perivascular inflammation in AAA.

scientific reports

Check for updates

OPEN

Computer tomography-based assessment of perivascular adipose tissue in patients with abdominal aortic aneurysms

Daniel Ginzburg[1,2], Sebastian Nowak[1,2], Ulrike Attenberger[1], Julian Luetkens[1], Alois Martin Sprinkart[1,2] & Daniel Kuetting[1,2]✉

This retrospective study investigates perivascular adipose tissue (PVAT) alterations in CT as a marker of inflammation in patients with abdominal aortic aneurysms (AAA). 100 abdominal CT scans of patients with abdominal aortic aneurysms and 100 age and sex matched controls without underlying aortic disease were included. Artificial Intelligence (AI) assisted segmentation of the aorta and the surrounding adipose tissue was performed. Adipose tissue density was measured in Hounsfield units (HU) close (2-5mm, HU_{close}) and distant (10-12mm, $HU_{distant}$) to the aortic wall. To investigate alterations in adipose tissue density close to the aorta (HU_{close}) as a potential marker of inflammation, we calculated the difference $HU_{\Delta} = HU_{close}$-$HU_{distant}$ and the fat attenuation ratio $HU_{ratio} = HU_{close}/HU_{distant}$ as normalized attenuation measures. These two markers were compared i) inter-individually between AAA patients and controls and ii) intra-individually between the aneurysmal and non-aneurysmal segments in AAA patients. Since most AAAs are generally observed infrarenal, the aneurysmal section of the AAA patients was compared with the infrarenal section of the aorta of the control patients. In inter-individual comparisons, higher HU_{Δ} and a lower HU_{ratio} were observed (aneurysmal: 8.9 ± 5.1 HU vs. control: 6.9 ± 4.8 HU, p-value = 0.006; aneurysmal: 89.8 ± 5.7% vs. control: 92.1 ± 5.5% p-value = 0.004). In intra-individual comparisons, higher HU_{Δ} and lower HU_{ratio} were observed (aneurysmal: 8.9 ± 5.1 HU vs. non-aneurysmal: 5.5 ± 4.1 HU, p-value < 0.001; aneurysmal: 89.8 ± 5.7% vs. non-aneurysmal 93.3 ± 4.9%, p-value < 0.001). The results indicate PVAT density alterations in AAA patients. This motivates further research to establish non-invasive imaging markers for vascular and perivascular inflammation in AAA.

Prevalence of abdominal aortic aneurysms (AAA) increases with age, affecting up to 5% of individuals older than 50 years[1]. While the condition is usually asymptomatic until rupture, AAA can be deadly, with a mortality rate of up to 85% in such cases[2]. Established risk factors for AAA development and progression include smoking, diabetes mellitus, aneurysmal size, and morphology[3]. However, understanding of aneurysmal progression is still limited. Although newer assessments such as biomechanical analyses, functional and molecular imaging, and assessment of circulating biomarkers are promising, they are unlikely to be adopted in practice in the near future[4]. Currently, AAA diameter and growth rate remain the only routinely employed markers of AAA risk estimation, despite their limitations. As a result, all AAA patients require regular follow-ups, although up to 50% of smaller AAA below 4 cm diameter remain stable[3]. Therefore, further non-invasive imaging markers are necessary to allow for discrimination between slow-growing or even stable and progressive AAA. As progression of atherosclerosis is directly associated with vascular wall and perivascular inflammation[5], the perivascular adipose tissue (PVAT) has been identified as a region of interest for monitoring vascular wall inflammation and atherosclerosis progression[6]. As inflammation also plays a critical role in the progression of AAA, inflammatory processes can be detected in PVAT during AAA development[6–9]. CT-based analysis of PVAT allows for deduction of imaging biomarkers for non-invasive evaluation of perivascular inflammation. These measurements are based on the assumption that inflamed adipose tissue shows higher density values in CT according to the Hounsfield (HU) scale. This assumption is supported by recent histological studies observing alterations of PVAT induced

[1]Department of Diagnostic and Interventional Radiology, University Hospital Bonn, Venusberg-Campus 1, 53127 Bonn, Germany. [2]These authors contributed equally: Daniel Ginzburg, Sebastian Nowak, Alois Martin Sprinkart and Daniel Kuetting. ✉email: Daniel.Kuetting@ukbonn.de

by vascular injury and in turn its pivotal role in regulating vascular remodeling[10]. Thus, a greater difference in density can be measured between adipose tissue adjacent to the inflamed vascular wall and adipose tissue distant to the vessel[11]. Although encouraging results have been reported for PVAT analysis in coronary artery disease[12], there is currently limited data analyzing PVAT density in patients with AAA. Furthermore, manual assessment of PVAT is time-consuming and lacks reproducibility, which hampers broader application while AI-assisted evaluation of PVAT may facilitate broader application.

The aim of this study was to investigate potential perivascular inflammatory alterations in patients with AAA as a basis for non-invasive imaging biomarkers and to demonstrate that the analyses method can be automated by AI to facilitate further research in this area.

Material and methods

The study was performed in accordance with the relevant guidelines and regulations including the Declaration of Helsinki. The study specific approval for retrospective studies has been exempted as per the IRB approval number 303/16 (Ethics Committee, University Hospital Bonn). The need for a written informed consent was waived.

Patient selection and characteristics

For the AAA group, 100 patients receiving CT angiographies for AAA evaluation during a time interval of 4 years (04/2016—04/2020) were retrospectively included in the study. Patients were selected consecutively in accordance to the prevalence of AAA in the general population regarding gender and age[13].

Exclusion criteria for the AAA group were defined as following:

- Aortic dissection/ruptured aortic aneurysms or contained rupture
- Surgical or interventional treatment of AAA
- Patients with artefacts due to movement during the scan or beam hardening artefacts (e.g. caused by lumbal spondylodesis or intraabominal vascular coils).
- Underlying connective tissue diseases associated with AAA development
- Additional non-aortic aneurysmal diseases.

Details regarding the patient selection are listed in the following Fig. 1.

To compile an age- and gender-matched control group, patients without underlying advanced vascular disease who underwent CT scans in the same period as the included AAA patients were considered. Of these, 100 individuals were selected who had only minimal non-calcified plaques and atherosclerotic changes and who had undergone CT scans as part of long-term follow-up for malignancy. Patients with abdominal scan artefacts and with conditions affecting the abdominal aorta or the associated retroperitoneal perivascular fat tissue such as M. Ormond or Takayasu arteritis, were excluded from the control group. Additionally, patients who had received immuno-, chemo-, or radiotherapy were also excluded from the control group due to the possibility of associated inflammatory vascular/perivascular alterations.

Image acquisition and aortic assessment

Contrast enhanced CT scans were performed on a Somatom Force dual source CT (Siemens Healthineers). Reconstructions of the data set were reformatted in sagittal and coronal planes with a slice thickness of 1 mm with a regular vascular kernel (BV40) and iterative reconstruction (SAFIRE, level 3). Image selection and interpretation was performed with a clinical PACS system viewer (Deep Unity, Dedalus).

The maximal abdominal aortic diameter was evaluated in both the AAA as well as the control group. Aneurysm shapes were subdivided into either saccular or fusiform. The extent of luminal thrombosis (0 cm; < 0,5 cm; 0,5 –1 cm; 1–2 cm; > 2 cm) and the degree of circumferential vasosclerotic changes of the aortic wall (0%; < 25%; 25–50%; 50–75%; 75–100%) were graded in 5 categories (grade 0–4) in AAA patients and controls. For patients with AAA the volume and length of the aneurysmal aortic section was determined.

Image annotations

In a first step, the aorta was segmented from the right coronary artery to the aortic bifurcation. To improve time efficiency of the annotation process, the segmentations were performed iteratively with assistance of artificial intelligence (AI) as described in Supplement S1. The segmented aorta was divided into different sections perpendicular to the central line of the vessel based on several anatomic landmarks placed in in 3D Slicer v4.11.2[14]. For sectioning the aorta of patients with AAA the aortic bifurcation, coeliac trunc, and the inferior and superior ends of the AAA were marked. In control patients, the aortic bifurcation and the junction of the right renal artery were marked to define the infrarenal section of the aorta as seen in Fig. 2.

As most AAAs are observed infrarenally, the aneurysmal section of the AAA patients was compared with the infrarenal section of the aorta of the control patients. To identify voxels surrounding the aorta at different distances from the vessel, the aortic segmentation was iteratively morphologically dilated. An established density-based threshold of −190 HU to −30 HU was subsequently applied for identification of adipose tissue[15]. Detailed information on the landmark based partitioning of the aorta and the extraction of PVAT density is provided in Supplement S2.

Investigation of the perivascular adipose tissue

Attenuation of the perivascular adipose tissue (PVAT) was measured in HU at two different distances from the aortic wall, namely 2 to 5 mm (referred to as HU_{close}) and 10 to 12 mm (referred to as $HU_{distant}$). The area that

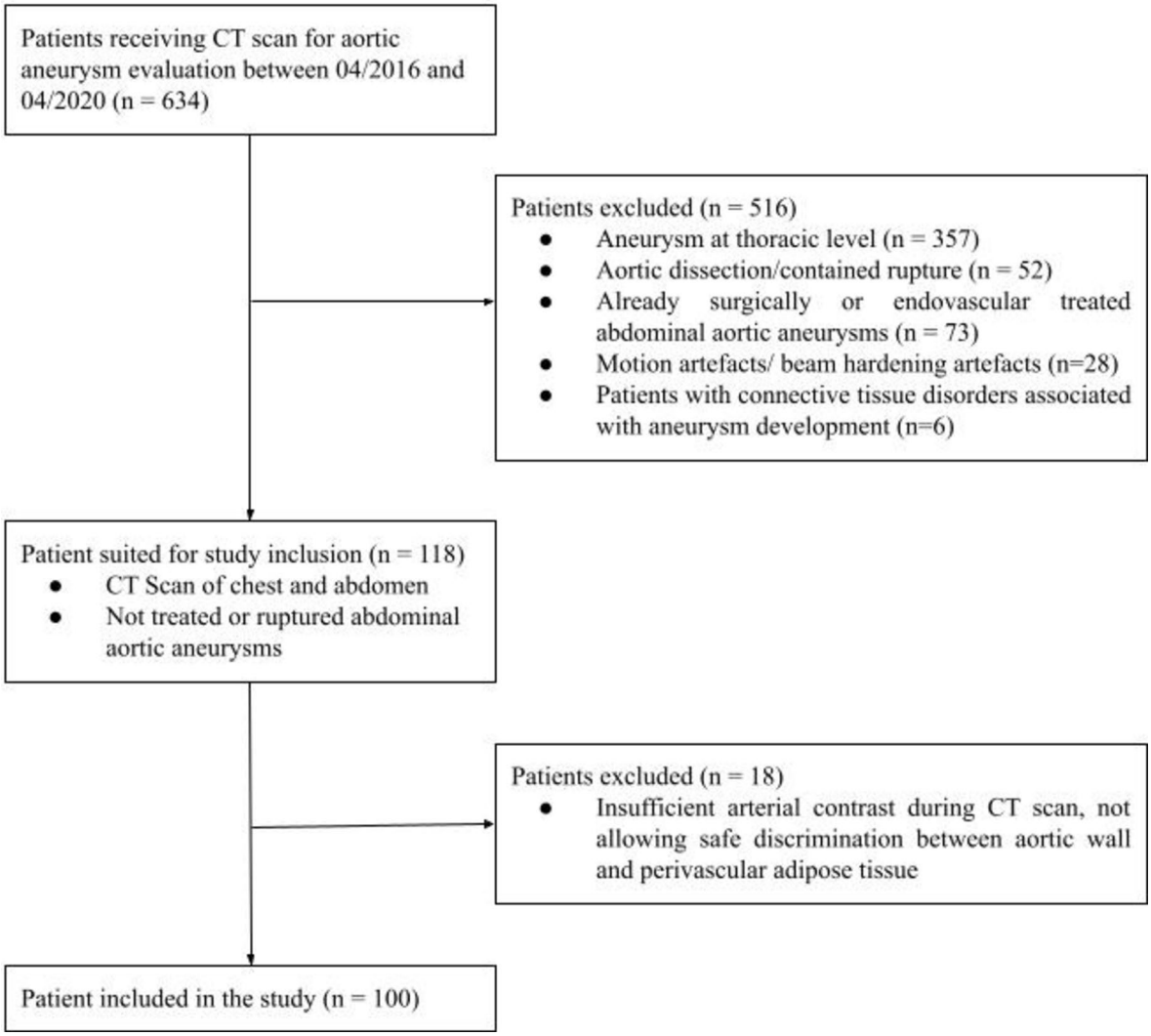


Fig.1. General study outline.

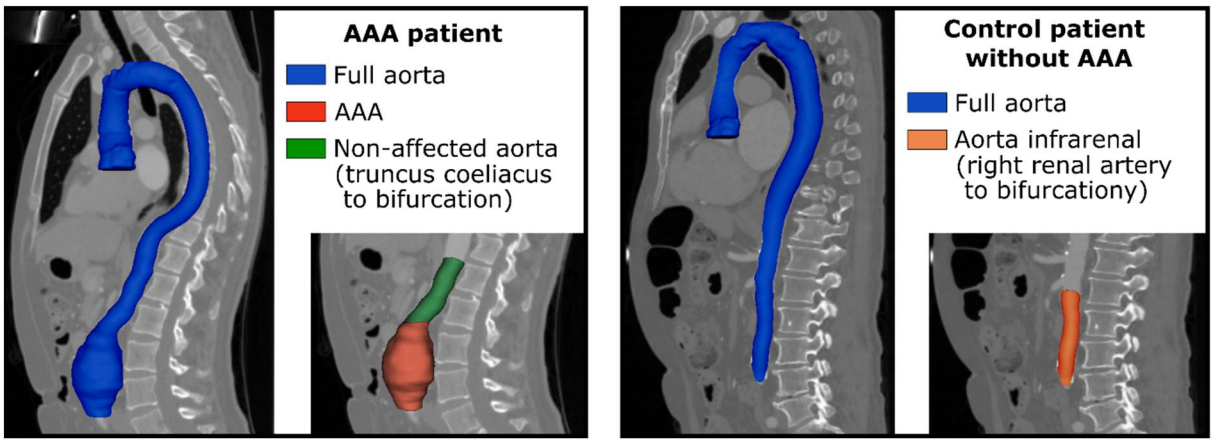


Fig.2. Overview of segmentation and partitioning of the aorta. The aorta was segmented from the right coronary artery to the aortic bifurcation (blue) in all patients. In patients with abdominal aortic aneurysm (AAA), the abdominal aorta was sectioned into the aneurysmal section of the aorta (red) and the non-aneurysmal section of the abdominal aorta between the coeliac trunc and the aortic bifurcation (green). In control patients, the infrarenal section of the aorta was additionally identified (orange).

was closest to the aortic wall including the adventitia, ranging from 0 to 2 mm, was deliberately excluded from the analyses to minimize the potential impact of partial volume effects. The reason for choosing 10–12 mm for $HU_{distant}$ was that fatty tissue should be included at a sufficient distance from the wall to measure an adequate effect, but not too far away to have a close spatial relation to HU_{close}. For intra-individual normalization, differences between HU_{close} and $HU_{distant}$ (HU_{Δ}) and the fat attenuation ratio $HU_{ratio} = HU_{close}/HU_{distant}$ were computed to assess density alterations in PVAT. Both measures were compared in two experiments.

- Inter-individually (AAA vs controls): Assessment of HU_{Δ} and HU_{ratio} in the aneurysmal section of the aorta of AAA patients compared against the infrarenal aortic section of controls.
- Intra-individually (aneurysmal vs non-aneurysmal sections): Assessment of HU_{Δ} and HU_{ratio} in the aneurysmal section of the aorta of AAA patients compared to non-aneurysmal segments inferior of the coeliac trunc to the aortic bifurcation in the same patients.

SciPy 1.6.3 was used for statistical analysis[16]. Differences between the groups were assessed by two-tailed t test for independent (i) and dependent (ii) samples. If at least one of the two analysed areas (close or distant PVAT) was smaller than 0.3 cm^3, the patient was excluded from the respective analysis. Figure 3 illustrates the investigated regions.

Automation by AI

To assess the potential of AI for automation of future analysis of PVAT in AAA, all annotated images of the study were used to train a (convolutional neural network) CNN for segmentation of the aorta used for centerline extraction and for segmenting aneurysmal sections for localization of AAA. The nnU-Net framework was used to implement a patch-wise generic U-Net with 3D convolutions that divides the input images into patches of size $112 \times 112 \times 192$ with resolution of $0.72 \times 0.72 \times 0.8$ mm, which corresponds to a field of view of $80 \times 80 \times 153$ mm[17,18]. This U-Net was optimized by stochastic gradient descent with Nesterov momentum of 0.99, with a decreasing learning rate starting from 0.01, a batch size of two patches, and oversampling of the foreground voxels. More detailed information on pre-processing, model architecture, and training hyperparameters are presented in Supplement S3.

Annotations from 180 patients (90 with AAA, 90 controls) were used in fivefold cross-validated training. Ensembles of the cross-validated models were evaluated on a hold-out test set (10 with AAA, 10 controls). The segmentation performance was assessed using the Dice score. The mean deviation of predicted and actual inferior and superior ends of the AAA was assessed as performance metric for the localization of the aneurysmal section. For the hold-out test set intraclass correlation coefficients (ICC) were determined for HU_{Δ} and HU_{ratio} calculated by model segmentation and by ground truth annotation. For AAA patients the metrics were determined in aneurysmal section, for the controls in the infrarenal aortic section. ICC was determined in SPSS 27.0.0 with 95% confidence intervals (CI). Also the mean difference per patient between HU_{Δ} and HU_{ratio} calculated by model segmentation and by ground truth is given with CI calculated by bootstrapping with 1000 resamples.

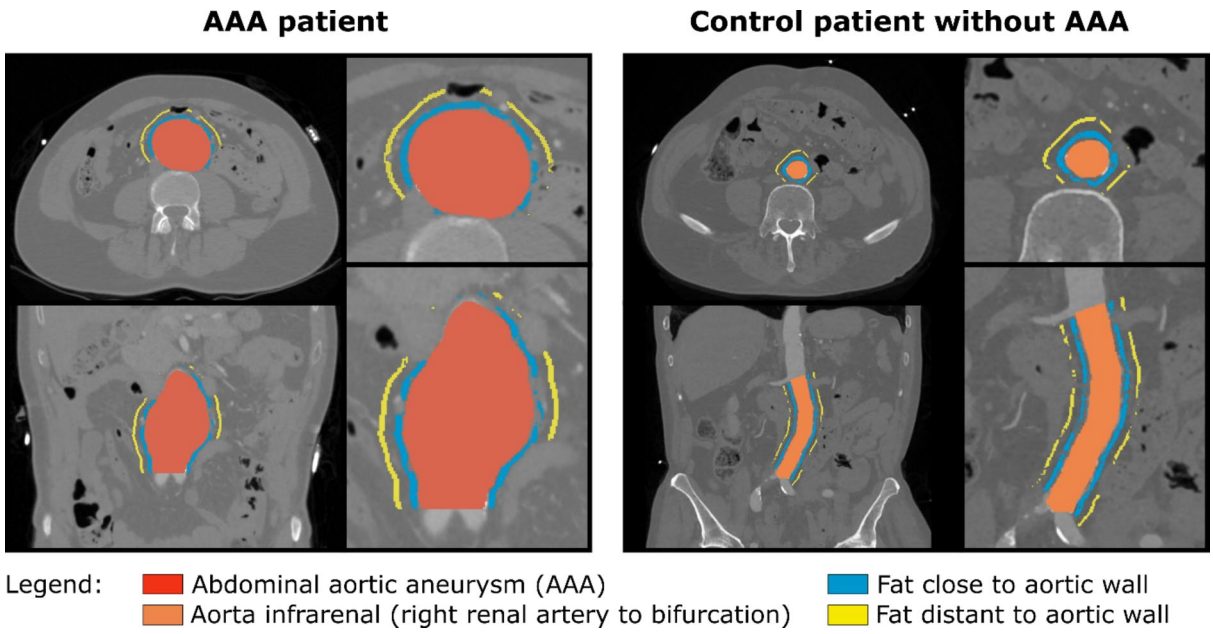


Fig.3. Illustration of the investigated perivascular adipose tissue (PVAT) areas. Mean Hounsfield units (HU) of the PVAT within 2 to 5 mm distance (blue) and within 10 to 12 mm distance (yellow) to the aortic wall were assessed and are referred to as HU_{close} and $HU_{distant}$.

Code availability

Code for training the U-Net model (https://github.com/MIC-DKFZ/nnUNet) and for evaluation of PVAT based on segmented aortas (https://github.com/ukb-rad-cfqiai/AAA_CT_fat_attenuation_evaluation) is publicly available.

Results

Patient characteristics and aortic assessment

Table 1 shows the general patient characteristics. In the AAA group (80% male, 20% female, age: 76.3 ± 9.3 years) the mean of the maximal aortic diameters was 49.8 ± 13.4 mm with 63% of aneurysm showing fusiform configuration, 37% saccular. The mean aneurysmal volume was 120.5 ± 117.6 ml with 25th / 75th quartiles of 49.3 ml / 135.0 ml. The mean aneurysmal length was 116.7 ± 89.4 mm with 25th / 75th quartiles of 59.3 mm / 137.4 mm. In the control group (80% male, 20% female, age: 76.8 ± 9.5 years) the mean of the maximal aortic diameters was 20.7 ± 2.7 mm. The majority of aneurysms showed an extensive degree of thrombosis and non-calcified plaques (75% > 1 cm, 35% > 2 cm max diameter of thrombosis). In the control group the amount of non-calcified plaque was low (90% none, 10% < 0.5 cm diameter). In the AAA group 54% of patients showed calcified atherosclerotic wall alteration affecting more than 50% of circumference, 41% more than 75%. In the control group 71% of patients showed less than 25% circumferential calcified atherosclerotic wall alterations.

Assessment of perivascular adipose tissue

Table 2 and Fig. 4 show the results of (i) inter-individual and (ii) intra-individual evaluation of HU_{Δ} and HU_{ratio}. For both analyses, one AAA patient was excluded due to very low PVAT volume of less than 0.3 cm^3 in the

	Abdominal aortic aneurysm	Control
Male	80/100 (80%)	80/100 (80%)
Female	20/100 (20%)	20/100 (20%)
Average age	76.3 ± 9.3 years	76.8 ± 9.5 years
Thrombosis/non-calcified plaque max. diameter*: 0 1 2 3 4	 8% 5% 12% 40% 35%	 90% 10% 0% 0% 0%
Aortic wall sclerosis percentage# 0 1 2 3 4	 7% 20% 19% 13% 41%	 16% 55% 15% 11% 3%
Average max. abdominal aortic diameter	49.8 ± 13.4 mm	20.7 ± 2.7 mm
Mean aneurysmal volume	120.5 ml ± 117.6 ml	–
25%/75% quartiles aneurysmal volume	49.3 ml / 135 ml	–
Mean aneurysmal length	116.7 mm ± 89.4 mm	–
25%/75% quartiles aneurysmal length	59.3 mm / 137.4 mm	–

Table 1. Patient characteristics of the abdominal aortic aneurysm and the control group. *Thrombosis/non-calcified plaque maximal diameter: 0 = none, 1 = < 0.5 cm 2 = 0.5–1 cm, 3 = 1–2 cm, 4 ≥ 2 cm. #Aortic wall sclerosis percentage: 0 = none, 1 = < 25%, 2 = 25–50%, 3 = 50–75%, 4 ≥ 75%.

Cohort	Investigated section	HU_{close}	$HU_{distant}$	HU_{Δ}	HU_{ratio}
	(i) Inter-individual (AAA vs controls)				
AAA	Aneurysmal segment	−75.0 ± 11.4	−83.9 ± 13.5	8.9 ± 5.1	89.8 ± 5.7
Control	Infrarenal segment	−76.5 ± 10.4	−83.4 ± 12.5	6.9 ± 4.8	92.1 ± 5.5
	P-values	0.337	0.808	**0.006**	**0.004**
	(ii) Intra-individual (aneurysmal vs non-aneurysmal segments)				
AAA	Aneurysmal segment	−75.0 ± 11.4	−83.9 ± 13.5	8.9 ± 5.1	89.8 ± 5.7
AAA	Unaffected segment	−73.0 ± 10.5	−78.5 ± 11.9	5.5 ± 4.1	93.3 ± 4.9
	P-values	**< 0.001**	**< 0.001**	**< 0.001**	**< 0.001**

Table 2. Inter-individual and intra-individual comparison of PVAT in AAA patients and controls. $HU_{close:}$ mean attenuation of PVAT in Hounsfield units within 2 to 5 mm distance from the aortic wall. $HU_{distant:}$ mean attenuation of PVAT in Hounsfield units within 10 to 12 mm distance from the aortic wall. HU_{Δ}: Mean difference of attenuation. $HU_{ratio} = HU_{close} / HU_{distant}$. Significant values are in bold.

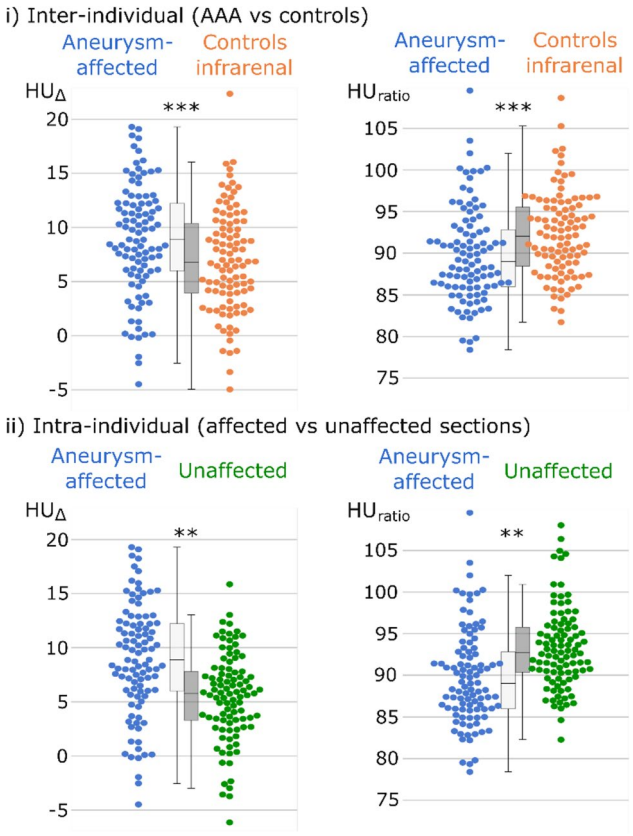


Fig.4. Swarm and boxplots of the investigated perivascular adipose tissues (PVAT). Difference of PVAT attenuation from close and distant measurements (HU_{Δ}), along with the perivascular adipose attenuation ratio (HU_{ratio}) in % displayed in swarm and box plots. Differences between aneurysm- aneurysmal sections of the aorta (blue) in AAA patients in comparison to the infrarenal section of control patients (orange) were investigated by two-tailed t-test for independent samples. Intra-individual differences between the aneurysmal (blue) and non-aneurysmal aortic sections (green) in AAA patients, were investigated by two-tailed t-test for related samples. **: P-value ≤ 0.01; ***: P-value ≤ 0.001.

aneurysmal segment. Another three patients with AAA were excluded from the intra-individual analysis due to PVAT volumes < 0.3 cm^3 in the non-aneurysmal segment.

In inter-individual comparisons between patients with AAA and controls (i), higher HU_{Δ} and a lower HU_{ratio} were observed in the aneurysmal segments in comparison to the infrarenal segments of control patients (AAA: 8.9 ± 5.1 HU vs. control: 6.9 ± 4.8 HU, p-value = 0.006; AAA: 89.8 ± 5.7% vs. control: 92.1 ± 5.5% p-value = 0.004).

In intra-individual comparisons patients with AAA (ii) showed higher HU_{Δ} and lower HU_{ratio} in the aneurysmal segments compared to non-aneurysmal segments of controls (aneurysmal: 8.9 ± 5.1 HU vs. non-aneurysmal: 5.5 ± 4.1 HU, p-value < 0.001; aneurysmal: 89.8 ± 5.7% vs. non-aneurysmal: 93.3 ± 4.9% p-value < 0.001).

Automation by AI

The nnU-Net model segmenting the aorta from the right coronary artery to bifurcation that can be used for automating centreline extraction achieved a mean Dice score of 0.972 ± 0.035 on the 180 patients of the validation sets in fivefold cross-validation and 0.977 ± 0.016 (patient with AAA 0.972 ± 0.021; control patients: 0.983 ± 0.007) on the hold-out test set.

The mean differences per patient between HU_{Δ} and HU_{ratio} calculated based on model segmentation and ground truth annotation evaluated on the test set were low (HU_{Δ} 0.21 [CI 0.01–0.82], HU_{ratio} 0.34% [CI 0.01–1.13%]). ICC analysis also revealed high agreement (HU_{Δ} 0.97 [CI 0.93–0.99], HU_{ratio} 0.96 [CI 0.90–0.98]). The nnU-Net model for locating aneurysmal sections by segmentation achieved a mean deviation between predicted and actual inferior ends of 3.54 ± 4.61 mm and superior ends of 8.22 ± 9.23 mm on the validation sets in fivefold cross-validation. In the hold-out test set mean deviations were 9.01 ± 13.41 mm for inferior and 8.34 ± 7.19 mm for superior ends of AAA.

Discussion

The main findings of this study are that the analysis of PVAT density, as a surrogate marker of perivascular inflammation, reveals differences in AAA patients not only when compared to a matched control group but also compared with non-diseased segments of the aorta in intra-individual comparison. This indicates the potential

of PVAT density analysis as a basis for non-invasive imaging biomarkers in AAA. Furthermore, AI assisted assessment of PVAT density appears feasible, which will facilitate the use and further analysis of respective image-based markers in future large-scale studies.

While aneurysmal diameter and the degree of intraluminal thrombosis are well-established parameters used to predict the risk of AAA growth and related events, the assessment of PVAT as a means to evaluate vascular inflammation in the formation and progression of AAA has scarcely been investigated[15,19]. The perivascular adipose tissue is actively involved in the maintenance of vascular homeostasis[20,21], and its inflammation has been linked to atherosclerotic changes and hypertension[6]. Additionally, inflammation of the juxta-aortic perivascular fat has been associated with AAA formation[6,8]. A persistent PVAT inflammation can lead to extracellular matrix degradation and vascular wall thinning[8]. While invasive assessment of PVAT is not clinically feasible due to inherent risks of biopsy, CT based analysis allows for non-invasive assessment of PVAT density.

Although PVAT density has been shown to correlate with the degree of perivascular inflammation and coronary atherosclerosis in several studies[11,22], its analysis for the thoracic aorta has failed to demonstrate any correlation with histopathological findings[23]. The composition of PVAT varies depending on anatomical location, with peripherally increased amounts of white-like adipose tissue compared to the predominantly brown adipose tissue surrounding the thoracic aorta[20,24]. Therefore, PVAT density measurements are expected to vary depending on the anatomical region and to have a variable degree of correlation with inflammatory changes of PVAT. Two CT-based studies examining PVAT of the abdominal aorta have reported increased density values for AAA[15,19]. In fact, Yamaguchi et al. found that PVAT density could serve as a predictor of AAA growth[15]. The present findings provide additional backing for the concept of PVAT alteration adjacent to aneurysmal sections of the aorta, thus requiring further research into PVAT composition as potential predictors of AAA development and progression.

In addition to demonstrating alterations in perivascular adipose tissue density in patients with AAA compared to patients without aortic pathology, similar differences in PVAT density could also be demonstrated by us in intra-individual comparisons of diseased and non-diseased sections of the AAA group. This supports the assumption of inflamed PVAT surrounding AAAs instead of a generally altered perivascular attenuation in patients with aneurysmal disease. The inclusion of age- and sex-matched controls further eliminated other potential biases, such as age-related changes in PVAT density.

In the current study, PVAT density measurements were performed slightly different to those reported in previous studies[15,19]. Partial volume artefacts from intra-aortic iodine, beam hardening and blooming artefacts from mural calcifications can potentially cause an artificial elevation of perivascular attenuation. Absolute attenuation differences found in this as well as previous studies are quite low[15,19,25]. Therefore, also minimal artefacts could potentially significantly influence PVAT analysis. Thus, to reduce artefact interference, the area directly adjacent to the aortic wall was excluded from analysis in our study. Furthermore, instead of the absolute value of PVAT density, we focused on the difference and the ratio between the densities of fat tissue closer to the vessel compared to distant fat tissue as normalized attenuation measures. Another strength of this study is the selection of the control group. The selection process included matching for age and sex, as well as exclusion of patients with potential aortic/periaortic disease.

This is the first study, to our knowledge, investigating AI assisted assessment of aortic PVAT. Incorporating AI in this study provided several benefits, including simplification and abbreviation of aortic annotation in the current cohort and the automation of future analysis. The AI algorithm could help in the standardization and accessibility of PVAT assessment in AAA and enable the transfer of the analysis to potential research collaborators. To ensure practicality for CT images of clinical routine, the aortic segmentation method ideally should allow for assessment of variable scan lengths, including only the thoracic/abdominal aorta or both regions. However, this poses a challenge as a CNN typically requires a fixed input shape. Moreover, segmentations should be performed on high-resolution examinations to allow for precise measurements of PVAT in a region surrounding the aorta of only a few millimetres. To address these challenges a patch-wise 3D CNN was developed with the nnU-Net framework that divides input images into fixed-size patches of high resolution, enabling the analysis of datasets with variable scan lengths. Furthermore, nnU-Net has demonstrated high performance in segmentation challenges, including the aorta[16].

It is important to note that this is a retrospective, single-center investigation. While the number of patients included is relatively high compared to previous studies investigating PVAT in patients with AAA, the sample size may still be considered low in relation to the absolute attenuation differences observed. Moreover, the patient cohort consisted mainly of individuals with advanced AAA, and follow-up data were not collected due to the fact that the majority of patients underwent endovascular aortic repair treatment, thereby limiting the reliability of follow-up measurements. Lastly while it has been shown that inflammatory changes of the PVAT are a major factor in AAA development it remains difficult to prove that alterations in density of PVAT are only based on inflammation. Therefore, long term follow-up studies are required to prove the prognostic implications of CT based PVAT analysis.

Conclusion

The results indicate PVAT density alterations in AAA patients that could have potential to provide valuable imaging markers of perivascular inflammatory alterations. This motivates further research to establish non-invasive imaging markers for vascular and perivascular inflammation in AAA for risk stratification regarding aneurysmal growth and rupture risk. Therefore, multicentre studies should be initiated to investigate AI-based PVAT assessment in AAA patients longitudinally.

Data availability

The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

Received: 5 October 2023; Accepted: 27 August 2024
Published online: 03 September 2024

References

1. Wanhainen, A. *et al.* Editor's choice – european society for vascular surgery (esvs) 2019 clinical practice guidelines on the management of abdominal aorto-iliac artery aneurysms. *Eur. J. Vasc. Endovasc. Surg.* **57**, 8–93. https://doi.org/10.1038/s41572-018-0030-7 (2019).
2. Sakalihasan, N. *et al.* Abdominal aortic aneurysms. *Nat. Rev. Dis. Primers* **4**, 1–22. https://doi.org/10.1038/s41572-018-0030-7 (2018).
3. Chaikof, E. L. *et al.* The society for vascular surgery practice guidelines on the care of patients with an abdominal aortic aneurysm. *J. Vasc. Surg.* **67**, 2-77.e2. https://doi.org/10.1016/j.jvs.2017.10.044 (2018).
4. Wanhainen, A., Mani, K. & Golledge, J. surrogate markers of abdominal aortic aneurysm progression. *Arterioscler Thromb. Vasc. Biol.* **36**, 236–244. https://doi.org/10.1161/ATVBAHA.115.306538 (2016).
5. Yuan, Z. *et al.* Abdominal aortic aneurysm: Roles of inflammatory cells. *Front. Immunol.* https://doi.org/10.3389/fimmu.2020.609161 (2021).
6. Hu, H., Garcia-Barrio, M., Jiang, Z., Chen, Y. E. & Chang, L. Roles of perivascular adipose tissue in hypertension and atherosclerosis. *Antioxid. Redox Signal* **34**, 736–749. https://doi.org/10.1089/ars.2020.8103 (2021).
7. Meekel, J. P. *et al.* Inflammatory gene expression of human perivascular adipose tissue in abdominal aortic aneurysms. *Eur. J. Vasc. Endovasc. Surge.* **61**, 1008–1016. https://doi.org/10.1016/j.ejvs.2021.02.034 (2021).
8. Ye, T. *et al.* Relationships between perivascular adipose tissue and abdominal aortic aneurysms. *Front. Endocrinol. (Lausanne)* **12**, 704845. https://doi.org/10.3389/fendo.2021.704845 (2021).
9. Gao, J.-P. & Guo, W. Mechanisms of abdominal aortic aneurysm progression: A review. *Vasc. Med.* **27**, 88–96. https://doi.org/10.1177/1358863X211021170 (2022).
10. Adachi, Y. *et al.* Beiging of perivascular adipose tissue regulates its inflammation and vascular remodeling. *Nat. Commun.* **13**, 5117. https://doi.org/10.1038/s41467-022-32658-6 (2022).
11. Antonopoulos, A. S. *et al.* Detecting human coronary inflammation by imaging perivascular fat. *Sci. Transl. Med.* **9**, eaal2658. https://doi.org/10.1126/scitranslmed.aal2658 (2017).
12. Goeller, M. *et al.* Pericoronary adipose tissue computed tomography attenuation and high-risk plaque characteristics in acute coronary syndrome compared with stable coronary artery disease. *JAMA Cardiol.* **3**, 858–863. https://doi.org/10.1001/jamacardio.2018.1997 (2018).
13. AWMF Leitlinienregister, S3 guideline for Screening, Diagnostics, Therapy and Follow up of Abdominal Aortic Aneurysms (2018) Available via: https://register.awmf.org/assets/guidelines/004-014m__S3_Bauchaortenaneurysma_2018-08.pdf (2018) Accessed September 30, 2023
14. Fedorov, A. *et al.* 3D slicer as an image computing platform for the quantitative imaging network. *Magn. Reson. Imaging* **30**, 1323–1341. https://doi.org/10.1016/j.mri.2012.05.001 (2012).
15. Yamaguchi, M. *et al.* Clinical significance of increased computed tomography attenuation of periaortic adipose tissue in patients with abdominal aortic aneurysms. *Circ. J.* **85**, 2172–2180. https://doi.org/10.1253/circj.CJ-20-1014 (2021).
16. Virtanen, P. *et al.* SciPy 1.0: Fundamental algorithms for scientific computing in python. *Nat. Methods* **17**, 261–272. https://doi.org/10.1038/s41592-019-0686-2 (2020).
17. Nowak, S. *et al.* End-to-end automated body composition analyses with integrated quality control for opportunistic assessment of sarcopenia in CT. *Eur. Radiol.* **32**, 3142–3151. https://doi.org/10.1007/s00330-021-08313-x (2022).
18. Isensee, F., Jaeger, P. F., Kohl, S. A. A., Petersen, J. & Maier-Hein, K. H. nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation. *Nat. Methods* **18**, 203–211. https://doi.org/10.1038/s41592-020-01008-z (2021).
19. Dias-Neto, M. *et al.* High density of periaortic adipose tissue in abdominal aortic aneurysm. *Eur. J. Vasc. Endovasc. Surg.* **56**, 663–671. https://doi.org/10.1016/j.ejvs.2018.07.008 (2018).
20. Gil-Ortega, M., Somoza, B., Huang, Y., Gollasch, M. & Fernández-Alfonso, M. S. Regional differences in perivascular adipose tissue impacting vascular homeostasis. *Trends Endocrinol. Metab.* **26**, 367–375. https://doi.org/10.1016/j.tem.2015.04.003 (2015).
21. Szasz, T., Bomfim, G. F. & Webb, R. C. The influence of perivascular adipose tissue on vascular homeostasis. *Vasc. Health Risk Manag.* **9**, 105–116. https://doi.org/10.2147/VHRM.S33760 (2013).
22. Mancio, J., Oikonomou, E. K. & Antoniades, C. Perivascular adipose tissue and coronary atherosclerosis. *Heart* **104**, 1654–1662. https://doi.org/10.1136/heartjnl-2017-312324 (2018).
23. Gaibazzi, N. *et al.* The histopathological correlate of peri-vascular adipose tissue attenuation on computed tomography in surgical ascending aorta aneurysms: Is this a measure of tissue inflammation?. *Diagnostics (Basel)* **11**, 1799. https://doi.org/10.3390/diagnostics11101799 (2021).
24. Li, X., Ma, Z. & Zhu, Y. Z. Regional heterogeneity of perivascular adipose tissue: Morphology, origin, and secretome. *Front. Pharmacol.* **12**, 697720. https://doi.org/10.3389/fphar.2021.697720 (2021).
25. Yuvaraj, J. *et al.* Pericoronary adipose tissue attenuation is associated with high-risk plaque and subsequent acute coronary syndrome in patients with stable coronary artery disease. *Cells* **10**, 1143. https://doi.org/10.3390/cells10051143 (2021).

Author contributions

SN, AMS, DK and DG contributed substantially to the study design. DG contributed substantially to the acquisition of patient data and the manual landmark definition, segmentation and optimization of the automated segmentations. SN developed the automatic segmentation method and guided DG in the use of 3D Slicer. SN contributed substantially to the computer vision analysis, statistical analysis and statistical interpretation. DK and AMS supervised data acquisition, statistical analysis, interpretation. DG and SN wrote the main article text. SN created all figures. DK, AMS, JAL, UA participated in drafting and revising the article. All authors revised the version to be published critically for important intellectual content and gave final approval.

Funding

Open Access funding enabled and organized by Projekt DEAL. The study was funded by the diagnostic and interventional department of radiology, university hospital Bonn.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-024-71283-9.

Correspondence and requests for materials should be addressed to D.K.

Reprints and permissions information is available at www.nature.com/reprints.

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

4 Discussion

In this thesis, the applicability of AI for automated analyses of radiological images and free-text reports was investigated, with the aim to provide insights into the potential of AI for advancing radiological workflows. Radiology has traditionally been at the forefront of digital medicine. This is due to the digitalization of the two basic steps of the radiological examination as in recent decades, namely image acquisition and the subsequent image assessment, where the radiologist writes a radiological report also considering other relevant clinical information. Consequently, AI has potential for improving various areas in radiology with respect to image and medical document analysis (Hosny et al. 2018; Hagiwara et al. 2020; Haase et al. 2023; Pierre et al. 2023). The following sections discuss a selected subset of the diverse use cases of AI in radiological workflows. It will be explained what the current state of research is, what future directions might be and if so, how this thesis contributes to this research area.

4.1 Processing of free-text radiological reports

After its release in 2023, ChatGPT by OpenAI was one of the fastest growing web/software applications of all time and shifted the public discourse to an anticipation that transformer-based LLMs will profoundly impact the day-to-day operations of numerous knowledge-based professions, including the medical and radiologic professions (Nowak and Sprinkart 2024; Eloundou et al. 2023). Since then diverse use cases of this new technology were investigated by radiology researches (Elkassem and Smith 2023; Rau et al. 2023; Li et al. 2023; Nowak and Sprinkart 2024).

Extracting structured information from free-text report databases

Traditionally, radiologists' assessments and diagnoses are recorded in free text. However, reporting in predefined, machine-readable structures is preferable as this offers improved clinical usability for the assigning physicians and also enables secondary use of clinical databases (Schwartz et al. 2011; ESR 2018). This would include conducting large, retrospective epidemiological studies or the development of predictive AI models. LLMs could be utilized to capture the content of free-text reports stored in unstructured, digital clinical databases to enable their secondary use.

A current systematic review, solely focusing on ChatGPT and OpenAI-based services, shows that the current radiological community strongly focuses on evaluating closed LLMs (Sacoransky et al. 2024). This is also reflected in a PubMed search conducted at 11/20/2024, which yields 501 results for the last 5 years for ("radiology") AND (("GPT-") OR ("chatGPT")), but only 19 studies are found for ("radiology") AND (("large language model" OR "LLM")) AND (("open-source" OR "open-weights")). This focus on closed LLMs ignores data protection issues associated with processing sensitive patient information on external servers, which would be resolved by local processing through open LLMs (Bhayana 2024). Consequently, the utility of zero-shot prompting the privacy-ensuring vicuna-13b with open-weights was investigated in a recent study published in Radiology (Mukherjee et al. 2023).

The first two papers included in this thesis, published in the journals Radiology and European Radiology, significantly extend the knowledge on the use of privacy-preserving open transformers for structured free-text report databases. The application of BERT, 17 open LLMs and 4 models of the closed GPT series of OpenAI were evaluated on two large datasets including English and German reports. Furthermore, different approaches were investigated, including zero-/one-shot prompting and also fine-tuning of 4-bit quantized, frozen, open LLMs with low-rank adaptation and different amounts of training data. Our work essentially provides a radiological zero-shot and fine-tuning benchmark of current open LLMs with Mistral-Large created by the French startup Mistral AI showing promising results comparable to GPT-4o, which was the frontier model of OpenAI at the time of experiments. Thereby, this thesis provides important insights to the current scientific state of using LLMs on radiological data.

The third work of this thesis investigates whether automated, transformer-based report content extractions can be used as labels for the supervised training of image-based AI methods for diagnosis support. The results demonstrate that transformers have the potential to unlock clinical databases for secondary use. However, our results also indicate possible limitations of employing labels derived from a single radiological report. We identified several reasons for mismatch of content between report and image including not mentioning of findings due to low relevance for the current indication, conclusions based on information that are not content of the radiological report e.g. clinical/laboratory parameters, mentioning of borderline image findings as definite, or simply by errors of the transformer. This indicates the need for more sophisticated labels based on patient-centred databases including numerous documents and information derived from all clinical databases, contrary to labels solely based on single radiological reports. In principle, LLMs could be used for content extraction of various medical documents beyond radiological reports.

Retrieving information by interaction with patients

The design of current LLMs, that are post-trained with methods like RLHF to act as helpful chatbots, could be employed for offering interactive chats with patients to retrieve information, e.g. about symptoms, family history, risk factors, or medications (Tam et al. 2024). Also, informative consultations of patients by LLM-based chatbots could boost operational efficiency of the radiological workflow. Furthermore, interactive LLMs could benefit the patient experience by comprehensive 24/7 support that allows patient education on diseases or treatments in a controlled environment. This could be preferable to unguided internet research by the patient that can result in increased anxiety and stress (Hao et al. 2024). However, recent research investigating the correctness of patient questions on head and neck cancer showed that answers from OpenAI's ChatGPT were incorrect in 8% of cases (Wei et al. 2024). This indicates that foundational models, such as the GPT-4o backed ChatGPT, still lack comprehensive knowledge on medical topics. Consequently, there is a need for studies that investigate whether incorporating task-specific knowledge into LLMs, e.g. by retrieval augmented generation (RAG) or model fine-tuning, can contribute to enhancing accuracy and preventing harm to patients by incorrect statements. In our work, we demonstrated that fine-tuning LLMs with a remarkably small dataset of just 500 examples

sourced from clinical databases can already enhance task-specific medical knowledge. This has led to improved accuracy in *comprehending* and extracting content from radiological reports. Notably, while this thesis advances the application of LLMs in clinical data analysis, its scope does not extend to the direct application of AI in patient interaction.

Communication between radiologist and patients

A major obstacle in communication between doctors and patients is often that patients without medical training do not understand even basic medical and radiological terms, leading to misunderstandings and potentially fueling patient concerns (Yi et al. 2019). A study examining over 90,000 radiological reports at a US clinic found that merely 4% of the reports were written at a level readable by an average educated adult (Martin-Carreras et al. 2019). However, in the spirit of patient-centered care, there appears to be an increasing need and benefit for direct communication between radiologists and patients (Kemp et al. 2017).

Therefore, a study of our clinic investigated the potential of ChatGPT (backed by GPT-4) to generate a simplified version of a cardiac MRI report for patients, while maintaining correct content (Salam et al. 2024). It could be shown that using GPT-4 doubled readability of reports from 5 to 10 for 12 laypersons, measured by the automated readability index. Also, radiologists strongly agreed with the factual correctness in 94% and completeness in 81% of the simplified reports. Although this seems as a promising application for improving radiological workflow, future works should investigate the applicability of locally implemented open LLMs for this task, as this would omit retrieving patient consent for processing radiological reports on external servers of OpenAI and would prevent patient risks when sharing health information with foreign companies. As indicated in our work included in this thesis, task-specific fine-tuning of open LLMs is feasible and could also lead to improved results with respect to completeness of simplified report versions. This could be investigated in future work.

Supporting the reporting process

The primary objective of a radiological examination is to determine an accurate diagnosis for the given clinical indication based on the imaging. This diagnosis, along with other relevant findings, is commonly communicated to the referring physician through a free-text report.

AI might have potential to improve quality and efficiency of this central part of the radiology workflow (Pierre et al. 2023). This can be done at different levels of AI involvement. Prior to image acquisition and reporting, LLMs could provide summarizations illustrating the patient history from previous reports and documents, e.g. to decrease risks due to missing information on potential allergies to medication or contrast agents (Lee et al. 2024). After creation of the report, LLMs could review the content of the report for completeness or discrepancies with respect to relevant clinical guidelines. Prior to giving suggestions for report improvement, RAG could be employed to infuse the prompt with relevant context from embedding vector databases that are based on relevant medical guidelines. Previous works have shown that this approach enables LLMs to accurately cite context of various imaging or disease focused guidelines (Rau et al. 2023; Kresevic et al. 2024; Ferber et al. 2024). However, most of the research focuses on applying closed models, like the GPT series by OpenAI. Again, it should be emphasized that this prohibits the use in real clinical scenarios with real reports

of patients due to data protection regulations. Future work should evaluate open LLMs for guideline-focused report review.

Directly generating free-text reports from imaging

A question that has been controversially discussed among radiologists since the first promising applications of AI on radiological images, is whether AI will replace radiologists in the future. Here, the prevailing and reasonable opinion has formed that task-specific and tool-based AI, as known in 2019, will help make workflows more efficient, and thereby “Radiologists who use AI will replace radiologists who don’t” (Langlotz 2019).

However, as also a recent review stated, generative transformer models, especially those capable of processing multimodal input, have sparked new excitement in radiological research. Ultimately, the direct generation of free-text reports from radiological images, with companioning text-based clinical information, would constitute a remarkable achievement, essentially mirroring the current core task of a radiologist. A recent qualitative analysis investigated the closed GPT-4 vision for its zero-shot ability to interpret radiological images (Wu et al. 2023). However, significant limitations with respect to accurate diagnoses were found within the free-text reports generated by the LLM. Importantly, the correctness and completeness of reports is crucial for the feasibility of report generating systems. This shortcoming of current, generally pre-trained frontier models indicates the need for multimodal transformer models that were directly trained to generate radiological reports with correct and complete content.

Existing research on this topic primarily focuses on training transformer models from scratch for this task. For example, some work propose to train a classical encoder-decoder transformer architecture as introduced in "Attention is all you need" in 2017, however with the encoder being a vision transformer (Vaswani et al. 2017; Sloan et al. 2024). These approaches essentially see the task as a machine translation from image to report. Other research utilizes a CLIP pre-training with radiological image/report pairs, as with the prominent text to image generating models for natural images, like DALL·E 2 by OpenAI (Sloan et al. 2024; Radford et al. 2021; Ramesh et al. 2022). During CLIP pre-training, the feature representations of an image-encoder and text-encoder model are aligned by mapping corresponding image-text pairs. In a post-training phase, the feature representations of the image encoder are used as a seed for a text generating model. Note, this is inverse to the use of the feature representations of the text encoder for generating images, as in DALL·E 2 (Ramesh et al. 2022).

With the recent release of Meta's Llama-3.2 and Mistral AI's Pixtral model series, extensively pre-trained open LLMs are now available, which can process both image and text data as input and generate coherent text output (Agrawal et al. 2024). Notably, Pixtral-Large, a multimodal version of the 123b Mistral-Large was released, which showed promising results in the first work of this thesis for report content comprehension and extraction. Open models, like Pixtral, enable local training and deployment directly within the clinical infrastructure in contrast to proprietary systems, like OpenAI's GPT or Google's Gemini models, which would require transmitting sensitive patient data to external servers. As mentioned repeatedly, this resolves data privacy issues that would prohibit AI-based report generation in clinical practice.

In future work, we will investigate local chest X-ray report generation by pre-training Pixtral with causal language modelling to learn about general aspects of chest X-ray images and reports. Then, in a RLHF inspired post-training phase, we will employ the fine-tune BERT from the work included in this thesis as a reward model to improve correctness of the report content. Low-rank adaption of frozen, quantized models, as used in the first work of this thesis, or novel approaches as unfrozen training using low-rank projection of the gradients, could enable training of Pixtral also with limited hardware resources (Zhao et al. 2024a).
The impact on clinical practice and radiological workflows of a local, privacy-compliant system for pre-reporting radiological imaging could be substantial. However, all described approaches, existing and planned, are currently limited to two-dimensional images only. It remains unclear if these ideas will be extendable to three-dimensional, cross-sectional images, such as to CT or MRI, in the future. Possible solutions could be concatenating slices into a very large two-dimensional input image, where models with immense context size, as seen with Googles Gemini model series, are required. Another solution would be training models from scratch using three-dimensional vision-encoders with reduced computational complexity, such as Swin transformers.

4.2 Generating synthetic radiological images

As described before, new generative AI methods have been at the center of public discourse since at least 2023. To a minor extent, this includes also image-generating models in addition to text-generating LLMs. One of the first approaches that showed convincing results in artificially generating synthetic images were GANs introduced in 2017 (Goodfellow et al. 2014). Later, CLIP pre-trained text-encoders paired with diffusion models, as introduced with the image-generating DALL·E 2, demonstrated outstanding capabilities in generating images from text prompts (Ramesh et al. 2022). Radiological researchers have investigated various methods for generating image content for application in radiological workflows.

Generation of virtual contrast agent reduced or contrast agent free images

This is one of the central areas where this thesis makes contributions to the current state of research. Contrast agents are required in radiological imaging to e.g. enhance the visibility of internal body structures or to evaluate increased contrast agent uptake of specific tissues or lesions. For example, in MRI, gadolinium-based contrast agents (GBCAs) are predominantly applied. Although the use of GBCAs is rated as predominantly safe, there are risks such as rare acute allergic reactions, nephrogenic systemic fibrosis especially with impaired renal function, or gadolinium deposits in brain or organs with unclear long-term consequences (Starekova et al. 2024). Beside patient safety, also unknown environmental effects of substantial GBCA excretion prompt reasonable usage of contrast agents (Dekker et al. 2024).
Consequently, radiological research has investigated the use of generative AI methods to create virtual signal enhancing effects of GBCAs in brain MRI based on native images without contrast agent use or based on images acquired with reduced contrast agent administration (Mallio et al. 2023). For example, researchers from the Department of Neuroradiology at the University Hospital Bonn employed imaging of 138 participants to develop and evaluate a GAN

consisting of an image-to-image generating U-Net and a PatchGAN discriminator. The authors train this model to transform a T1-weighted MRI image acquired with 10% of GBCA dose to an image appearing at full contrast-enhancement (Haase et al. 2023). Although the virtual images did not achieve comparable contrast-enhancements to the sequence with full GBCA dose, results compared to zero contrast images were promising.

Most of the current research on AI for contrast agent reduced or contrast agent free MRI focuses on brain imaging. This is likely due to the need of precise alignment of tissues shown in the input and target image to leverage training losses that do pixel-wise comparison. This alignment is substantially easier to achieve in brain imaging compared to abdominal or thoracic imaging due to breathing motion and non-rigid deformable organs and tissues (Pasquini et al. 2022). This challenge further increases for cardiovascular imaging, where the image acquisition also has to be triggered at a particular phase of the beating heart. Beside these challenges, a recent study employs multiple U-Net generators to create virtual late gadolinium enhancement (LGE) images from pre-contrast cine frames and native T1-maps using cardiac MRI of 1,348 patients (Zhang et al. 2021). In a later work, the same group could show the potential of virtual LGE for diagnosis of myocardial infarction (Zhang et al. 2022).

The fourth work included in this thesis and published in the Journal of American Heart Association contributes to the current state of research on AI-based virtual contrast in cardiac MRI. This work demonstrates the potential of a GAN with a U-Net generator and PatchGAN discriminator to generate virtual CE T1 maps from native T1 maps using data from 1,000 patients from the University Hospital Bonn and Cologne. We anticipated challenges due to patient-dependent factors influencing how rapidly the contrast agent is distributed and excreted, such as renal function (Haaf et al. 2016). This introduces variabilities to the quantitative values of the CE T1 maps that should not be predictable solely from the native maps. To tackle this challenge, we trained the GAN on CE T1 values in driven equilibrium 10 minutes after contrast administration. Also, our primary goal was the use of AI-generated, virtual CE T1 for contrast agent free ECV, for which the variability stemming from contrast agent distribution is further normalized by bringing differences of native and CE R1 values of the myocardium and the blood pool into relation (Haaf et al. 2016). Future work, should investigate a direct prediction of ECV from native T1 maps, also omitting the requirement of hematocrit blood values.

As described in the beginning of this section, new AI approaches driven by diffusion models emerged for image generation and showed exceptional results also compared to GANs. Similarly to LLMs, many of these models are transformer based, have billions of parameters, are extensively pre-trained and some models are released under open licenses for local implementation. Prominent examples are the Stable Diffusion model series by Stability.ai or the FLUX model series by Black Forest Labs. Future work could investigate if current results on virtual contrast imaging can be exceeded by fine-tuning these novel image generating models. Similar to the first work of this thesis, low-rank adaptation of frozen, quantized models could be employed also for fine-tuning diffusion models to lower computational requirements, as done in a recent study that published their method under open-source license on Github (Parmar et al. 2024).

Improving image quality or image resolution

AI can also benefit radiological imaging by improving image quality, i.e. by suppressing image noise or artifacts, or by enhancing image resolution (Hosny et al. 2018). The work included in this thesis does not make contributions to this field. However, a brief introduction to this research area will be given in the following sections, due to its importance for improving radiological workflows.

Data from MRI and CT detector hardware is acquired in different spaces compared to the human-readable, pixel-based image space and thereby must be reconstructed. For CT, this is the projection space, which essentially represents the attenuation of the X-rays measured at different angles of the rotated detector. For MRI, this is the k-space that represents the measured MR signal in the frequency domain, where each point in k-space corresponds to a particular spatial frequency of the object being imaged. Sampling density in k-space and extension of the k-space is crucial for the quality of the MRI image after reconstruction, but increased sampling directly translates to longer acquisition times (Pal and Rathi 2022).

In general, AI-supported image reconstructions aim for three possible improvements that are often achieved in combination by a single or by multiple AI models in a pipeline. These are improved image quality by e.g. suppressing image noise or artifacts, improved image resolution e.g. by up-sampling images from 256x256 to 512x512 and lastly, especially for MRI, improved acquisition speed, by e.g. generating high quality images from under-sampled k-space data. To achieve this, AI models can be applied for image improvements solely in image space. For example, researchers trained a CNN to up-sample from lower resolution MRI images of the knee to higher resolution images with improved results compared to classical interpolation methods (Chaudhari et al. 2018). Another approach is to use AI models for improved reconstruction, i.e. the translation of the image from k-space to image space. As an example, researchers trained a GAN architecture to reconstruct cardiac MRI images from k-space with suppressed motion artifacts (Oksuz et al. 2018). Lastly, AI models can also be trained for data manipulations directly in k-space. For example, researchers improved brain MRI acquisition speed while maintaining sufficient quality by training two GAN models. The first, acting solely in the k-space, improves sparse sampling. Then, after inverse Fourier transformation, a second GAN improves image quality by noise reduction solely in the image space (Shaul et al. 2020).

Previous work of our clinic investigated the clinical utility of proprietary super-resolution reconstruction algorithms for prostate and cardiac MRI (Bischoff et al. 2023; Kravchenko et al. 2024). It was found that speed ups of up to 35% can be achieved compared to standard reconstruction of MRI sequences while maintaining diagnostic image quality.

As improving image quality and resolution is also of great interest for general application on natural images, there are several extensively pre-trained diffusion models with open weights and code on Github (Huang et al. 2024). Future work could investigate if these more generally pre-trained methods can also be applied for improving resolution of medical images in a zero-

shot matter or are a good basis for task-specific model adaption with reduced requirements on the amount of training data.

4.3 Supporting diagnostic and treatment decisions based on radiological imaging

The capabilities of AI models for image recognition and pattern detection can also be valuable for diagnostic image-based workflows, for the subsequent treatment decisions, and the monitoring of treatment success or disease progression (Hosny et al. 2018). Furthermore, AI-based extraction of quantitative measurements, and thereby the distillation of information of medical images into single numerical values, can aid in more objective evaluations of pathological tissue alterations. This can contribute to the identification of image-based biomarkers, supporting precision and personalized medicine (Hagiwara et al. 2020). The last four publications included in this thesis contribute to the current state of radiological research on the following topics.

Detection and further characterization of diseases

A primary objective of medical imaging is to detect and interpret pathological alterations of tissues. Examples are the detection of diffuse tissue alterations, like fibrosis, or the interpretation of the contrast uptake of focal lesions to evaluate malignity (Kim et al. 2021). Such systems have potential to complement the radiologist's diagnostic procedure and improve reading accuracy and efficiency (Hosny et al. 2018; Mango et al. 2020). In fact, there are numerous commercial systems available to date that are Food and Drug Administration and/or Conformité Européenne approved (Tadavarthi et al. 2020).

When many retrospective cases with a disease of interest are available, as in the clinical routine databases of large clinics, CNNs can be trained in a supervised fashion to learn to identify disease-specific image features. When applied for inference in future patients, these systems can aid by providing a probability of a specific disease or abnormality being present. In a previous work that was content of my Ph.D. thesis, we hypothesized that a publicly, ImageNet pre-trained CNN can detect a liver cirrhosis in clinical, T2-weighted MRI (Nowak et al. 2021).

Interestingly, we found that using the pre-trained CNN encoder as frozen feature extractor already resulted in classification performance comparable to the readings of a radiologist on a hold-out test set and that further training the CNN encoder did not lead to significant improvement, but rather rapid overfitting.

Besides the detection of pathologies within imaging, CNNs may be able to help answer further important questions about the present condition. Here, examples are the classification of a lesion as benign or malignant, the identification of an unknown primary tumor type of a metastasis, or the etiology, i.e. the underlying cause of a diffuse disease (Herent et al. 2019; Samani et al. 2021; Yasaka et al. 2018).

Therefore, in the fifth work included in this thesis, we have extended our insights on AI-based liver cirrhosis evaluation by further characterization of the etiology of the diffuse liver disease. We investigated if an ImageNet pre-trained CNN encoder, similar to the one applied in the previous work, can distinguish between an alcohol-induced cirrhosis versus a disease caused

by other etiologies. The model demonstrated a reasonable area under receiver operating curve of 0.83 on the hold-out test data. Although there are atrophies and hypertrophies of certain liver lobes that are described as more frequently occurring in alcohol-related disease, physicians currently do not employ information from imaging for their conclusions on disease etiology. Therefore, we could not compare the AI to the reading performance of a radiologist. Conversely, this highlights a central benefit of AI-driven research in radiology, specifically the potential to uncover image-based features of a disease that may be used in visual evaluations by radiologists in future patients, but are simply unknown yet. Therefore, we aimed to identify the features that were relevant for the model's prediction by methods of explainable AI. However, as also stated by other research, the current methods of explainable AI are highly limited and do not provide sufficient, fine-detailed explanation of relevant features, but rather coarse highlighted areas of importance (Ghassemi et al. 2021; Kashefi et al. 2023). As also discussed in later sections, improvements in explainable AI are required to identify novel image features, and to fully understand and trust individual predictions of automated systems, especially in scenarios where failures pose high risk for patient harm.

Prognostic value by opportunistically derived image markers

In radiological imaging, vast amounts of information about the patient are embedded within each image, including sizes, and composition of various tissues or organs. However, in clinical routine, only a fraction of the information embedded in imaging is evaluated visually by the radiologist, except for simple quantitative measurements when required by the specific indication. Therefore, a substantial amount of potentially valuable information is unleveraged, which is the basis for a paradigm termed "opportunistic imaging" (Pickhardt 2022). Opportunistic imaging and body composition analysis became increasingly important in radiological research in recent years, also due to automation of tissue quantifications by AI and retrospective research indicating prognostic value of body composition for various oncologic, cardiovascular, or even liver diseases (Cruz-Jentoft et al. 2019; Lenchik and Boutin 2018; Faron et al. 2020; Luetkens et al. 2020; Praktiknjo et al. 2023; Salam et al. 2023; Nowak et al. 2024).

In multiple previous works conducted during my Ph.D. and at our clinic, we developed AI-tools for quantifying fat, muscle, and organ tissues in CT and MRI, also to enable retrospective clinical research on body composition (Nowak et al. 2020; Nowak et al. 2022; Nowak et al. 2021; Nowak et al. 2023).

In the sixth work included in this thesis and published in European Radiology, we investigated associations of AI-based muscle size and composition measurements to basic clinical attributes, as well as to the prognosis of patients with pancreatic cancer after treatment with high-intensity focused ultrasound. In this cohort, we could report that lower muscle size was associated with female patients, patients with lower body-mass-index (BMI) and patients with higher eastern cooperative oncology group (ECOG) status. However, lower muscle density measurements were not associated with sex, BMI and ECOG, but with higher age. This underscored that muscle waste, described as sarcopenia, and fatty degeneration of the muscles, described as myosteatosis, are not synonymous to each other and that sarcopenia might have stronger prognostic implications based on the investigated cohort.

Commonly the prognostic value of image quantifications and markers is investigated in radiological research by uni- and multivariable CPH regression. CPH results are evaluated with focus on significant p-values of the respective hazard-ratios after optimizing log-risk function that indicate an association of the corresponding variable with patient survival (Cox 1972). However, one can criticize this p-value focused approach, which has several frequently overlooked pitfalls. First, p-values might indicate significant association of a variable to patient prognosis, however the effect size on the model's capability to predict clinical outcomes, e.g. by correctly ranking the survival times of pairs of individuals (concordance index), can still be small to irrelevant. This can lead to overestimating relevancy of markers (Schober et al. 2018; Lee 2016). Second, collinearities between features affect the CPH internal matrix inversion, which can increase the variance of regression coefficients, affecting the conclusiveness of p-values (Suchting et al. 2019; Babalola and Yahya 2020; Xue et al. 2007). Furthermore, with this approach the testing of the significance of hazard-ratios is conducted on the data with which the CPH is optimized, i.e. training data, which is affected by overfitting effects and can limit generalizability of results (Babyak 2004). Lastly, a study design focused on hypothesis testing with evaluation of p-values is affected to wrong positive results due to multiple testing.

Therefore, it might be preferable to apply more sophisticated statistical approaches that are not focused on detecting associations based on p-values, but aim for evaluations on predictive outcome. In a current study, we developed a data-driven approach of selecting markers using elastic-net penalized CPH with cross-validation that offers outcome-oriented, comparable evaluation of marker importance based on concordance-index and feature normalization. Therefore, we aim to contribute to improving the current state of research on image-based markers derived from opportunistic imaging by also releasing the code under open-source license for free use.

Contrary to the above-described extraction of human-defined scalar metrics for analysis in simple CPH models, recent works aim to leverage the capabilities of CNNs to independently identify and extract relevant features from medical images that could be relevant for the prognosis of the patient. Some works achieve this by training CNNs to binary classify the occurrence of a certain event. This could include progression/remission of disease or emergence of metastases from a primary tumor (Xu et al. 2019). However, in contrast to the optimization algorithm used for CPH models, the optimization algorithm of supervised classification does not account for censoring of data, i.e. data for which information about the occurrence of the event is no longer available after a certain period. Therefore, other studies do combine a negative log likelihood loss with CNNs, which can account for data censoring and is used during optimization of CPH models, to create patient hazard predicting AI models that are more flexible to imperfect training data from clinical routine (Kim et al. 2020).

In the seventh work presented in this thesis, published in the European Journal of Radiology, we used a similar approach to investigate the capabilities of AI to directly predict the patient hazard solely based on abdominal CT of patients undergoing transcatheter aortic valve implantation. We hypothesized that given a large cohort of 760 patients, the pattern recognizing capabilities of a CNN could leverage additional information from abdominal CT compared to the human-defined body composition metrics and thereby offer improved

predictive performance. We could show that a CNN encoder, which was pre-trained in an autoencoder setting, showed higher concordance-index on hold-out test data compared to CPH including common muscle-based body composition metrics, age, and sex. Age and sex were not input to the CNN. Interestingly, the CNNs´ predictive capabilities were also higher compared to a CPH model including age, sex, and EuroSCORE2, an established surgical risk score based on thorough information on patient history (Nashef et al. 2012). These results indicate the feasibility of future AI models that leverage additional, currently unused, and moreover, even unknown features from radiological imaging that could aid estimation of disease trajectory or response to specific therapies. These estimations could help in identifying high-risk patients, who may benefit from more aggressive interventions or closer monitoring, or help estimating the risk of treatments for patients in a personalized manner.

In future work using the same patient cohort, we want to extend the insights by investigating the hypothesis if improvements could be achieved by directly predicting the patient hazard based on a 3D crop of the heart.

However, the acceptance of a system that influences treatment decisions based on unknown imaging features would be highly questionable. As also previously stated, methods of explainable AI that give insights into reasons for a specific prediction are in high demand, with current approaches demonstrating limitations (Ghassemi et al. 2021; Kashefi et al. 2023). One may hypotheses that the current limitations of methods that aim to highlight image regions, e.g. by model agnostic gradient-based methods or vision-transformer based attention visualizations, may be overcome by developing multimodal LLMs, as proposed earlier in this discussion. These models could be able to indicate the reasons for predictions in an inherently human understandable format, i.e. by language.

In the last study included in this thesis, we explored the potential of AI for automating Hounsfield density analysis of perivascular adipose tissue. Previous studies demonstrated that inflammatory changes in the coronary arteries are associated with increased radiodensity of the surrounding adipose tissue. It is assumed that this effect could be utilized to define quantitative markers providing information on elevated cardiovascular mortality risk based on imaging of the coronary arteries, thus potentially providing valuable information for treatment decisions (Oikonomou et al. 2018). In our study, we aimed to extend insights into radiodensity alterations of perivascular adipose tissue due to inflammatory effects in abdominal aortic aneurysms.

Highly accurate segmentation of the diseased vessel is crucial, as the inflammatory effect is expected within a few millimeters around the vessel. Additionally, partial volume effects, originating from the vascular wall, influence the radiodensity of the surrounding perivascular adipose tissue in CT imaging. Existing open-source CT segmentation tools enable segmentation of the aorta, but these tools are not designed to precisely capture aortic aneurysms (Wasserthal et al. 2023). To address the precise segmentation requirement for analysis of the perivascular adipose tissue, we developed a CNN for aneurysm segmentation using nnUNet in an iterative AI-assisted process (Isensee et al. 2021). This allowed us to demonstrate a significant difference in radiodensity of perivascular adipose tissue, both

interindividually between patients with aneurysms and healthy patients, as well as intraindividually between diseased and healthy sections of the aorta in affected patients.
Publicly available CNN-based segmentation tools, such as the TotalSegmentator, exhibit the limitation of being task-specific to the segmentation classes and modalities they were trained on using supervised learning (Wasserthal et al. 2023). Although the TotalSegmentator increasingly offers more segmentation classes for CT and has recently been expanded to MRI, developing methods that can perform segmentations on any modality in a zero-shot manner, like LLMs with text-based tasks, would be highly beneficial (D'Antonoli et al. 2024). This would significantly simplify or even eliminate the need for labor-intensive manual annotation in developing segmentation tools for specialized applications.
Promising recent advancements in this research area include the "Segment Anything Model" proposed by researchers from Meta AI, which is a transformer-based model that can precisely segment any structure in a semi-manual setting. A recent study verified the feasibility of zero-shot segmentations in medical images and further fine-tuning on medical images (Ma et al. 2024). Also, researchers from Microsoft introduced a transformer-based foundational model capable of segmenting 82 organs and objects across 9 different medical modalities, ranging from MRI to pathological image data (Zhao et al. 2024b). As with multimodal LLMs, this model allows for communication over segmentation adaptation via language.

4.4 Conclusion

In this thesis, the applicability of AI for automated analyses of radiological image and free-text radiological reports was investigated. The work provides insights into the potential of AI for advancing three categories of radiological workflows, namely the processing of free-text radiological reports, the artificial generation of virtual CE images, and the support of diagnostic and treatment decisions. It was demonstrated that previously unstructured clinical text databases could be unlocked by open LLMs that can be implemented in secured clinical infrastructures. Furthermore, it could be shown that image-based generative AI has the potential to benefit patient safety, the environment and the efficiency of cardiac MRI by replacing or reducing the need for contrast agent use. Lastly, it was demonstrated that AI can aid in identifying and extracting currently unused image features from radiological images that may be opportunistically derived and applied for precision- and personalized medicine in liver, oncologic or cardiovascular disease.

5 Summary

Traditionally, radiology has been at the forefront of digital medicine due to the digitalization of the two fundamental steps of the radiological workup in recent decades. This includes image acquisition and subsequent image reporting, which represents the visual assessment of the imaging by the radiologist alongside relevant clinical information to create a free-text document. Naturally, recent developments in AI promise high value for improving radiological workflows starting from task-specific, tool-based AI methods known prior to the release of ChatGPT in 2023. Since then, the public discourse increasingly shifted to an anticipation that task-agnostic, transformer-based LLMs will profoundly impact the day-to-day operations of numerous knowledge-based tasks, including radiology.

Therefore, this thesis aims to give an introduction into the AI developments that have led to the current promising AI tools of today. Furthermore, the potential of modern AI to improve diverse radiological workflows were investigated, namely the extraction of structured information from free-text report databases, the feasibility of contrast agent free ECV in cardiac MRI, the detection and further characterization of diseases and the use of AI for patient prognosis directly based on imaging or based on opportunistically derived scalar tissue quantifications.

The first two works of this thesis provide extensive insights on employing privacy-ensuring open transformers to extract content from free-text radiological reports and thereby to structure clinical document databases. Our work essentially provides a zero-shot and fine-tuning benchmark of BERT, 17 open LLMs, and 4 models of the closed GPT series of OpenAI for radiological documents using thousands of English and German language reports. Also, we compare LLM usage to conventional report analysis by simple rule-based systems. Key results were that zero-shot application of open LLMs is preferable to simple-rule based report content evaluation and comparable to GPT-4o, the current frontier model of OpenAI. Furthermore, the Mistral-Large model with 123 billion parameters created by the French startup Mistral AI showed higher capabilities than the larger Llama-3.1 model with 405 billion parameters for retrieving structured information from English and German reports. This indicates that the open models developed in the European Union can compete with prominent open and closed models from large US companies with respect to medical text analysis, and thereby could contribute to unlocking radiological databases for secondary use.

In the third work included in this thesis, we demonstrate the utility of transformer-based report content labels for training image-based AI for detecting findings in chest X-ray images of patients from the intensive care units of the University Hospital Bonn. We could identify shortcomings of some report-based labels that have led to content mismatch between report and image, including the not mentioning of findings due to low relevance for the current indication, reaching conclusion based on information that is not content of the radiological report e.g. clinical/laboratory parameters, or the mentioning of borderline image findings as definite, or simply by errors of the transformer. This indicates the need for more sophisticated labels based on patient-centred databases including multiple documents and information derived from all clinical databases contrary to labels solely based on single radiological reports.

The fourth work included in this thesis contributes to the current state of research on contrast agent free cardiac MRI using generative AI. We investigated the development of a GAN that transforms a native T1-map into a virtual CE T1 map using data from 1,000 patients from the University Hospitals of Bonn and Cologne. We anticipated challenges due to the requirement for precise image alignment of both T1-maps for training and due to patient-dependent factors influencing how rapidly the contrast agent is distributed and excreted. Although significant differences between contrast-free and conventional ECV were observed for amyloidosis, we could demonstrate comparable diagnostic utility for detecting both amyloidosis and myocarditis. This indicates the potential of generative AI to benefit patient safety, reduce environmental effects of contrast agent usage, and to decrease acquisition times in cardiac MRI by replacing or reducing the need for contrast agent use.

The fifth work indicates the potential of AI for supporting in disease characterizations. In this work, including T2-weighted MRI of 465 patients with liver cirrhosis, we demonstrated that an ImageNet pre-trained CNN encoder employed as frozen feature extractor, can be trained to differentiate between alcohol to non-alcohol related etiology with reasonable performance. This work highlights the potential of AI to indicate that image-based features with diagnostic value could exist for a given disease, even if they have simply not been discovered by radiological research yet. Therefore, reliable methods of explainable AI are required, which currently presents strong limitations for identifying individual reasons for model prediction.

The sixth and seventh work investigated the utility of AI for creating models with capabilities to predict the treatment success and patient prognosis by either AI-based extraction of scalar body composition measurements from CT of patients with pancreatic cancer for use in conventional CPH models, or by direct AI-based hazard estimation from abdominal CT of patients undergoing transcatheter aortic valve implantation. In the light of these works, we discuss downsides of the commonly applied hypothesis driven and p-value focused study design of clinical research, and the need for outcome-centered analysis to identify models with predictive capabilities that could be of true benefit for prognosis estimation and thereby for patient care.

In the last work included in the thesis, we demonstrated a significant difference in radiodensity of perivascular adipose tissue by enabling precise AI-based segmentations of abdominal aortic aneurysms. The results could contribute to identifying image markers with prognostic value based on perivascular inflammatory changes measured in CT. As a dedicated segmentation tool had to be developed to achieve analysis of perivascular adipose tissue of abdominal aortic aneurysms, we discuss the need for general segmentation methods that can perform segmentations on any modality in a zero-shot manner, like LLMs can perform numerous text-based tasks.

6 Overlap by shared authorships

This thesis includes eight original works, one as single first author, six with me in joint first authorship (four times in the first position, twice in the second position), and one with me in joint last authorship in second position. Generally, the interdisciplinary nature of the included works often required expertise in the field of data science/AI, as well as medical/radiological expertise, and involved very extensive areas of responsibility. Notably, the included works are not part of another cumulative habilitation dissertation. In the following sections the contributions of the authors with whom I share joint authorships will be outlined in detail.

The second work in this habilitation thesis was developed in collaboration with the Fraunhofer Institute for Intelligent Analysis and Information Systems. On the Fraunhofer side, Mr. David Biesner was responsible for conducting the project. In the following, the contributions of Mr. Biesner and myself are outlined in clear terms. Study concept and project idea: While the general study concept and project idea was mainly contributed by the last authors of the work, I contributed with the idea of investigating the use of BERT transformer with public and custom German pre-training by masked language modelling. Data acquisition: I was involved in data acquisition by preparing the free text reports exported by Dr. Wolfgang Block from the clinic systems for annotation, which was ultimately performed by medical assistants and Dr. Yannik Layer. From a technical standpoint, I implemented the annotation software (with the help of Mr. Benjamin Wulff), introduced it to the students, and provided technical support during the annotation process. Subsequently, I prepared the data for training the deep learning methods with the support of scripts provided by Mr. Biesner. Experiments: Mr. Biesner initially prepared Python scripts for training the transformers and classical deep learning models. I used these scripts as a basis for further optimizations and, with their aid, executed the experiments on hardware of the University Hospital Bonn, as the report data could not be shared with the Fraunhofer Institute due to data protection reasons. Interpretation and manuscript preparation: The evaluation of the results, creation of tables and figures, and the initial version of the manuscript were created by me with support from Mr. Biesner. The interpretation of the results and the critical review of the experiments and manuscript took place in regular meetings among all authors. Corresponding to Mr. Biesner's engagement, the authorship was divided accordingly.

The third project of this thesis is the follow-up to the previous work, therefore also conducted in collaboration with the Fraunhofer Institute for Intelligent Analysis and Information Systems,

with Ms. Helen Schneider responsible on the Fraunhofer side. In the following, the contributions of Ms. Schneider and myself are clearly described. Study concept and project idea: The study concept and project idea were mainly contributed by the last authors. Data acquisition: Both Ms. Schneider and I were involved in data acquisition. In addition to the contributions described above, I applied the deep learning methods developed in the previous work at the University Hospital Bonn to over 90,000 reports to structure them. Subsequently, I processed the data to enable identification of the corresponding X-ray images on the hospital systems, which was carried out using a tool developed by Dr. Alois Sprinkart and myself. The final export and transfer to the Fraunhofer Institute were performed by Mr. Sprinkart. On the Fraunhofer side, Ms. Schneider created scripts to prepare the data for training image-based methods, supported by Mr. Benjamin Wulff, Mr. David Biesner, and student assistants. Experiments: Ms. Helen Schneider, with support from me and David Biesner, implemented Python code for training of the deep learning methods. Regular meetings between Ms. Schneider and myself were held to plan experiments and discuss intermediate results. Interpretation and manuscript preparation: The evaluation of results, creation of tables and figures, and the first version of the manuscript were conducted by both Ms. Schneider and me, with accompanying regular meetings. The interpretation of results and critical review of experiments and the manuscript were conducted in regular meetings among all authors. Accordingly, authorship was shared reflecting Ms. Schneider's engagement.

The fourth work included in this thesis describes an interdisciplinary project that required both a data science/AI expertise and medical/radiological expert in leading roles. Below, the contributions of Dr. Leon Bischoff and myself are clearly described. Study concept and project idea: The project idea of contrast-free extracellular volume by AI was contributed by Dr. Julian Luetkens and Dr. Alois Sprinkart, while I designed the methodological concept. Data acquisition: Both Dr. Bischoff and I were involved in data acquisition. With support from Dr. Sprinkart and Dr. Block, Dr. Bischoff identified, exported, and performed an initial quality analysis of T1 images on the hospital systems, excluding poor-quality images. Subsequently, I prepared the data for deep learning training by writing Python code to perform necessary registrations between native T1 maps and T1 maps with contrast agents. I then subjected all images to a second visual quality control, excluding failed registrations. Lastly, I divided the data into training and test sets. Experiments: I conducted the training of the deep learning models and the quantitative evaluation of contrast-free images regarding diseases. Dr.

Bischoff visually evaluated the test dataset using his radiological expertise to assess whether focal lesions were represented. Interpretation and manuscript preparation: I led the creation of tables and figures and the first version of the manuscript, supported by Dr. Bischoff. The interpretation of results and critical review of experiments and the manuscript were conducted among all authors. Accordingly, authorship was shared reflecting Dr. Bischoff's engagement.

The fifth interdisciplinary work also required both data science/AI expertise and radiological/medical expertise in liver diseases. The contributions of Dr. Julian Luetkens and myself are described below. Study concept and project idea: The project idea of AI-based prediction of liver cirrhosis etiology was introduced by Dr. Luetkens and Dr. Sprinkart, while I designed the methodological concept. Data collection: Both Dr. Luetkens and I were involved in data collection. Dr. Luetkens was involved in identifying patients and supervising the extraction of data from clinical systems, executed by Dr. Narine Mesropyan. I subsequently prepared the data for deep learning training by applying AI for liver segmentation, exclusion criteria, and dividing the data into training and test sets. Experiments: I created Python code for training the deep learning methods and developed the models. I wrote Python code for applying explainable AI methods on the test data, which were visually evaluated by Dr. Anton Faron and Dr. Luetkens. I assisted Dr. Faron in statistical evaluation. Interpretation and manuscript creation: The preparation of tables, figures, and the initial manuscript version was led by Dr. Luetkens and Dr. Faron, with support by me. The interpretation of results and critical review of experiments and the manuscript were performed by all authors. Accordingly, authorship was shared reflecting Dr. Luetkens and my engagement.

The sixth work also required interdisciplinary knowledge of data science/statistics and medical expertise. The contributions of Mr. Christoph Kloth and myself are described below. Study concept and project idea: The project idea was introduced by the last authors of the study. Data collection: Both Mr. Kloth and I were involved in data collection. Mr. Kloth, under the supervision of Dr. Luetkens and with support from Dr. Milka Marinova, did the workup of the clinical cohort. I, with support from Dr. Sprinkart, exported the corresponding CT images from the clinical systems and performed tissue quantification using deep learning methods developed in our previous studies by me. Experiments: I executed the entire statistics of the study using self-created Python code. Interpretation and manuscript creation: The preparation of tables, figures, and the initial manuscript version was created by me with

support from Dr. Sprinkart and Mr. Kloth. The interpretation of results and critical review of experiments and the manuscript were performed by all authors. Accordingly, authorship was shared based on Mr. Kloth's engagement.

The seventh work shares joint last authorship with Dr. Sprinkart, with the following contributions. Study concept and project idea: I initially proposed the project idea of directly applying AI for survival prediction based on abdominal imaging, which was further developed by Dr. Sprinkart as supervisor of our working group. Ms. Maike Theis led the methodological conceptualization with support by me. Data collection: The data originated from previous studies, in which Dr. Sprinkart made central contributions on the technical side, and Dr. Luetkens contributed through processing the clinical cohort. Experiments: Ms. Theis created the Python code for training the AI models with my support and supervision. Ms. Theis performed the statistics. Interpretation and manuscript creation: Ms. Theis led the preparation of tables, figures, and the initial manuscript version. Subsequently, after a first critical revision and adaptation by Dr. Sprinkart and me, all authors reviewed the results and the manuscript. Accordingly, the last authorship was shared reflecting Dr. Sprinkart's and my engagement in the study.

Again, the last work required interdisciplinary knowledge of data science, AI, and statistics, as well as radiological expertise. The contributions of Dr. Daniel Ginzburg and myself are described below. Study concept and project idea: The project idea was introduced by the last authors of the study. Data collection: Both Dr. Ginzburg and I were involved in data collection. Dr. Ginzburg, under the supervision of Dr. Daniel Kütting, did the workup of the clinical cohort. Dr. Sprinkart, with my support, exported the corresponding CT images from the clinical systems. Dr. Ginzburg performed manual annotations of the aorta and adapted AI-based segmentations in 3D Slicer with technical support from me. I performed AI-based segmentations through the development of CNN models. Experiments: I developed the Python code for evaluating aortic segmentations regarding perivascular fat, including statistical tests. Interpretation and manuscript creation: The preparation of tables and the initial manuscript version was equally divided between Dr. Ginzburg (introduction and medical discussion) and me (methodology and results). The preparation of figures was conducted by me. The interpretation of results and critical review of experiments and the manuscript were performed by all authors. Accordingly, authorship was divided based on Dr. Ginzburg's engagement.

7 Bibliography

Agrawal, Pravesh; Antoniak, Szymon; Hanna, Emma Bou; Bout, Baptiste; Chaplot, Devendra; Chudnovsky, Jessica et al. (2024): Pixtral 12B. Available online at http://arxiv.org/pdf/2410.07073v2.

Babalola, B. T.; Yahya, W. B. (2020): Effects of Collinearity on Cox Proportional Hazard Model with Time Dependent Coefficients: A Simulation Study. In *jbe*. DOI: 10.18502/jbe.v5i2.2348.

Babyak, Michael A. (2004): What you see may not be what you get: a brief, nontechnical introduction to overfitting in regression-type models. In *Psychosomatic medicine* 66 (3), pp. 411–421. DOI: 10.1097/01.psy.0000127692.23278.a9.

Bhayana, Rajesh (2024): Chatbots and Large Language Models in Radiology: A Practical Primer for Clinical and Research Applications. In *Radiology* 310 (1), e232756. DOI: 10.1148/radiol.232756.

Bischoff, Leon M.; Peeters, Johannes M.; Weinhold, Leonie; Krausewitz, Philipp; Ellinger, Jörg; Katemann, Christoph et al. (2023): Deep Learning Super-Resolution Reconstruction for Fast and Motion-Robust T2-weighted Prostate MRI. In *Radiology* 308 (3), e230427. DOI: 10.1148/radiol.230427.

Brown, Tom B.; Mann, Benjamin; Ryder, Nick; Subbiah, Melanie; Kaplan, Jared; Dhariwal, Prafulla et al. (2020): Language Models are Few-Shot Learners. Available online at http://arxiv.org/pdf/2005.14165v4.

Chaudhari, Akshay S.; Fang, Zhongnan; Kogan, Feliks; Wood, Jeff; Stevens, Kathryn J.; Gibbons, Eric K. et al. (2018): Super-resolution musculoskeletal MRI using deep learning. In *Magnetic resonance in medicine* 80 (5), pp. 2139–2154. DOI: 10.1002/mrm.27178.

Chiang, Wei-Lin; Zheng, Lianmin; Sheng, Ying; Angelopoulos, Anastasios Nikolas; Li, Tianle; Li, Dacheng et al. (2024): Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference. Available online at http://arxiv.org/pdf/2403.04132v1.

Cox, D. R. (1972): Regression Models and Life-Tables. In *Journal of the Royal Statistical Society Series B: Statistical Methodology* 34 (2), pp. 187–202. DOI: 10.1111/j.2517-6161.1972.tb00899.x.

Cruz-Jentoft, Alfonso J.; Bahat, Gülistan; Bauer, Jürgen; Boirie, Yves; Bruyère, Olivier; Cederholm, Tommy et al. (2019): Sarcopenia: revised European consensus on definition and diagnosis. In *Age and ageing* 48 (1), pp. 16–31. DOI: 10.1093/ageing/afy169.

D'Souza, Rinaldo D.; Wang, Quanxin; Ji, Weiqing; Meier, Andrew M.; Kennedy, Henry; Knoblauch, Kenneth; Burkhalter, Andreas (2022): Hierarchical and nonhierarchical features of the mouse visual cortical network. In *Nat Commun* 13 (1). DOI: 10.1038/s41467-022-28035-y.

D'Antonoli, Tugba Akinci; Berger, Lucas K.; Indrakanti, Ashraya K.; Vishwanathan, Nathan; Weiß, Jakob; Jung, Matthias et al. (2024): TotalSegmentator MRI: Sequence-Independent Segmentation of 59 Anatomical Structures in MR images. Available online at http://arxiv.org/pdf/2405.19492v1.

Dechter, Rina (1986): Learning while searching in constraint-satisfaction-problems. In : AAAI'86: Proceedings of the Fifth AAAI National Conference on Artificial Intelligence. Philadelphia, Pennsylvania: AAAI Press, pp. 178–183.

Dekker, Helena M.; Stroomberg, Gerard J.; van der Molen, Aart J.; Prokop, Mathias (2024): Review of strategies to reduce the contamination of the water environment by gadolinium-based contrast agents. In *Insights into imaging* 15 (1), p. 62. DOI: 10.1186/s13244-024-01626-7.

Devlin, Jacob; Chang, Ming-Wei; Lee, Kenton; Toutanova, Kristina (2018): BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Available online at http://arxiv.org/pdf/1810.04805v2.

Dominic, S.; Das, R.; Whitley, D.; Anderson, C. (1991): Genetic reinforcement learning for neural networks. In : IJCNN-91-Seattle International Joint Conference on Neural Networks. IJCNN-91-Seattle International Joint Conference on Neural Networks. Seattle, WA, USA, 8-14 July 1991: IEEE, pp. 71–76.

Dosovitskiy, Alexey; Beyer, Lucas; Kolesnikov, Alexander; Weissenborn, Dirk; Zhai, Xiaohua; Unterthiner, Thomas et al. (2020): An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. Available online at http://arxiv.org/pdf/2010.11929v2.

Elkassem, Asser Abou; Smith, Andrew D. (2023): Potential Use Cases for ChatGPT in Radiology Reporting. In *AJR. American journal of roentgenology* 221 (3), pp. 373–376. DOI: 10.2214/AJR.23.29198.

Eloundou, Tyna; Manning, Sam; Mishkin, Pamela; Rock, Daniel (2023): GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models.

ESR (2018): ESR paper on structured reporting in radiology. In *Insights into imaging* 9 (1), pp. 1–7. DOI: 10.1007/s13244-017-0588-8.

Faron, Anton; Sprinkart, Alois M.; Kuetting, Daniel L. R.; Feisst, Andreas; Isaak, Alexander; Endler, Christoph et al. (2020): Body composition analysis using CT and MRI: intra-individual intermodal comparison of muscle mass and myosteatosis. In *Scientific reports* 10 (1), p. 11765. DOI: 10.1038/s41598-020-68797-3.

Ferber, Dyke; Wiest, Isabella C.; Wölflein, Georg; Ebert, Matthias P.; Beutel, Gernot; Eckardt, Jan-Niklas et al. (2024): GPT-4 for Information Retrieval and Comparison of Medical Oncology Guidelines. In *NEJM AI* 1 (6). DOI: 10.1056/AIcs2300235.

Ghassemi, Marzyeh; Oakden-Rayner, Luke; Beam, Andrew L. (2021): The false hope of current approaches to explainable artificial intelligence in health care. In *The Lancet. Digital health* 3 (11), e745-e750. DOI: 10.1016/S2589-7500(21)00208-9.

Gong, Chengyue; Di He; Tan, Xu; Qin, Tao; Wang, Liwei; Liu, Tie-Yan (2018): FRAGE: Frequency-Agnostic Word Representation. Available online at http://arxiv.org/pdf/1809.06858v2.

Goodfellow, Ian J.; Pouget-Abadie, Jean; Mirza, Mehdi; Xu, Bing; Warde-Farley, David; Ozair, Sherjil et al. (2014): Generative Adversarial Networks. Available online at http://arxiv.org/pdf/1406.2661v1.

Haaf, Philip; Garg, Pankaj; Messroghli, Daniel R.; Broadbent, David A.; Greenwood, John P.; Plein, Sven (2016): Cardiac T1 Mapping and Extracellular Volume (ECV) in clinical practice: a comprehensive review. In *Journal of cardiovascular magnetic resonance : official journal of*

the Society for Cardiovascular Magnetic Resonance 18 (1), p. 89. DOI: 10.1186/s12968-016-0308-4.

Haase, Robert; Pinetz, Thomas; Kobler, Erich; Paech, Daniel; Effland, Alexander; Radbruch, Alexander; Deike-Hofmann, Katerina (2023): Artificial Contrast: Deep Learning for Reducing Gadolinium-Based Contrast Agents in Neuroradiology. In *Investigative radiology* 58 (8), pp. 539–547. DOI: 10.1097/RLI.0000000000000963.

Hagiwara, Akifumi; Fujita, Shohei; Ohno, Yoshiharu; Aoki, Shigeki (2020): Variability and Standardization of Quantitative Imaging: Monoparametric to Multiparametric Quantification, Radiomics, and Artificial Intelligence. In *Investigative radiology* 55 (9), pp. 601–616. DOI: 10.1097/RLI.0000000000000666.

Hao, Yuexing; Holmes, Jason; Waddle, Mark; Yu, Nathan; Vickers, Kirstin; Preston, Heather et al. (2024): Outlining the Borders for LLM Applications in Patient Education: Developing an Expert-in-the-Loop LLM-Powered Chatbot for Prostate Cancer Patient Education. Available online at http://arxiv.org/pdf/2409.19100v1.

Herent, P.; Schmauch, B.; Jehanno, P.; Dehaene, O.; Saillard, C.; Balleyguier, C. et al. (2019): Detection and characterization of MRI breast lesions using deep learning. In *Diagnostic and Interventional Imaging* 100 (4), pp. 219–225. DOI: 10.1016/j.diii.2019.02.008.

Hosny, Ahmed; Parmar, Chintan; Quackenbush, John; Schwartz, Lawrence H.; Aerts, Hugo J. W. L. (2018): Artificial intelligence in radiology. In *Nature reviews. Cancer* 18 (8), pp. 500–510. DOI: 10.1038/s41568-018-0016-5.

Hu, Edward J.; Shen, Yelong; Wallis, Phillip; Allen-Zhu, Zeyuan; Li, Yuanzhi; Wang, Shean et al. (2021): LoRA: Low-Rank Adaptation of Large Language Models. Available online at http://arxiv.org/pdf/2106.09685v2.

Huang, Jen-Yuan; Wang, Haofan; Wang, Qixun; Bai, Xu; Ai, Hao; Xing, Peng; Huang, Jen-Tse (2024): InstantIR: Blind Image Restoration with Instant Generative Reference. Available online at http://arxiv.org/pdf/2410.06551v1.

Hubel, D. H.; Wiesel, T. N. (1962): Receptive fields of single neurones in the cat's striate cortex. In *The Journal of physiology* (160), pp. 106–154. DOI: 10.1113/jphysiol.2009.174151.

Isensee, Fabian; Jaeger, Paul F.; Kohl, Simon A. A.; Petersen, Jens; Maier-Hein, Klaus H. (2021): nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. In *Nature methods* 18 (2), pp. 203–211. DOI: 10.1038/s41592-020-01008-z.

Kashefi, Rojina; Barekatain, Leili; Sabokrou, Mohammad; Aghaeipoor, Fatemeh (2023): Explainability of Vision Transformers: A Comprehensive Review and New Perspectives. Available online at http://arxiv.org/pdf/2311.06786v1.

Kemp, Jennifer L.; Mahoney, Mary C.; Mathews, Vincent P.; Wintermark, Max; Yee, Judy; Brown, Stephen D. (2017): Patient-centered Radiology: Where Are We, Where Do We Want to Be, and How Do We Get There? In *Radiology* 285 (2), pp. 601–608. DOI: 10.1148/radiol.2017162056.

Kim, Dong Wook; Lee, Gaeun; Kim, So Yeon; Ahn, Geunhwi; Lee, June-Goo; Lee, Seung Soo et al. (2021): Deep learning-based algorithm to detect primary hepatic malignancy in multiphase CT of patients at high risk for HCC. In *European radiology* 31 (9), pp. 7047–7057. DOI: 10.1007/s00330-021-07803-2.

Kim, Hyungjin; Goo, Jin Mo; Lee, Kyung Hee; Kim, Young Tae; Park, Chang Min (2020): Preoperative CT-based Deep Learning Model for Predicting Disease-Free Survival in Patients with Lung Adenocarcinomas. In *Radiology* 296 (1), pp. 216–224. DOI: 10.1148/radiol.2020192764.

Kravchenko, Dmitrij; Isaak, Alexander; Mesropyan, Narine; Peeters, Johannes M.; Kuetting, Daniel; Pieper, Claus C. et al. (2024): Deep learning super-resolution reconstruction for fast and high-quality cine cardiovascular magnetic resonance. In *European radiology*. DOI: 10.1007/s00330-024-11145-0.

Kresevic, Simone; Giuffrè, Mauro; Ajcevic, Milos; Accardo, Agostino; Crocè, Lory S.; Shung, Dennis L. (2024): Optimization of hepatological clinical guidelines interpretation by large language models: a retrieval augmented generation-based framework. In *NPJ digital medicine* 7 (1), p. 102. DOI: 10.1038/s41746-024-01091-y.

Krizhevsky, Alex; Sutskever, Ilya; Hinton, Geoffrey E. (2012): ImageNet Classification with Deep Convolutional Neural Networks. In F. Pereira, C. J. Burges, L. Bottou, K. Q. Weinberger (Eds.): Advances in Neural Information Processing Systems, vol. 25: Curran Associates, Inc. Available online at https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

Langlotz, Curtis P. (2019): Will Artificial Intelligence Replace Radiologists? In *Radiology. Artificial intelligence* 1 (3), e190058. DOI: 10.1148/ryai.2019190058.

LeCun, Y. (1989): Generalization and network design strategies. In *Connections in Perspective*.

LeCun, Y.; Boser, B.; Denker, J. S.; Henderson, D.; Howard, R. E.; Hubbard, W.; Jackel, L. D. (1989): Backpropagation Applied to Handwritten Zip Code Recognition. In *Neural Computation* 1 (4), pp. 541–551. DOI: 10.1162/neco.1989.1.4.541.

LeCun, Yann; Bengio, Yoshua; Hinton, Geoffrey (2015): Deep learning. In *Nature* 521 (7553), pp. 436–444. DOI: 10.1038/nature14539.

LeCun, Yann; Bottou, L.; Bengio, Y.; Haffner, P. (1998): Gradient-based learning applied to document recognition. In *Proc. IEEE* 86 (11), pp. 2278–2324. DOI: 10.1109/5.726791.

Lee, Chanseo; Vogt, Kimon A.; Kumar, Sonu (2024): Prospects for AI clinical summarization to reduce the burden of patient chart review. In *Front. Digit. Health* 6, Article 1475092. DOI: 10.3389/fdgth.2024.1475092.

Lee, Dong Kyu (2016): Alternatives to P value: confidence interval and effect size. In *Korean Journal of Anesthesiology* 69 (6), pp. 555–562. DOI: 10.4097/kjae.2016.69.6.555.

Lenchik, Leon; Boutin, Robert D. (2018): Sarcopenia: Beyond Muscle Atrophy and into the New Frontiers of Opportunistic Imaging, Precision Medicine, and Machine Learning. In *Seminars in musculoskeletal radiology* 22 (3), pp. 307–322. DOI: 10.1055/s-0038-1641573.

Li, Binbin; Meng, Tianxin; Shi, Xiaoming; Zhai, Jie; Ruan, Tong (2023): MedDM:LLM-executable clinical guidance tree for clinical decision-making.

Liu, Ze; Lin, Yutong; Cao, Yue; Hu, Han; Wei, Yixuan; Zhang, Zheng et al. (2021): Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. Available online at http://arxiv.org/pdf/2103.14030v2.

Luetkens, Julian A.; Faron, Anton; Geissler, Helena L.; Al-Kassou, Baravan; Shamekhi, Jasmin; Stundl, Anja et al. (2020): Opportunistic Computed Tomography Imaging for the Assessment of Fatty Muscle Fraction Predicts Outcome in Patients Undergoing Transcatheter Aortic Valve Replacement. In *Circulation* 141 (3), pp. 234–236. DOI: 10.1161/CIRCULATIONAHA.119.042927.

Ma, Jun; He, Yuting; Li, Feifei; Han, Lin; You, Chenyu; Wang, Bo (2024): Segment anything in medical images. In *Nat Commun* 15 (1), p. 654. DOI: 10.1038/s41467-024-44824-z.

Mallio, Carlo A.; Radbruch, Alexander; Deike-Hofmann, Katerina; van der Molen, Aart J.; Dekkers, Ilona A.; Zaharchuk, Greg et al. (2023): Artificial Intelligence to Reduce or Eliminate the Need for Gadolinium-Based Contrast Agents in Brain and Cardiac MRI: A Literature Review. In *Investigative radiology* 58 (10), pp. 746–753. DOI: 10.1097/RLI.0000000000000983.

Mango, Victoria L.; Sun, Mary; Wynn, Ralph T.; Ha, Richard (2020): Should We Ignore, Follow, or Biopsy? Impact of Artificial Intelligence Decision Support on Breast Ultrasound Lesion Assessment. In *AJR. American journal of roentgenology* 214 (6), pp. 1445–1452. DOI: 10.2214/AJR.19.21872.

Martin-Carreras, Teresa; Cook, Tessa S.; Kahn, Charles E. (2019): Readability of radiology reports: implications for patient-centered care. In *Clinical imaging* 54, pp. 116–120. DOI: 10.1016/j.clinimag.2018.12.006.

Minsky, Marvin; Papert, Seymour (1988): Perceptions. An introduction to computational geometry. Exp. ed. The MIT Press: Cambridge.

Mukherjee, Pritam; Hou, Benjamin; Lanfredi, Ricardo B.; Summers, Ronald M. (2023): Feasibility of Using the Privacy-preserving Large Language Model Vicuna for Labeling Radiology Reports. In *Radiology* 309 (1), e231147. DOI: 10.1148/radiol.231147.

Nashef, Samer A. M.; Roques, François; Sharples, Linda D.; Nilsson, Johan; Smith, Christopher; Goldstone, Antony R.; Lockowandt, Ulf (2012): EuroSCORE II. In *European journal of cardio-thoracic surgery : official journal of the European Association for Cardio-thoracic Surgery* 41 (4), 734-44; discussion 744-5. DOI: 10.1093/ejcts/ezs043.

Nowak, Sebastian; Faron, Anton; Luetkens, Julian A.; Geißler, Helena L.; Praktiknjo, Michael; Block, Wolfgang et al. (2020): Fully Automated Segmentation of Connective Tissue Compartments for CT-Based Body Composition Analysis: A Deep Learning Approach. In *Investigative radiology* 55 (6), pp. 357–366. DOI: 10.1097/RLI.0000000000000647.

Nowak, Sebastian; Henkel, Andreas; Theis, Maike; Luetkens, Julian; Geiger, Sergej; Sprinkart, Alois M. et al. (2023): Deep learning for standardized, MRI-based quantification of subcutaneous and subfascial tissue volume for patients with lipedema and lymphedema. In *European radiology* 33 (2), pp. 884–892. DOI: 10.1007/s00330-022-09047-0.

Nowak, Sebastian; Kloth, Christoph; Theis, Maike; Marinova, Milka; Attenberger, Ulrike I.; Sprinkart, Alois M.; Luetkens, Julian A. (2024): Deep learning-based assessment of CT markers of sarcopenia and myosteatosis for outcome assessment in patients with advanced pancreatic cancer after high-intensity focused ultrasound treatment. In *European radiology* 34 (1), pp. 279–286. DOI: 10.1007/s00330-023-09974-6.

Nowak, Sebastian; Mesropyan, Narine; Faron, Anton; Block, Wolfgang; Reuter, Martin; Attenberger, Ulrike I. et al. (2021): Detection of liver cirrhosis in standard T2-weighted MRI

using deep transfer learning. In *European radiology* 31 (11), pp. 8807–8815. DOI: 10.1007/s00330-021-07858-1.

Nowak, Sebastian; Sprinkart, Alois M. (2024): Große Sprachmodelle von OpenAI, Google, Meta, X und Co. : Die Rolle von „closed“ und „open“ Modellen in der Radiologie. In *Radiologie (Heidelberg, Germany)* 64 (10), pp. 779–786. DOI: 10.1007/s00117-024-01327-8.

Nowak, Sebastian; Theis, Maike; Wichtmann, Barbara D.; Faron, Anton; Froelich, Matthias F.; Tollens, Fabian et al. (2022): End-to-end automated body composition analyses with integrated quality control for opportunistic assessment of sarcopenia in CT. In *European radiology* 32 (5), pp. 3142–3151. DOI: 10.1007/s00330-021-08313-x.

Oikonomou, Evangelos K.; Marwan, Mohamed; Desai, Milind Y.; Mancio, Jennifer; Alashi, Alaa; Hutt Centeno, Erika et al. (2018): Non-invasive detection of coronary inflammation using computed tomography and prediction of residual cardiovascular risk (the CRISP CT study): a post-hoc analysis of prospective outcome data. In *Lancet (London, England)* 392 (10151), pp. 929–939. DOI: 10.1016/S0140-6736(18)31114-0.

Oksuz, Ilkay; Clough, James; Bustin, Aurelien; Cruz, Gastao; Prieto, Claudia; Botnar, Rene et al.: Cardiac MR Motion Artefact Correction from K-space Using Deep Learning-Based Reconstruction. In : Machine Learning for Medical Image Reconstruction: First International Workshop, MLMIR 2018, Held in Conjunction with MICCAI 2018, vol. 11074, pp. 21–29.

Ouyang, Long; Wu, Jeff; Jiang, Xu; Almeida, Diogo; Wainwright, Carroll L.; Mishkin, Pamela et al. (2022): Training language models to follow instructions with human feedback. Available online at http://arxiv.org/pdf/2203.02155v1.

Pal, Arghya; Rathi, Yogesh (2022): A review and experimental evaluation of deep learning methods for MRI reconstruction. In *The journal of machine learning for biomedical imaging* 1.

Parmar, Gaurav; Park, Taesung; Narasimhan, Srinivasa; Zhu, Jun-Yan (2024): One-Step Image Translation with Text-to-Image Models. Available online at http://arxiv.org/pdf/2403.12036v1.

Pasquini, Luca; Napolitano, Antonio; Pignatelli, Matteo; Tagliente, Emanuela; Parrillo, Chiara; Nasta, Francesco et al. (2022): Synthetic Post-Contrast Imaging through Artificial Intelligence: Clinical Applications of Virtual and Augmented Contrast Media. In *Pharmaceutics* 14 (11). DOI: 10.3390/pharmaceutics14112378.

Pickhardt, Perry J. (2022): Value-added Opportunistic CT Screening: State of the Art. In *Radiology* 303 (2), pp. 241–254. DOI: 10.1148/radiol.211561.

Pierre, Kevin; Haneberg, Adam G.; Kwak, Sean; Peters, Keith R.; Hochhegger, Bruno; Sananmuang, Thiparom et al. (2023): Applications of Artificial Intelligence in the Radiology Roundtrip: Process Streamlining, Workflow Optimization, and Beyond. In *Seminars in roentgenology* 58 (2), pp. 158–169. DOI: 10.1053/j.ro.2023.02.003.

Praktiknjo, Michael; Zhou, Taotao; Krüsken, Maximiliane; Jacob, Torid; Sprinkart, Alois M.; Nowak, Sebastian et al. (2023): Myosteatosis independently predicts transplant-free survival in patients with primary sclerosing cholangitis. In *Digestive and liver disease : official journal of the Italian Society of Gastroenterology and the Italian Association for the Study of the Liver* 55 (11), pp. 1543–1547. DOI: 10.1016/j.dld.2023.08.037.

Radford, Alec; Kim, Jong Wook; Hallacy, Chris; Ramesh, Aditya; Goh, Gabriel; Agarwal, Sandhini et al. (2021): Learning Transferable Visual Models From Natural Language Supervision. Available online at http://arxiv.org/pdf/2103.00020v1.

Ramesh, Aditya; Dhariwal, Prafulla; Nichol, Alex; Chu, Casey; Chen, Mark (2022): Hierarchical Text-Conditional Image Generation with CLIP Latents. Available online at http://arxiv.org/pdf/2204.06125v1.

Rau, Alexander; Rau, Stephan; Zoeller, Daniela; Fink, Anna; Tran, Hien; Wilpert, Caroline et al. (2023): A Context-based Chatbot Surpasses Trained Radiologists and Generic ChatGPT in Following the ACR Appropriateness Guidelines. In *Radiology* 308 (1), e230970. DOI: 10.1148/radiol.230970.

Ronneberger, Olaf; Fischer, Philipp; Brox, Thomas (2015): U-Net: Convolutional Networks for Biomedical Image Segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells, Alejandro F. Frangi (Eds.): Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, vol. 9351. Cham: Springer International Publishing (Lecture Notes in Computer Science), pp. 234–241.

Rumelhart, David E.; Hinton, Geoffrey E.; Williams, Ronald J. (1986): Learning representations by back-propagating errors. In *Nature* 323 (6088), pp. 533–536. DOI: 10.1038/323533a0.

Sacoransky, Ethan; Kwan, Benjamin Y. M.; Soboleski, Donald (2024): ChatGPT and assistive AI in structured radiology reporting: A systematic review. In *Current problems in diagnostic radiology* 53 (6), pp. 728–737. DOI: 10.1067/j.cpradiol.2024.07.007.

Salam, Babak; Al Zaidi, Muntadher; Sprinkart, Alois M.; Nowak, Sebastian; Theis, Maike; Kuetting, Daniel et al. (2023): Opportunistic CT-derived analysis of fat and muscle tissue composition predicts mortality in patients with cardiogenic shock. In *Scientific reports* 13 (1), p. 22293. DOI: 10.1038/s41598-023-49454-x.

Salam, Babak; Kravchenko, Dmitrij; Nowak, Sebastian; Sprinkart, Alois M.; Weinhold, Leonie; Odenthal, Anna et al. (2024): Generative Pre-trained Transformer 4 makes cardiovascular magnetic resonance reports easy to understand. In *Journal of cardiovascular magnetic resonance : official journal of the Society for Cardiovascular Magnetic Resonance* 26 (1), p. 101035. DOI: 10.1016/j.jocmr.2024.101035.

Samani, Zahra Riahi; Parker, Drew; Wolf, Ronald; Hodges, Wes; Brem, Steven; Verma, Ragini (2021): Distinct tumor signatures using deep learning-based characterization of the peritumoral microenvironment in glioblastomas and brain metastases. In *Scientific reports* 11 (1). DOI: 10.1038/s41598-021-93804-6.

Schober, Patrick; Bossers, Sebastiaan M.; Schwarte, Lothar A. (2018): Statistical Significance Versus Clinical Importance of Observed Effect Sizes: What Do P Values and Confidence Intervals Really Represent? In *Anesthesia and analgesia* 126 (3), pp. 1068–1072. DOI: 10.1213/ANE.0000000000002798.

Schwartz, Lawrence H.; Panicek, David M.; Berk, Alexandra R.; Li, Yuelin; Hricak, Hedvig (2011): Improving communication of diagnostic radiology findings through structured reporting. In *Radiology* 260 (1), pp. 174–181. DOI: 10.1148/radiol.11101913.

Shaul, Roy; David, Itamar; Shitrit, Ohad; Riklin Raviv, Tammy (2020): Subsampled brain MRI reconstruction by generative adversarial neural networks. In *Medical image analysis* 65, p. 101747. DOI: 10.1016/j.media.2020.101747.

Silver, David; Huang, Aja; Maddison, Chris J.; Guez, Arthur; Sifre, Laurent; van den Driessche, George et al. (2016): Mastering the game of Go with deep neural networks and tree search. In *Nature* 529 (7587), pp. 484–489. DOI: 10.1038/nature16961.

Sloan, Phillip; Clatworthy, Philip; Simpson, Edwin; Mirmehdi, Majid (2024): Automated Radiology Report Generation: A Review of Recent Advances. In *IEEE reviews in biomedical engineering* PP. DOI: 10.1109/RBME.2024.3408456.

Starekova, Jitka; Pirasteh, Ali; Reeder, Scott B. (2024): Update on Gadolinium-Based Contrast Agent Safety, From the AJR Special Series on Contrast Media. In *AJR. American journal of roentgenology* 223 (3), e2330036. DOI: 10.2214/AJR.23.30036.

Suchting, Robert; Hébert, Emily T.; Ma, Ping; Kendzor, Darla E.; Businelle, Michael S. (2019): Using Elastic Net Penalized Cox Proportional Hazards Regression to Identify Predictors of Imminent Smoking Lapse. In *Nicotine & tobacco research : official journal of the Society for Research on Nicotine and Tobacco* 21 (2), pp. 173–179. DOI: 10.1093/ntr/ntx201.

Tadavarthi, Yasasvi; Vey, Brianna; Krupinski, Elizabeth; Prater, Adam; Gichoya, Judy; Safdar, Nabile; Trivedi, Hari (2020): The State of Radiology AI: Considerations for Purchase Decisions and Current Market Offerings. In *Radiology. Artificial intelligence* 2 (6), e200004. DOI: 10.1148/ryai.2020200004.

Tam, Bernard; Chng, Sue Inn; Quiroz Aguilera, Juan (2024): Leveraging Large Language Models (LLMs) to create a chatbot assistant for the retrieval of medication information. With assistance of Ritu Agarwal, Gabriel Brat, Guodong Gordon Gao, Jeffrey McCullough, UNSW Sydney.

Tappert, Charles C. (2019): Who Is the Father of Deep Learning? In : 2019 International Conference on Computational Science and Computational Intelligence (CSCI). 2019 International Conference on Computational Science and Computational Intelligence (CSCI). Las Vegas, NV, USA, 05.12.2019 - 07.12.2019: IEEE, pp. 343–348.

van der Malsburg, C. (1986): Frank Rosenblatt: Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms. In Günther Palm, Ad Aertsen (Eds.): Brain Theory. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 245–248.

Vaswani, Ashish; Shazeer, Noam; Parmar, Niki; Uszkoreit, Jakob; Jones, Llion; Gomez, Aidan N. et al. (2017): Attention Is All You Need. Available online at http://arxiv.org/pdf/1706.03762v7.

Wasserthal, Jakob; Breit, Hanns-Christian; Meyer, Manfred T.; Pradella, Maurice; Hinck, Daniel; Sauter, Alexander W. et al. (2023): TotalSegmentator: Robust Segmentation of 104 Anatomic Structures in CT Images. In *Radiology. Artificial intelligence* 5 (5), e230024. DOI: 10.1148/ryai.230024.

Wei, Kimberly; Fritz, Christian; Rajasekaran, Karthik (2024): Answering head and neck cancer questions: An assessment of ChatGPT responses. In *American journal of otolaryngology* 45 (1), p. 104085. DOI: 10.1016/j.amjoto.2023.104085.

Wu, Chaoyi; Lei, Jiayu; Zheng, Qiaoyu; Zhao, Weike; Lin, Weixiong; Zhang, Xiaoman et al. (2023): Can GPT-4V(ision) Serve Medical Applications? Case Studies on GPT-4V for Multimodal Medical Diagnosis. Available online at http://arxiv.org/pdf/2310.09909v3.

Xu, Yiwen; Hosny, Ahmed; Zeleznik, Roman; Parmar, Chintan; Coroller, Thibaud; Franco, Idalid et al. (2019): Deep Learning Predicts Lung Cancer Treatment Response from Serial Medical

Imaging. In *Clinical cancer research : an official journal of the American Association for Cancer Research* 25 (11), pp. 3266–3275. DOI: 10.1158/1078-0432.CCR-18-2495.

Xue, Xiaonan; Kim, Mimi Y.; Shore, Roy E. (2007): Cox regression analysis in presence of collinearity: an application to assessment of health risks associated with occupational radiation exposure. In *Lifetime data analysis* 13 (3), pp. 333–350. DOI: 10.1007/s10985-007-9045-1.

Yasaka, Koichiro; Akai, Hiroyuki; Abe, Osamu; Kiryu, Shigeru (2018): Deep Learning with Convolutional Neural Network for Differentiation of Liver Masses at Dynamic Contrast-enhanced CT: A Preliminary Study. In *Radiology* 286 (3), pp. 887–896. DOI: 10.1148/radiol.2017170706.

Yi, Paul Hyunsoo; Golden, Sean Kenney; Harringa, John B.; Kliewer, Mark A. (2019): Readability of Lumbar Spine MRI Reports: Will Patients Understand? In *AJR. American journal of roentgenology* 212 (3), pp. 602–606. DOI: 10.2214/AJR.18.20197.

Zeiler, Matthew D.; Fergus, Rob (2014): Visualizing and Understanding Convolutional Networks. In David Fleet, Tomas Pajdla, Bernt Schiele, Tinne Tuytelaars (Eds.): Computer Vision – ECCV 2014, vol. 8689. Cham: Springer International Publishing (Lecture Notes in Computer Science), pp. 818–833.

Zhang, Qiang; Burrage, Matthew K.; Lukaschuk, Elena; Shanmuganathan, Mayooran; Popescu, Iulia A.; Nikolaidou, Chrysovalantou et al. (2021): Toward Replacing Late Gadolinium Enhancement With Artificial Intelligence Virtual Native Enhancement for Gadolinium-Free Cardiovascular Magnetic Resonance Tissue Characterization in Hypertrophic Cardiomyopathy. In *Circulation* 144 (8), pp. 589–599. DOI: 10.1161/CIRCULATIONAHA.121.054432.

Zhang, Qiang; Burrage, Matthew K.; Shanmuganathan, Mayooran; Gonzales, Ricardo A.; Lukaschuk, Elena; Thomas, Katharine E. et al. (2022): Artificial Intelligence for Contrast-Free MRI: Scar Assessment in Myocardial Infarction Using Deep Learning-Based Virtual Native Enhancement. In *Circulation* 146 (20), pp. 1492–1503. DOI: 10.1161/CIRCULATIONAHA.122.060137.

Zhao, Jiawei; Zhang, Zhenyu; Chen, Beidi; Wang, Zhangyang; Anandkumar, Anima; Tian, Yuandong (2024a): GaLore: Memory-Efficient LLM Training by Gradient Low-Rank Projection. Available online at http://arxiv.org/pdf/2403.03507v2.

Zhao, Theodore; Gu, Yu; Yang, Jianwei; Usuyama, Naoto; Lee, Ho Hin; Naumann, Tristan et al. (2024b): BiomedParse: a biomedical foundation model for image parsing of everything everywhere all at once. Available online at http://arxiv.org/pdf/2405.12971v3.

8 Acknowledgments

I would like to express my deepest gratitude to PD Dr. med. Julian Alexander Luetkens, who enabled me to conduct this thesis under his supervision at the Department of Diagnostic and Interventional Radiology and who was always supportive during this time.

Also, I want to express my deepest appreciation especially to PD Dr.-Ing. Alois Martin Sprinkart, who has been a mentor to me since supervising my Bachelor thesis, investing in me and advocating for me. I also want to express my sincere gratitude to PD Dr. rer. nat. Wolfgang Block for his support in personal and regulatory problems along all years. I want to deeply thank Mr. Benjamin Wulff for his technical support with respect to the high-performance systems required to train the language models investigated in this thesis. Many thanks also to my colleagues Maike Theis and Laura Garajová, with whom I had good cooperation, but also a lot of fun during the time of writing this thesis. Also many thanks to all of my other colleagues and co-authors at the University Hospital Bonn.

I want to thank my parents, who have encouraged me throughout my life, supported me in good and bad times and thus played a major role in my personal and professional development. I would like to thank my big brother, who has always been and will always be a role model for me. Lastly, I would like to thank my girlfriend, who supported me during the period of this thesis and patiently endured one or the other fanatical raving about current deep learning methods.