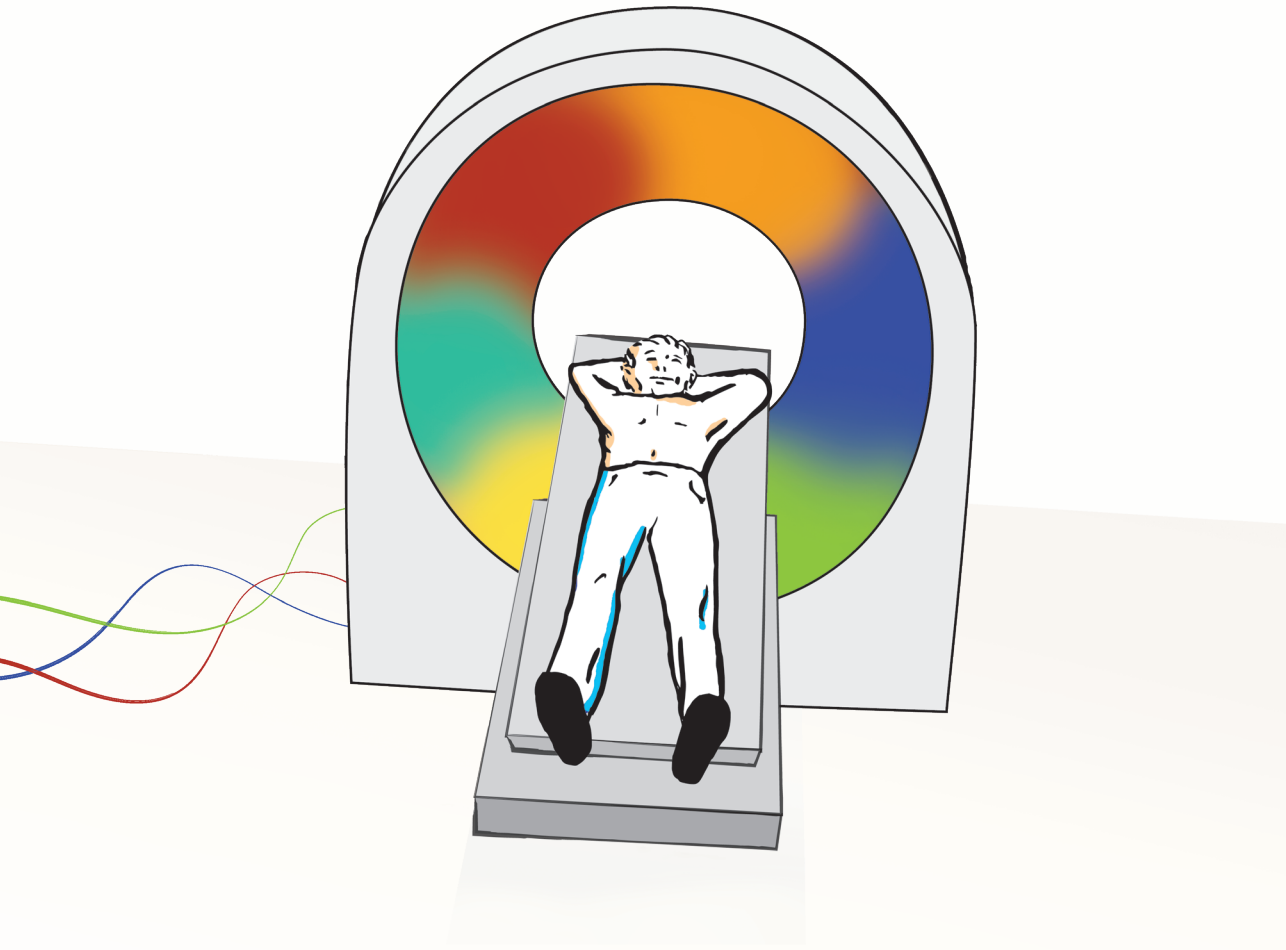
Machine learning based quantification of extrapulmonary diseases in chest CT



Nikolas Leßmann

The publication of this thesis was financially supported by Thirona.

Financial support by the Dutch Heart Foundation for the publication of this thesis is gratefully acknowledged.

ISBN: 978-94-6323-607-2

Copyright © N. Leßmann 2019. All rights reserved.

No part of this publication may be reproduced or transmitted in any form by any means without prior permission from the copyright owner. The copyright of the articles that have been published has been transferred to the respective journals.

Cover design by Remco Wetzels. Printed by Gildeprint.

Machine learning based quantification of extrapulmonary diseases in chest CT

Kwantificering van extrapulmonale ziekten in
CT beelden van de longen op basis van machine learning
(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de Universiteit Utrecht
op gezag van de rector magnificus, prof. dr. H. R. B. M. Kummeling,
ingevolge het besluit van het college voor promoties
in het openbaar te verdedigen op dinsdag 25 juni 2019 des middags te 2.30 uur

door

Nikolas Leßmann
geboren op 5 juli 1988
te Münster, Duitsland

Promotoren:	Prof. dr. ir. M. A. Viergever
	Prof. dr. B. van Ginneken
	Prof. dr. P. A. de Jong
Copromotor:	Dr. I. Išgum

Contents

CHAPTER 1
Introduction 7

CHAPTER 2
Automatic arterial calcification scoring in low-dose chest CT 13

CHAPTER 3
Sex differences in arterial calcification in heavy smokers 37

CHAPTER 4
Automatic vertebra segmentation and identification 55

CHAPTER 5
Vertebra partitioning with thin-plate splines steered by neural networks 83

CHAPTER 6
Summary and discussion 107

Nederlandse samenvatting 115
Bibliography 119
Acknowledgments 139
Publications 143
Biography 147

1

CHAPTER 1

Introduction

1.1 Lung cancer screening

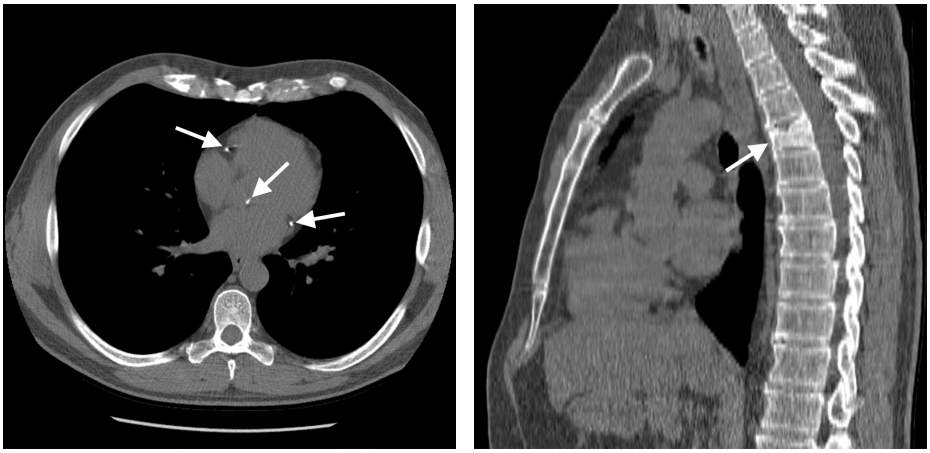
Lung cancer is responsible for more deaths than any other type of cancer. In 2018, lung cancer was responsible for 18.4 % of all cancer deaths worldwide, followed by colorectal cancer which was responsible for only half as many deaths.[1] Because treatment options and prognosis are a lot more favorable when lung cancer is detected in an early stage,[2] substantial efforts have been made toward finding an appropriate screening test. Randomized trials were not able to show a benefit of screening with chest radiography and/or sputum cytology.[3] However, the results of the National Lung Screening Trial in 2011 were a breakthrough. This trial found a 20 % reduction in lung cancer mortality in a high risk population of heavy smokers who received screening with low-dose chest CT.[4] While similar trials are still underway to verify these findings,[5] lung cancer screening with low-dose chest CT has already become part of official guidelines in several countries, including the United States and China, and screening programs are being implemented.[6–8]

Lung cancer is predominantly caused by cigarette smoking.[9] In 2000, smoking accounted for approximately 71 % of lung cancer deaths worldwide and even around 90 % in the developed countries where environmental factors such as air pollution play a less important role.[10,11] An extensive history of cigarette smoking is therefore one of the inclusion criteria for participation in lung cancer screening programs and research trials. While lung cancer is the primary target of these programs, the screened population presents also with high prevalence of a whole array of other chronic diseases that are caused by heavy smoking, such as chronic obstructive pulmonary disease (COPD)[12] and atherosclerosis.[13] Many of these diseases can be detected and quantified in chest CT as well. Lung cancer screening with low-dose chest CT therefore provides opportunities for concurrent evaluation of multiple diseases in addition to lung cancer using the existing screening images. This thesis focuses on quantification of extrapulmonary diseases in chest CT scans, namely cardiovascular disease and osteoporosis.

1.2 Quantification of extrapulmonary diseases in chest CT

1.2.1 Cardiovascular disease

While lung cancer is the leading cause of cancer death worldwide, cardiovascular disease (CVD) is the overall leading cause of death and responsible for more deaths than all types of cancer combined.[14] Cardiovascular diseases are diseases that affect the heart and the blood vessels, such as myocardial infarction and stroke. Smoking is a major risk factor for these diseases, and both fatal and non-fatal cardiovascular events are common in heavy smokers.[4,15] Cardiovascular risk assessment in conjunction with lung cancer screening has the potential to improve the identification of heavy smokers who can benefit from preventive measures, such as treatment with aspirin or statins.



(a) Arterial calcification, axial view (b) Vertebral fracture, sagittal view

Figure 1.1: Examples of low-dose chest CT scans. Even though the lungs make up most of the three-dimensional image volumes, many other anatomical structures are visualized as well, including the heart and the thoracic sections of aorta and spine.

The risk of cardiovascular events can be assessed with CT imaging by quantifying the amount of arterial calcification, most importantly of the coronary arteries[16] and the aorta.[17] An example of arterial calcification visible on a low-dose chest CT scan is shown in Figure 1.1a.

1.2.2 Osteoporosis

Smoking is also a risk factor for osteoporosis and osteoporotic fractures.[18] These fractures manifest frequently in the spine as vertebral compression fractures and can be detected on chest CT images if they affect the thoracic vertebrae.[19] Furthermore, chest CT can be used to quantify other markers for osteoporosis and fracture risk, such as bone mineral density.[20] These markers could prove useful for personalized osteoporotic fracture risk assessment. Effective drug therapies are available for treating osteoporosis and reducing fractures, such as bisphosphonates or raloxifene.[21] Heavy smokers could therefore benefit from assessment of their risk for osteoporosis in conjunction with lung cancer screening, especially if this could help identify the disease before it causes fractures. An example of a vertebral compression fracture visible on a low-dose chest CT scan is shown in Figure 1.1b.

1.3 Automatic quantification with machine learning

Although theoretically appealing, there are some practical limitations to using lung cancer screening CT scans to screen not only for lung cancer but also for other chronic

diseases. With millions of current and former heavy smokers meeting the eligibility criteria for lung cancer screening,[22] the additional efforts required to evaluate these diseases in addition to lung cancer would put a considerable extra burden on radiologists. Exact quantification of the amount of arterial calcification and of shape or density of the vertebrae are time-consuming and tedious tasks, but may be more beneficial for precise and objective risk stratification than simple visual estimates.

Unfortunately, there is currently no strong evidence that screening for cardiovascular disease and osteoporosis as part of lung cancer screening provides benefits. Screening trials such as the National Lung Screening Trial (NLST) with ~54 000 participants[23] or the NELSON trial with ~16 000 participants,[24] and studies such as the Genetic Epidemiology of COPD study (COPDGene) with ~10 000 participants,[25] collected large datasets with chest CT images and associated demographic and follow-up data. These datasets can potentially provide insights into prevalence, development and progression of extrapulmonary diseases and in the predictive value of disease markers quantified on CT images. However, manual analysis of these large datasets is hardly feasible.

These shortcomings are addressed in this thesis by developing automatic methods for quantification of cardiovascular disease and osteoporosis markers in chest CT images with the goal of enabling both routine analysis of screening CT scans as well as retrospective analysis of large existing datasets for clinical research. The developed methods are based on machine learning techniques, namely deep learning, and in particular convolutional neural networks.

1.4 Outline of the thesis

The first half of the thesis, the following two chapters, addresses quantification of arterial calcification for cardiovascular risk prediction in heavy smokers:

CHAPTER 2 describes a method for automatic quantification of calcifications of the coronary arteries, the thoracic aorta and the heart valves in CT scans. The method uses two consecutive convolutional neural networks to detect and label calcifications and is evaluated with more than 500 manually annotated low-dose chest CT scans.

CHAPTER 3 describes an analysis of the prevalence and severity of arterial calcification visible on lung cancer screening CT scans. The automatic quantification method from the previous chapter was used to quantify arterial calcifications in a large set of low-dose chest CT scans.

The second half of the thesis presents steps toward automatic quantification of vertebral characteristics:

CHAPTER 4 describes a method for segmentation and anatomical labeling of the vertebrae

in low-dose chest CT scans, but also other types of CT and MR scans. The method relies on a fully convolutional neural network with an integrated memory component to search for the spine and iteratively segment the vertebrae one after the other.

CHAPTER 5 describes a method that partitions the segmentation mask of a vertebra, obtained with the segmentation method from the previous chapter, into the vertebral body and the posterior elements. A convolutional neural network analyzes the segmentation mask and generates a smooth surface representing the boundary between the two parts of the vertebra.

Finally, **CHAPTER 6** summarizes the presented methods and findings and discusses the results as well as future directions.

2

CHAPTER 2

Automatic arterial calcification scoring in low-dose chest CT

BASED ON: N. Lessmann, B. van Ginneken, M. Zreik, P. A. de Jong, B. D. de Vos, M. A. Viergever, and I. Išgum. "Automatic calcium scoring in low-dose chest CT using deep neural networks with dilated convolutions", *IEEE Transactions on Medical Imaging*, vol. 37 (2018), pp. 615–625.

Abstract

Heavy smokers undergoing screening with low-dose chest CT are affected by cardiovascular disease as much as by lung cancer. Low-dose chest CT scans acquired in screening enable quantification of atherosclerotic calcifications and thus enable identification of subjects at increased cardiovascular risk. This chapter presents a method for automatic detection of coronary artery, thoracic aorta and cardiac valve calcifications in low-dose chest CT using two consecutive convolutional neural networks. The first network identifies and labels potential calcifications according to their anatomical location and the second network identifies true calcifications among the detected candidates. This method was trained and evaluated on a set of 1744 CT scans from the National Lung Screening Trial. To determine whether any reconstruction or only images reconstructed with soft tissue filters can be used for calcification detection, we evaluated the method on soft and medium/sharp filter reconstructions separately. On soft filter reconstructions, the method achieved F_1 scores of 0.89, 0.89, 0.67, and 0.55 for coronary artery, thoracic aorta, aortic valve and mitral valve calcifications, respectively. On sharp filter reconstructions, the F_1 scores were 0.84, 0.81, 0.64, and 0.66, respectively. Linearly weighted kappa coefficients for risk category assignment based on per subject coronary artery calcium were 0.91 and 0.90 for soft and sharp filter reconstructions, respectively. These results demonstrate that the presented method enables reliable automatic cardiovascular risk assessment in all low-dose chest CT scans acquired for lung cancer screening.

2.1 Introduction

Screening with low-dose chest CT has been found effective in reducing mortality from lung cancer in current or former heavy smokers.[4] However, smoking is not only a major risk factor for lung cancer, but also for cardiovascular disease (CVD).[26] The presence of CVD can be detected in CT scans by measuring the amount of coronary artery calcification (CAC), a strong and independent predictor of cardiovascular events. CAC is usually quantified in dedicated cardiac CT images and expressed as a calcium score.[27] Recent studies have shown that calcium scores are also able to predict cardiovascular events if quantified in low-dose chest CT for lung cancer screening.[15,28,29] Calcium scoring could therefore complement lung cancer screening programs to help identify subjects at elevated cardiovascular risk without the need for further imaging.[30,31] However, manual calcium scoring in addition to lung screening would impose a considerable extra burden on screening programs due to the large number of subjects, the high average calcium burden in a high risk population, the suboptimal image quality of low-dose screening scans, and the lack of ECG synchronization leading to cardiac motion artifacts. Automatic calcium scoring could therefore be a viable alternative that would enable routine cardiovascular risk prediction from low-dose chest CT scans.

Automatic methods for CAC scoring have been developed mostly for dedicated non-contrast enhanced cardiac CT[32–38] or cardiac CT angiography (CTA) scans.[39–45] Only few methods have been developed specifically for coronary calcium scoring in chest CT.[46–49]

In non-contrast cardiac CT scans, the coronary arteries are visible only when calcified or embedded in fat. Automatic scoring methods therefore typically rely on segmentation or rough localization of larger structures such as the heart and the aorta to derive a region of interest[35,36] or to derive spatial features for classification of candidate lesions using machine learning.[32–34] Other methods that use machine learning derive spatial features from segmentations of the coronary arteries, which are obtained by registration of the non-contrast scan with a CTA-based atlas.[37,38] Commonly used features besides spatial features are texture, lesion volume and shape. Spatial features are consistently reported to be most important.

In cardiac CTA scans, the coronary arteries are well visible due to the arterial contrast enhancement. Automatic scoring methods therefore typically perform a segmentation of the coronary artery tree. The segmentation is used to detect calcifications by searching for strong intensity gradients along the segmented vessel[39–42] because calcifications are typically brighter than the contrast enhanced lumen of the vessel. In a similar approach, Mittal et al.[43] employ a classifier to detect calcifications based on features that describe the texture along the vessel. In contrast to such approaches, Wolterink et al.[38] proposed a method without prior segmentation of the coronary arteries. The method uses a convolutional neural network (CNN) in combination with simple spatial features based on image coordinates to classify candidate voxels in the

image. To reduce the false positive rate of the CNN, connected groups of voxels that were detected by the CNN are reclassified by a random forest classifier. In their later publication,[45] a second CNN replaces this step.

In non-contrast chest CT scans, segmentation of the coronary artery tree is not feasible due to the lack of contrast enhancement and due to cardiac motion artifacts caused by the lack of ECG synchronization. Automatic scoring methods therefore typically rely on other means to identify a region of interest, similar to methods for CAC detection in non-contrast cardiac CT. Išgum et al.[46] obtain spatial features from a coronary calcium probability atlas and use these in combination with volume and texture features in a multi-classifier approach. Xie et al.[47] and González et al.[49] segment or roughly localize the heart and identify coronary calcifications in the detected region of interest based on decision rules. In our preliminary work, we proposed to use a CNN to classify candidate voxels within a bounding box around the heart.[48]

In addition to coronary calcifications, thoracic aorta calcifications (TAC) and calcifications of the cardiac valves have been related to cardiovascular risk.[17,50,51] For calcium scoring in the thoracic aorta in chest CT, few automatic methods have been published. All methods first perform a segmentation of the aorta followed by either rule-based calcification detection using auxiliary segmentations of trachea and spine[52,53] or calcification detection based on machine learning.[54] The machine learning approach uses kNN classifiers with features similar to those used by methods for coronary calcium detection: various spatial features derived from the segmentation of the aorta, volume of the potential calcification and texture features. For detection and quantification of cardiac valve calcifications, automatic methods have not been published.

We propose an automatic system for concurrent detection of CAC, TAC and cardiac valve calcifications in low-dose chest CT. These calcifications likely show different aspects of atherosclerotic disease and their quantities can potentially complement each other in detecting the presence of CVD and in predicting cardiovascular events. In contrast to simple combination of the output of multiple systems, e.g., one system for CAC and another for TAC detection, a single method for concurrent detection avoids ambiguous results.

Furthermore, we propose to label each voxel separately according to the affected vessel rather than labeling calcified lesions, i.e., connected voxels above a certain intensity value. Lesion labeling is standardly performed in clinically used commercial software and most previous automatic methods. However, voxel labeling allows separation of single lesions that extend to multiple vascular beds, e.g., in the aorta and a coronary artery. This is important because calcifications in different arteries carry different prognostic value.[17]

Next, we propose to use a CNN to directly identify potential calcifications within the entire image, without the need to explicitly segment or localize anatomical structures. The majority of methods in the literature for CAC or TAC detection rely on segmentation to restrict the region of interest or to infer spatial features as these have

Table 2.1: Overview of scanner models and reconstruction filters

Vendor	Models	Reconstruction filters
General Electric (GE)	LightSpeed 16 LightSpeed Pro 16 LightSpeed Ultra LightSpeed Plus LightSpeed QX/i Discovery QX/i HiSpeed QX/i	Standard (soft tissue) Bone, Lung (medium/sharp)
Siemens	Sensation 4 Sensation 16 Volume Zoom	B30f (soft tissue) B50f, B80f (medium/sharp)
Philips	MX8000 MX8000 IDT	C (soft tissue) D (medium/sharp)
Toshiba	Aquilion	FC10 (soft tissue) FC51 (medium/sharp)

been found important for accurate calcium detection. Instead, context information is not provided by segmentations, but rather the CNN has to be able to infer context information directly from the image. We therefore use a CNN with a particularly large receptive field, which is achieved by using an architecture based on dilated convolutions. This large receptive field furthermore enables the network to label candidates based on their anatomical location. Similar to,[45] we subsequently employ a second CNN to identify calcifications among the candidates identified and labeled by the first CNN. The main contribution is therefore a network architecture tailored specifically to the problem of calcium detection in low-dose chest CT scans.

Finally, we evaluated our method on a large and diverse dataset from the to date largest lung cancer screening trial with low-dose chest CT. Most other methods for automatic CAC or TAC scoring have been evaluated only on relatively small and homogeneous datasets. However, a method that is applied clinically will face images acquired with a multitude of scanner models and reconstructed with a wide range of reconstruction algorithms. In this work, scans were therefore selected such that a wide variety of screening sites, scanner models and reconstruction algorithms are present.

2.2 Dataset

For training and evaluation, we used low-dose chest CT scans acquired in the National Lung Screening Trial (NLST). The NLST was a large lung cancer screening trial in the United States that enrolled 53 454 current or former heavy smokers aged 55 to 74.[4] To develop and evaluate the proposed automatic calcium scoring method on a diverse dataset, we selected 1744 scans from 6000 available baseline scans by randomly

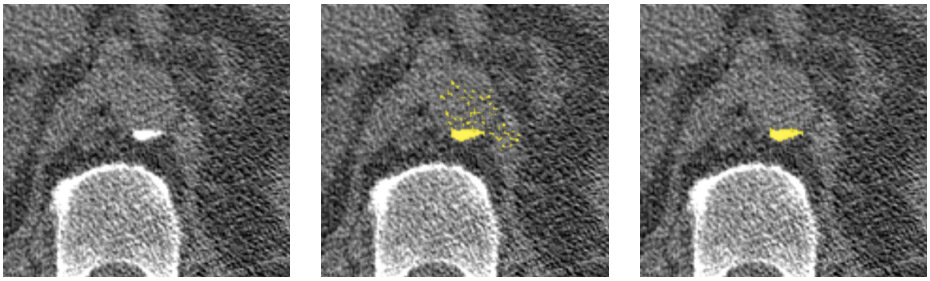


Figure 2.1: Segmentation of a TAC lesion in a noisy scan (left) using 3D region growing (center) and fully manual segmentation (right). Region growing here leads to segmentation of a large amount of noise together with the calcium.

sampling from the scans acquired with the 25 most common imaging settings with respect to scanner model and reconstruction algorithm. Specifically, 100 scans were selected for each of the 10 most common settings, and up to 50 for each of the 15 next most common settings.

The selected scans were acquired in 31 medical centers on 13 different scanner models from four major CT scanner vendors (Table 2.1). The scans were acquired at breath-hold after inspiration in helical scanning mode without contrast enhancement and without ECG synchronization. Tube voltage was set to 120 kVp, or 140 kVp for large subjects (5 %). In-plane resolution ranged from 0.49 mm to 0.98 mm, slice thickness from 1 mm to 3 mm and slice spacing from 0.6 mm to 3.0 mm. Since calcium scoring is typically performed on 3 mm thick slices, we reconstructed 3 mm thick axial slices with 1.5 mm slice spacing from all scans.

To establish a reference standard, calcifications were manually labeled in all scans. Scans were distributed among four trained observers and one radiologist with extensive experience in calcium scoring. To measure interobserver agreement, a subset of 100 scans (four scans from each of the 25 different scanner models and reconstruction algorithms) was annotated by two of the trained observers and the radiologist. Manual calcium annotation usually requires the observer to select only a single voxel per lesion. The lesion is then automatically segmented with region growing using the standard intensity threshold of 130 HU. In low-dose scans, however, intensity based region growing often leads to large amounts of noise being segmented with the calcium (Figure 2.1). Moreover, it can lead to the spine and ribs being segmented together with calcium, or calcifications in arteries branching off the aorta being segmented together with calcium in the aorta. The observers therefore marked calcifications voxel-by-voxel (≥130 HU) in the coronary arteries, the aorta and the aortic and mitral valves, including the annulus. Coronary calcifications were labeled as either left anterior descending artery (LAD), left circumflex artery (LCX) or right coronary artery (RCA). The left main coronary artery was considered part of LAD because these are difficult to

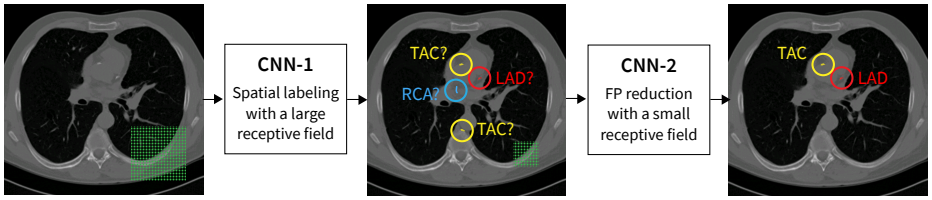


Figure 2.2: Overview of the proposed calcium detection method. Two CNNs are applied consecutively to first identify and label candidates (CNN-1), and to finally identify true calcifications among the candidates (CNN-2). The size of the receptive fields of the networks are indicated by green dotted areas.

distinguish on ungated scans. Motion artifacts caused by calcifications were annotated as calcifications because an exact separation of true calcification and artifact is often not possible. Depending on the amount of calcification and the image quality, the annotation effort varied from 5–10 minutes for images with soft reconstruction and little calcium to 60–90 minutes for images with sharp reconstruction and/or large amounts of calcium.

2.3 Method

The proposed method for automatic detection of CAC, TAC and calcifications of the aortic and mitral valves consists of two steps. Each step uses a CNN to classify voxels in the image. The first CNN (CNN-1) has a large receptive field to be able to detect calcium based on the anatomical context and to label calcium voxels according to their anatomical location. The second CNN (CNN-2) has a smaller receptive field and discards false positives based on local image information. Only voxels that CNN-1 considers calcium are classified by CNN-2 as either true-positive or false-positive (Figure 2.2).

2.3.1 First stage network (CNN-1)

CNN-1 classifies all voxels in the image that exceed the standard calcium threshold of 130 HU. The number of voxels that need to be classified in each scan is therefore high. Classification voxel-by-voxel with a sliding window approach would be highly inefficient as many identical convolutional operations would be repeated unnecessarily. CNN-1 is therefore constructed as a purely convolutional network,[55,56] i.e., with all layers implemented as convolutions, which allows arbitrary-sized inputs so that entire slices or volumes can be classified at once (Figure 2.3). CNN-1 classifies voxels as either LAD (including the left main coronary artery), LCX, RCA, TAC, aortic valve calcification, mitral valve calcification or background.

Previous publications showed that spatial information is particularly important for

calcium detection. To allow CNN-1 to infer spatial information from the image area covered by its receptive field, its receptive field needs to be relatively large. However, CNNs with large receptive fields, such as very deep networks or networks with large convolution kernels, often suffer from overfitting due to large numbers of trainable parameters. To allow for a large receptive field while keeping the number of trainable parameters low, we rely on dilated convolutions, which are based on convolution kernels with spacing between their elements. By stacking convolutions with exponentially growing dilation, the receptive field of the network grows exponentially while the number of trainable parameters only grows linearly.[57] At the same time, the network still includes all information within its receptive field in the analysis.

The architecture of CNN-1 is similar to the network proposed by Yu et al.,[57] but extends it from 2D inputs to three orthogonal 2D inputs (often referred to as 2.5D).[58] We chose 2.5D inputs over 3D inputs because of the previously reported superior performance for calcium scoring.[45] The receptive field of CNN-1 is 131 × 131 pixels, which corresponds to roughly a quarter of an axial slice.

The input of CNN-1 is a set of three orthogonal patches, which always intersect in a single voxel regardless of the patch size. This contradicts the idea of purely convolutional networks as larger inputs do then not lead to larger outputs. However, the processing of 2.5D input patches can be divided into 2D subtasks by processing the axial, sagittal and coronal inputs independently.[45] Each 2D input is first separately processed with the subnetwork for the respective orientation, allowing us to obtain a feature representation per patch in an efficient manner. The remaining layers of the network are applied to the concatenated feature vectors from all three orientations to obtain posterior probabilities for each voxel.

Inspired by the concept of deep supervision,[59] each subnetwork has an auxiliary softmax output layer (O_A, O_S and O_C in Figure 2.3). These are used during training to enable learning from input patches larger than the receptive field, i.e., learning from multiple labeled pixels per patch. Using these auxiliary output layers, auxiliary loss terms L_A, L_S and L_C are defined. A loss term L_N is defined with the output of the entire network for only the voxel in the intersection of the three orthogonal input patches. These loss terms are combined into an overall loss term L using a weight factor γ:

$$L = L_N + \gamma (L_A + L_S + L_C) . \tag{2.1}$$

After training, the auxiliary output layers are used together with the output layer of the entire network for classification. The corresponding posterior probabilities $\mathbf{p}_A$, $\mathbf{p}_S$ and $\mathbf{p}_C$ from the auxiliary output layers are combined with the posterior probabilities $\mathbf{p}_N$ by computing the weighted average of the probabilities with weights ω_A, ω_S, ω_C, and ω_N.

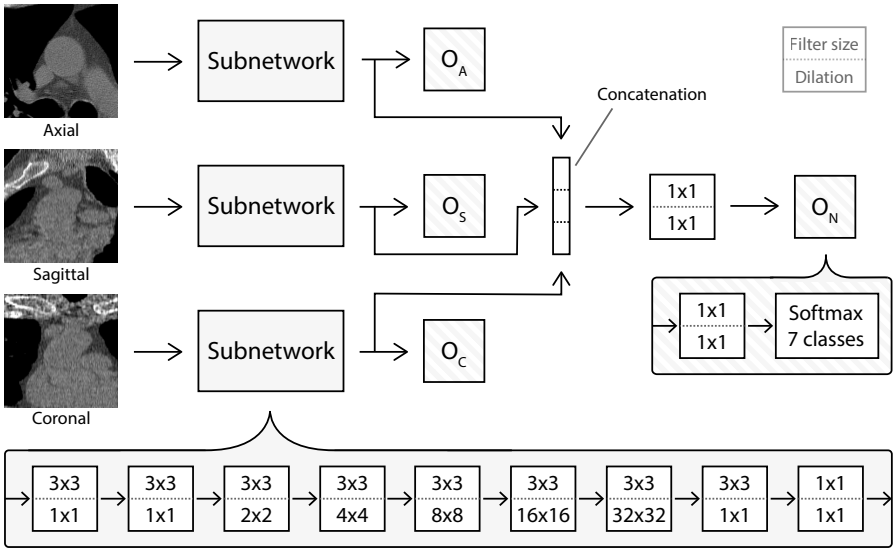


Figure 2.3: Architecture of CNN-1. The three orthogonal patches are analyzed by subnetworks with identical structure (bottom). The seven output classes are LAD, LCX, RCA, TAC, aortic valve, mitral valve and background. Convolutional layers are shown as boxes with filter size (top) and dilation factor (bottom). All convolutional layers consist of 32 filters, only the layer before O_N consists of 128 filters.

2.3.2 Second stage network (CNN-2)

CNN-1 detects potential calcifications based on appearance and spatial context, and furthermore determines whether the calcification is located in the coronary arteries (LAD including left main, LCX or RCA), the aorta, or the aortic or mitral valve. However, metal artifacts, image noise or other high intensity structures, such as parts of the spine in direct proximity to the aorta, can result in false positive voxel detections.

CNN-2 refines the output of CNN-1 by distinguishing between true calcifications and false positive voxels with similar appearance and location. In contrast to CNN-1, CNN-2 does therefore not need to focus on the spatial context but can focus on local information and finer details. CNN-2 does not use dilated convolutions, but instead non-dilated convolutions with max-pooling between convolutions. CNN-2 is not purely convolutional like CNN-1 as it only needs to analyze a limited number of voxels. CNN-2 analyzes 2.5D inputs and has a receptive field of 65 × 65 pixels. Opposed to the multi-class output of CNN-1, the output of CNN-2 is binary as its purpose is false positive reduction and not categorization of the detected calcifications. The architecture of CNN-2 (Figure 2.4) is motivated by our preliminary work on coronary calcium scoring,[48] in which a similar network achieved good performance when the problem was restricted to a region of interest. In this work, CNN-1 detects and labels candidate voxels, i.e.,

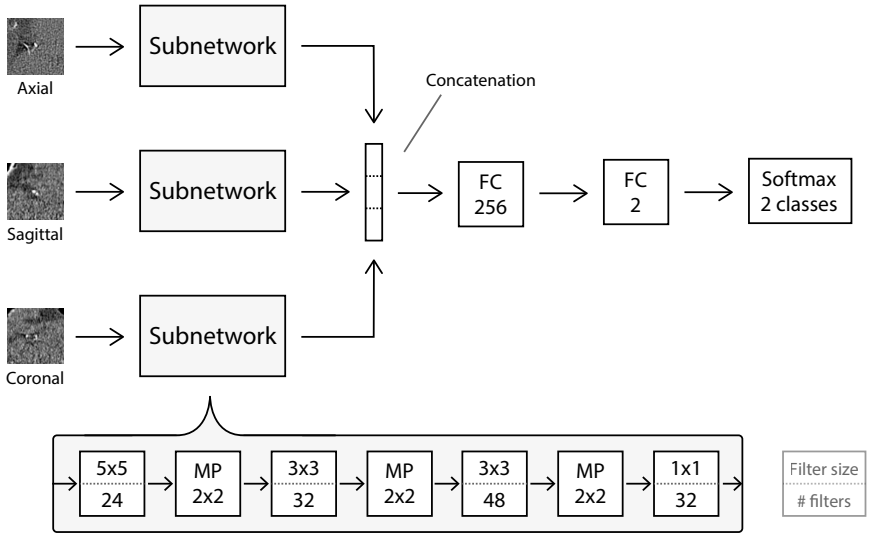


Figure 2.4: Architecture of CNN-2. The three orthogonal patches are analyzed by subnetworks with identical structure (bottom). The two output classes are calcium and background. Convolutional layers are shown as boxes with filter size (top) and number of filters (bottom). **MP** denotes max-pooling layers with the specified pooling region. **FC** denotes fully-connected (dense) layers with the specified number of units.

objects of interest, instead of a region of interest.

2.4 Evaluation

Calcium scoring is normally performed in images with soft tissue reconstruction. However, lung cancer screening data also includes images reconstructed with sharper filter kernels, in which edges but also noise appear more prominent (Figure 2.5). To evaluate whether our method needs to be trained with images that are reconstructed with parameters recommended for calcium scoring (soft tissue filter kernels) or whether it can be trained with all images acquired in the screening (see Table 2.1), two pairs of CNN-1 and CNN-2 were trained: one pair using soft reconstructions only and the other pair using both soft and sharp reconstructions. Furthermore, to evaluate whether calcium detection can be performed in images reconstructed using soft and sharp filter kernels, and to evaluate which of the two training settings leads to best performance, we evaluated our method separately on soft and sharp reconstructions.

To evaluate the performance of the method, calcifications were quantified per subject and per label using volume and Agatston scores. Agatston scores were normalized to account for overlapping slices.[60] The agreement between automatically and manually determined calcium volumes was assessed using sensitivity, average false positive

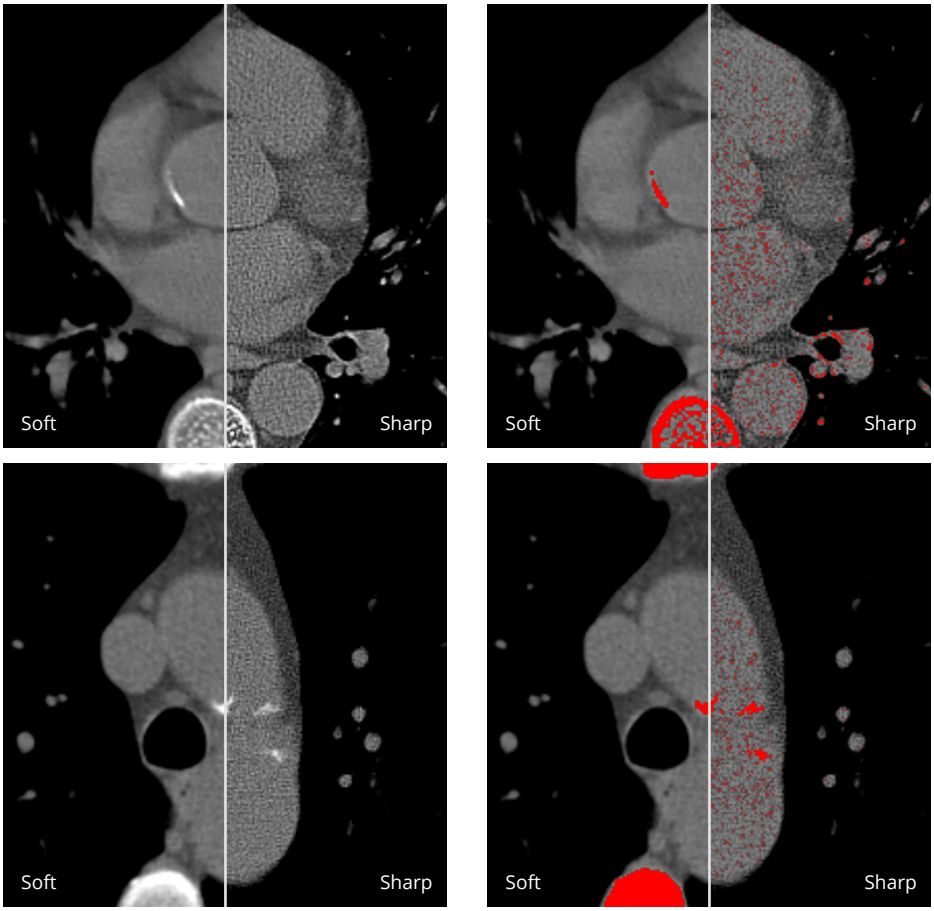


Figure 2.5: Comparison of low-dose chest CT scans reconstructed with soft filter kernel (left half of each image) and sharp filter kernel (right half of each image). The figures on the right show the same scans as the figures on the left, but with voxels ≥130 HU highlighted in red, indicating all voxels above the standard calcium threshold.

Table 2.2: Manual and automatic CAC scoring performance on 310 soft and 196 sharp reconstructions. *Reference standard* refers to the amount of calcium manually identified by the observers. The performance of a *second and third observer* on a subset of 100 scans is additionally reported. The performance of the automatic method is reported when the networks were *trained on soft reconstructions* and when they were *trained on soft and sharp reconstructions*.

	Soft reconstructions				*Sharp reconstructions*			
	CAC	LAD	LCX	RCA	CAC	LAD	LCX	RCA
Reference standard								
Scans with any calcification (%)	74.5	67.7	41.6	48.1	76.5	71.9	36.7	49.5
Calcium volume / scan (mm^3)	303.4	150.2	50.4	102.9	377.65	171.3	51.1	155.3
Second observer[†]								
Sensitivity (%)	94.8	93.5	83.6	95.8	86.8	88.5	79.4	84.1
False positive volume / scan (mm^3)	20.9	9.9	9.2	12.3	23.3	15.0	9.2	9.7
F_1 score calcium volume	0.95	0.94	0.87	0.96	0.85	0.86	0.75	0.83
Third observer[†]								
Sensitivity (%)	97.6	98.9	87.0	96.7	91.0	91.2	86.0	87.8
False positive volume / scan (mm^3)	19.7	12.3	8.1	10.2	13.5	7.3	11.1	8.5
F_1 score calcium volume	0.97	0.96	0.89	0.97	0.94	0.94	0.86	0.92
Trained on soft reconstructions only								
Sensitivity (%)	91.3	92.5	71.8	91.5	53.7	54.0	44.9	54.1
False positive volume / scan (mm^3)	35.5	18.3	13.6	11.2	19.4	10.6	4.7	7.4
F_1 score calcium volume	0.90	0.90	0.72	0.90	0.68	0.67	0.58	0.68
Trained on soft and sharp reconstructions								
Sensitivity (%)	91.2	93.7	65.8	90.4	84.4	86.9	61.6	83.0
False positive volume / scan (mm^3)	40.7	26.6	12.0	11.8	62.8	37.0	9.9	25.2
F_1 score calcium volume	0.89	0.89	0.69	0.90	0.84	0.83	0.68	0.83

[†]subset of 100 scans

volume per scan and the F_1 score. Interobserver agreement was assessed using the same metrics by comparing the annotations of the second and third observers to the radiologist as reference.

Additionally, we evaluated cardiovascular risk classification. Each subject was assigned one of four risk categories (I–IV: 0–10, 11–100, 101–1000, >1000) based on their total CAC Agatston score. Reliability of the risk category assignment was assessed using the linearly weighted κ coefficient.

2.5 Experiments and Results

The image quality of 57 scans (3 %) was considered inadequate for manual annotation due to severe metal artifacts or excessive image noise. The remaining 1687 scans with manual reference standard were divided into subsets for training (60 % = 1012 scans), validation (10 % = 169 scans) and testing (30 % = 506 scans). These were scans of 1459 different participants as sometimes multiple reconstructions of the same baseline

Table 2.3: Agreement in risk categorization of subjects based on their total CAC Agatston score (I: 0–10, II: 11–100, III: 101–1000, IV: >1000) between manual reference standard and automatically determined scores. The agreement is reported separately for soft and sharp reconstructions. The table on the left side specifies the agreement when the automatic method was *trained on soft reconstructions only* and the table on the right side when the automatic method was *trained on soft and sharp reconstructions*.

Trained on soft reconstructions only

	Soft reconstructions				*Sharp reconstructions*			
	Automatic				Automatic			
Reference	I	II	III	IV	I	II	III	IV
I	**90**	17	1	0	**62**	2	1	0
II	3	**59**	4	0	12	**18**	0	0
III	0	2	**99**	2	9	12	**51**	1
IV	0	0	1	**32**	3	4	6	**15**
	$\kappa = 0.91$				$\kappa = 0.70$			

Trained on soft and sharp reconstructions

	Soft reconstructions				*Sharp reconstructions*			
	Automatic				Automatic			
Reference	I	II	III	IV	I	II	III	IV
I	**92**	12	4	0	**55**	10	0	0
II	3	**58**	5	0	0	**25**	5	0
III	0	2	**99**	2	1	1	**68**	3
IV	0	0	1	**32**	0	0	1	**27**
	$\kappa = 0.91$				$\kappa = 0.90$			

scan were included. The division into subsets was random, but all scans of the same participant were assigned to the same subset. Among the 1687 scans, 58 % were reconstructed with soft filters and 42 % with sharp filters. Among the 100 scans with multiple annotations, 52 % were reconstructed with soft filters and 48 % with sharp filters.

The areas covered by the receptive fields of both networks are not necessarily comparable across scans due to different resolutions. We therefore resampled all scans in-plane with bilinear interpolation to 0.66 mm × 0.66 mm, which is the average resolution across the dataset. Resampling was only performed in-plane because the slice spacing was already standardized to 1.5 mm before manual annotation (see Section 2.2). Both networks CNN-1 and CNN-2 thus analyzed images at a standardized resolution. The predicted label maps were finally resampled to the original resolution using nearest neighbor interpolation.

The two networks CNN-1 and CNN-2 were trained sequentially. CNN-1 was trained with high density voxels (≥130 HU) in the training scans. CNN-2 was trained with high density voxels classified as any type of calcification by CNN-1. The validation set was used to ensure there was no substantial overfitting and to determine convergence of the networks. We trained both networks on balanced minibatches, which consisted half of randomly selected calcium voxels of any class and half of randomly selected background voxels. Both networks used exponential linear units[61] as activation function. Adam[62] was used as optimizer with a learning rate of 5×10^{-4} and the categorical cross-entropy as loss function. CNN-1 was trained on patches of 155 × 155 pixels, i.e., larger than its receptive field, with $\gamma = 0.05$ and CNN-2 on patches of 65 × 65 pixels. During training of both networks, Dropout[63] (CNN-1: 35 %, CNN-2: 50 %) and L2 weight decay (CNN-1: 5×10^{-5}, CNN-2: 1×10^{-5}) were used for regularization. Since CNN-2 has many more trainable parameters, we added batch normalization[64] between all layers to provide additional regularization. For both networks, the output class was determined as the class with highest activation. CNN-1 used $\omega_N = \frac{1}{2}$ and $\omega_{A,S,C} = \frac{1}{6}$ to weight the output layers, which corresponds to averaging between the normal and the auxiliary outputs.

The networks were implemented using the Theano[65] and Lasagne[66] frameworks and trained on NVIDIA Titan X GPUs. The total computation time is 5–7 minutes, depending on the size of the image volume and the number of candidate objects in the image. With our non-optimized implementation, CNN-1 needs on average 5.5 ± 1 minutes to scan an entire image, and CNN-2 needs on average 8.6 ± 8.1 seconds to classify the detected candidate voxels.

2.5.1 Detection of CAC

The performance of automatic CAC detection was evaluated based on scores per artery and per subject. Per artery and per subject sensitivities, average false positive volumes and F_1 scores for CAC detection are listed in Table 2.2. Examples of detected calcifications are shown in Figure 2.6.

In images reconstructed with soft filter kernels, the automatic method detected more than 90 % of the calcium in LAD and RCA, and 72 % in LCX. Training on soft and sharp reconstructions compared to only soft reconstructions led to similar performance for CAC with F_1 scores of 0.89 and 0.90. In images reconstructed with sharp filter kernels, the F_1 score for CAC increased from 0.68 to 0.84 when sharp reconstructions were added to the training data.

Risk categories derived from per-subject CAC scores agreed with the manual reference annotation in images reconstructed with soft filter kernels in 90 % of the subjects when trained only on soft reconstructions and in 91 % when sharp reconstructions were added to the training data. In images reconstructed with sharp filter kernels, agreement increased from 75 % to 89 % when sharp reconstructions were added to the training data. Confusion matrices for the risk category assignment are shown in Table III.

Interobserver agreement was high for per-subject CAC in soft reconstructions with F_1 scores of 0.95 and 0.97 for the second and third observer, respectively. Interobserver agreement was overall slightly lower in sharp reconstructions compared to soft reconstructions. For risk category assignment, κ was 0.98 and 1.00 for the second and third observer, respectively, in soft reconstructions and 0.92 and 0.99 in sharp reconstructions.

2.5.2 Detection of TAC

The observers identified TAC in the majority of scans, in 450 of 506 (89 %). The detection performance of the automatic method in terms of sensitivity, average false positive volume and F_1 score is listed in Table 2.4. In images reconstructed with soft filter kernels, the automatic method achieved an F_1 score of 0.89 regardless whether only soft reconstructions or soft and sharp reconstructions were used for training. In images reconstructed with sharp filter kernels, the F_1 score increased from 0.71 to 0.81 when sharp reconstructions were added to the training data. Interobserver agreement was high for TAC and, overall, the observers had higher sensitivity than the automatic method, but also higher average false positive volume.

2.5.3 Detection of cardiac valve calcifications

Cardiac valve calcifications were infrequently identified by the observers, aortic valve calcifications in 92 of 506 scans (18.2 %) and mitral valve calcifications in 58 of 506 scans (11.5 %). The detection performance of the automatic method in terms of sensitivity, average false positive volume and F_1 score is listed in Table 2.5.

The automatic method achieved lower performance for detection of cardiac valve calcifications compared to detection of CAC and TAC. However, similar was that adding sharp reconstructions to the training data improved F_1 scores in sharp reconstructions, from 0.59 to 0.64 for aortic valve calcifications and from 0.52 to 0.66 for mitral valve calcifications. Interobserver agreement in terms of F_1 score was particularly low in

Table 2.4: Manual and automatic TAC scoring performance on 310 soft and 196 sharp reconstructions. *Reference standard* refers to the amount of calcium manually identified by the observers. The performance of a *second and third observer* on a subset of 100 scans is additionally reported. The performance of the automatic method is reported when the networks were *trained on soft reconstructions* and when they were *trained on soft and sharp reconstructions*.

Reconstruction filter	*Soft*	*Sharp*
Reference standard		
Scans with any calcium (%)	89.4	88.3
Calcium volume / scan (mm^3)	1875.0	1476.4
Second observer[†]		
Sensitivity (%)	97.6	90.2
False positive volume / scan (mm^3)	300.6	367.7
F_1 score calcium volume	0.93	0.79
Third observer[†]		
Sensitivity (%)	93.7	86.5
False positive volume / scan (mm^3)	366.2	537.4
F_1 score calcium volume	0.90	0.75
Trained on soft reconstructions only		
Sensitivity (%)	89.8	60.1
False positive volume / scan (mm^3)	213.3	152.7
F_1 score calcium volume	0.89	0.71
Trained on soft and sharp reconstructions		
Sensitivity (%)	92.0	85.1
False positive volume / scan (mm^3)	258.9	379.3
F_1 score calcium volume	0.89	0.81

[†]subset of 100 scans

Table 2.5: Performance of manual and automatic scoring of aortic and mitral valve calcifications, reported separately for 310 soft and 196 sharp reconstructions. *Reference standard* refers to the amount of calcium manually identified by the observers. The performance of a *second and third observer* on a subset of 100 scans is additionally reported. The performance of the automatic method is reported when the networks were *trained on soft reconstructions* and when they were *trained on soft and sharp reconstructions*.

	Soft reconstructions		*Sharp reconstructions*	
	Aortic valve	Mitral valve	Aortic valve	Mitral valve
Reference standard				
Scans with any calcium (%)	20.6	12.6	14.3	9.2
Calcium volume / scan (mm^3)	44.9	31.1	62.1	38.8
Second observer[†]				
Sensitivity (%)	86.8	83.2	21.9	9.0
False positive volume / scan (mm^3)	36.5	8.7	4.1	1.7
F_1 score calcium volume	0.74	0.88	0.18	0.09
Third observer[†]				
Sensitivity (%)	97.1	88.0	87.3	9.8
False positive volume / scan (mm^3)	64.1	19.1	15.1	1.6
F_1 score calcium volume	0.69	0.88	0.51	0.17
Trained on soft reconstructions only				
Sensitivity (%)	55.3	64.2	44.8	39.4
False positive volume / scan (mm^3)	5.4	18.2	3.6	4.3
F_1 score calcium volume	0.66	0.58	0.59	0.52
Trained on soft and sharp reconstructions				
Sensitivity (%)	58.7	67.1	57.1	60.8
False positive volume / scan (mm^3)	7.4	24.5	12.9	9.5
F_1 score calcium volume	0.67	0.55	0.64	0.66

[†]subset of 100 scans

images reconstructed with sharp filter kernels low with F_1 scores of 0.18 / 0.51 and 0.09 / 0.17 for aortic and mitral valve calcifications, respectively.

2.5.4 Single vs. two-stage performance

To evaluate the contribution of CNN-2 to the overall performance, we compared the performance of CNN-1 followed by CNN-2 against CNN-1 alone. Trained using only images with soft reconstruction kernels, CNN-1 achieved a sensitivity of 96.6 % with an average false-positive volume of 5574 mm^3 on soft reconstructions, and 90.7 % sensitivity with 18 739 mm^3 average false-positive volume on sharp reconstructions. Reclassification of the positive detections by CNN-2 reduced the sensitivity by 6.1 % and 32.0 %, but at the same time reduced the average false-positive volume per scan by 95.7 % and 99.1 % in soft and sharp reconstructions, respectively.

When images with sharp reconstruction kernels were added to the training data, CNN-1 achieved a sensitivity of 97.3 % at an average false-positive volume of 6495 mm^3 in soft reconstructions. In sharp reconstructions, the sensitivity was 92.6 % at 14 802 mm^3 average false-positive volume. Reclassification by CNN-2 reduced the sensitivity by 4.8 % and 7.7 %, but at the same time reduced the average false-positive volume per scan by 95.4 % and 97.0 % in soft and sharp reconstructions, respectively.

2.5.5 Effect of receptive field size

To evaluate the influence of the size of the receptive field of CNN-1 on the detection performance, we trained networks with various maximum dilation factors. A larger maximal dilation factor results in a larger receptive field and also a deeper network (Figure 2.3;[57]). For receptive field sizes of 35 × 35, 67 × 67, 131 × 131 and 259 × 259, CNN-1 achieved F_1 scores of 0.27, 0.29, 0.40 and 0.15, respectively, in images reconstructed with soft reconstruction kernels. In images with sharp reconstruction kernel, the F_1 scores were 0.18, 0.14, 0.20 and 0.07. The largest network had to be trained on smaller batches of samples due to hardware limitations (32 instead of 64), which made the network more difficult to train.

2.5.6 Effect of auxiliary output layers and loss terms

We additionally evaluated whether the auxiliary softmax output layers of CNN-1 with the corresponding auxiliary loss terms had a positive effect on training time and performance. We observed that the network learned slower with the auxiliary loss terms. However, the detection performance improved. The performance of CNN-1 with and without auxiliary output layers and loss terms was compared using training images reconstructed with both soft and sharp reconstruction kernels. In images reconstructed with soft filter kernels, F_1 scores for calcium detection (binary, disregarding the label)

were 0.17 without auxiliary outputs and 0.40 with auxiliary outputs. In images reconstructed with sharp filter kernels, the F_1 scores were 0.07 without auxiliary outputs and 0.20 with auxiliary outputs.

2.5.7 Comparison with other methods

The number of other publications on automatic calcium scoring methods in low-dose chest CT is low. No other methods have been published that concurrently detected CAC, TAC and cardiac valve calcifications, and also no methods for detection of cardiac valve calcifications. Hence, the results of the proposed combined system can only be compared to the simpler tasks of detecting either only CAC or only TAC.

For automatic CAC scoring in low-dose chest CT, Išgum et al.[46] reported a sensitivity of 79 % at 4 mm^3 false positive volume per scan in 231 scans. In the same scans, Lessmann et al.[48] achieved a sensitivity of 97 % at an average false positive volume of 10 mm^3. However, the dataset on which these methods were tested was much less diverse than the dataset used in this chapter. All scans were acquired in the same hospital with CT scanners from a single vendor and reconstructed using a single soft filter kernel. Furthermore, the average CAC burden per subject was considerably lower (198 mm^3 vs. 303 mm^3). We therefore evaluated the better performing method[48] on the current test data after retraining with the current training data. The method described in[48] classifies individual voxels, but originally the average posterior probabilities across connected voxels was calculated to classify lesions rather than individual voxels. This averaging was now omitted to allow for a comparison on voxel-level. The best performance was achieved on soft reconstructions, using both soft and sharp reconstructions for training: the average F_1 score per scan for CAC detection was 0.23 (95 % CI: 0.20–0.25), mostly attributed to many false positive detections. In comparison, the proposed method achieved an average F_1 score of 0.87 (95 % CI: 0.85–0.90) on the same test data using the same training data. This difference was statistically significant ($p < 0.001$, paired samples t-test).

Other methods for automatic CAC scoring were only evaluated in terms of their correlations with manual scores: Xie et al.[47] performed linear regression with CAC Agatston scores in 41 scans and reported $R^2 = 0.91$. González et al.[49] reported a Pearson correlation coefficient of 0.86 for CAC Agatston scores in 1000 scans.

For automatic TAC scoring in low-dose chest CT, Išgum et al.[54] reported a sensitivity of 98 % at 64 mm^3 false positive volume per scan in 93 scans. Kurugol et al.[52] reported a sensitivity of 94 % and a positive predictive value of 91 % for TAC volume in 45 scans. In comparison, we reported here a sensitivity of 90 % at an average false positive volume of 213 mm^3 and a positive predictive value of 89 % in 310 scans reconstructed with soft filter kernels. Xie et al.[53] reported only the correlation with manual scores as $R^2 = 0.98$ after performing linear regression for TAC volume in 45 scans.

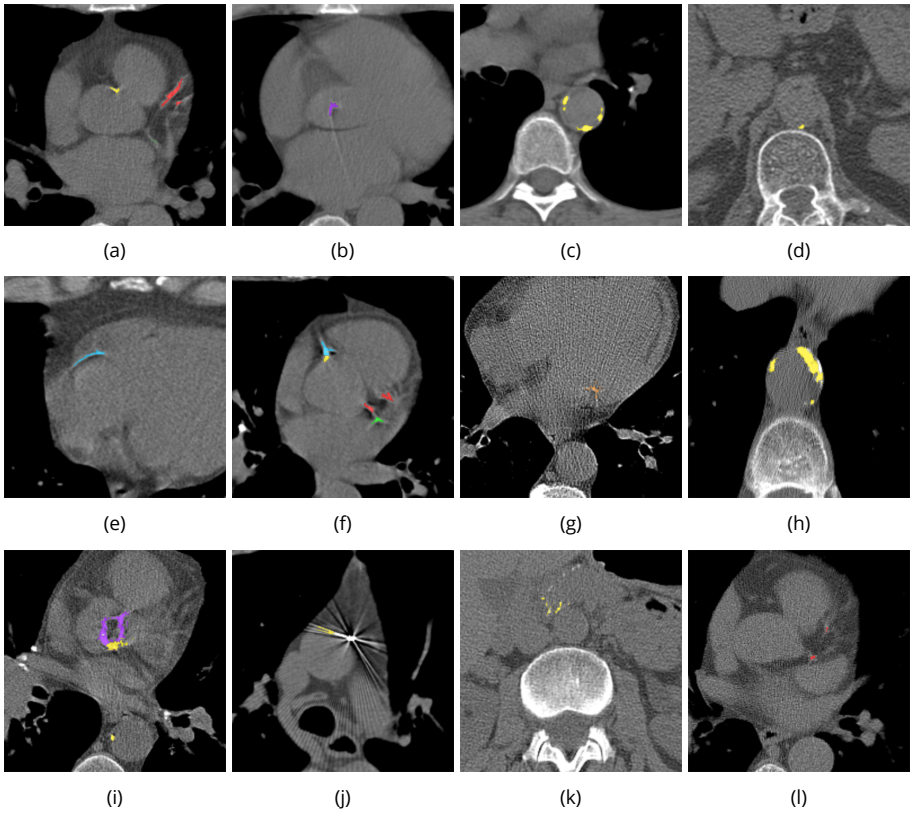


Figure 2.6: Example cases overlayed with color coded automatic detection results. The color scheme is as follows: red=LAD including LM, green=LCX, blue=RCA, yellow=TAC, purple=aortic valve calcification, orange=mitral valve calcification. Examples of (a)–(f) correctly detected and labeled calcifications, (g) and (h) correctly labeled calcifications of which a few voxels were missed, (i) an artificial aortic valve that was labeled as calcium, (j) a metal artifact next to the aortic wall that was partially labeled as calcium, (k) a false positive detection in proximity of the aorta, (l) a correctly detected calcification in LCX that was incorrectly labeled as LAD.

2.5.8 Voxel-level vs. lesion-level annotation

The manual reference annotation was performed voxel-by-voxel to enable annotation of scans with poor image quality. To assess how much voxel-level annotation differs from the standard lesion-level annotation, we converted the voxel-level annotations into lesion-level annotations using 3D region growing with the standard calcium threshold (130 HU). Lesions were labeled using majority voting if they contained voxels with different labels. Scans with a more than five times increase in calcium volume were excluded. These were 21 % of the scans with soft reconstruction kernel and 64 % of the scans with sharp reconstruction kernel. In the remaining scans, the overall Agatston score increased on average by 85 in soft reconstructions and by 155 in sharp reconstructions. This clearly indicates that in low-dose CT scans, lesion-level annotation leads to an overestimation of the calcium score.

2.6 Discussion

We proposed a method for automatic detection of CAC subdivided into LAD, LCX and RCA calcifications, TAC and cardiac valve calcifications in low-dose chest CT. The method is the first that detects these calcifications concurrently. The approach is based on two consecutive CNNs: The first CNN uses stacked dilated convolutions to facilitate a large receptive field, which enables identification and spatial labeling of high density voxels. The second CNN discards false positive detections of the first CNN.

For CAC and TAC detection, the method achieved a performance close to the level of interobserver agreement. The method was furthermore able to separate calcifications in the coronary arteries into LAD, LCX and RCA calcifications (Figure 2.6 (f)). The method as well as the observers were more successful in identification of LAD and RCA calcifications than LCX calcifications. The course of LCX is particularly difficult to follow in non-contrast scans. Hence, LCX calcifications can be difficult to differentiate from LM and LAD calcifications (Figure 2.6 (l)), as well as from those in the mitral valves. In comparison to CAC and TAC, calcifications of the aortic and mitral valves were less common in our dataset. Performance of the automatic method was below the performance of CAC and TAC detection. However, this is also a difficult task for experts. The observers especially disagreed on mitral valve calcifications, which is in line with findings of previous studies.[67] The disagreement is mainly caused by confusion with LCX calcifications and the lack of soft tissue contrast in the mitral valve region. For the aortic valve, confusion with TAC was the most common cause of disagreement.

False positive detections were mostly caused by mislabeling of calcifications with respect to their location (e.g., LAD and LCX), low-dose and motion artifacts, and other calcifications such as calcified lymph nodes or calcifications in other vessels (Figure 2.6 (i)–(k)). False positive detections outside the heart and the aorta occurred infrequently and usually in proximity to the heart or the aorta. This demonstrates

that CNN-1 was able to implicitly learn to recognize the typical spatial context of calcifications in the image. The individual evaluation of CNN-1 additionally showed that CNN-2 substantially contributes to reducing false-positive detections while maintaining a high sensitivity. However, future work could aim at unifying the two networks into a single network.

False negative detections were sometimes partially misclassified lesions (Figure 2.6 (h)–(i)). Partial misclassification can occur because the method performs voxel classification rather than the standardly used lesion classification. Even though voxel labeling occasionally causes partial misclassification of calcifications, it enables splitting of calcifications that are contained in more than one arterial bed, such as those partly located in the aorta and partly in the coronary arteries (Figure 2.6 (f)). Assigning a calcification that is partially contained in the aorta to the coronary artery could affect cardiovascular risk categorization. Similarly, assigning LM calcifications to the aorta would result in missing high risk lesions. To the best of our knowledge, this is the first method enabling splitting of the calcifications according to their arterial bed.

Other methods for calcification detection often first detect a region of interest in the image. The proposed method is able to omit this step and instead searches the entire image for calcifications without the need for any preprocessing steps. Moreover, the method does not require explicit spatial features, even though these have been reported in the literature as crucial for automatic calcium scoring. The results demonstrate that a CNN with dilated convolutions is able to recognize the spatial context in three orthogonal 2D patches.

In contrast, other commonly used network architectures have various shortcomings that make them less suited for calcium scoring in low-dose chest CT. For example, U-net[68] is not well suited for sparse problems because it can process 3D volumes only in smaller tiles, of which most would not contain any calcium voxels. Residual networks[69] and the similar densely connected networks[70] use pooling to increase their receptive field, which is not compatible with the idea of purely convolutional networks. Classifying all voxels in a typical 3D chest CT volume with these networks would be inefficient and time-consuming. Hough networks[71] are based on the idea of enforcing learned shape priors, which is useful in segmentation tasks of structures with relatively homogeneous shape. However, the shape of calcifications is rather heterogeneous, especially if scans are distorted by cardiac motion. Spatial transformer networks[72] address alignment issues, but chest CT scans are fairly standardized. The scanned subject typically lies on the back and the FOV of the reconstructed image is standardly configured using the outer body contour or the ribs, and the apex/base of the lungs as landmarks.

A particular strength of this chapter is the large, diverse and realistic dataset of low-dose chest CT scans from the NLST that we used for training and evaluation. Even though this data is challenging due to low radiation dose, the lack of ECG-synchronization and the high diversity of image acquisition parameters, the automatic

method achieved good detection performance and high agreement in risk categorization. The separate evaluation in images reconstructed with soft and sharp filter kernels additionally demonstrates that the performance on soft reconstructions does not suffer when sharp reconstructions are added to the training data. This indicates that the networks were able to generalize to both types of reconstructions.

The high reliability of the risk categorization indicates that this method can be used for cardiovascular risk assessment in lung cancer screening. While standardized risk categories are defined for CAC scores, TAC and cardiac valve calcium scores are currently not commonly used for cardiovascular risk assessment. Automatic scoring of these calcifications enables evaluation of their predictive value using available large datasets from lung screening trials, or other screening trials with CT imaging visualizing the heart.

Acknowledgments

We are grateful to the United States National Cancer Institute (NCI) for providing access to NCI's data collected by the National Lung Screening Trial. The statements contained herein are solely ours and do not represent or imply concurrence or endorsement by NCI. To enable other researchers to request the same data and perform a comparison with the here presented results, NCI has been provided with the list of scans used in this study.

3

CHAPTER 3

Sex differences in arterial calcification in heavy smokers

BASED ON: N. Lessmann, P. A. de Jong, C. Celeng, R. A. P. Takx, M. A. Viergever, B. van Ginneken, and I. Išgum. "Sex differences in coronary artery and thoracic aorta calcification and their association with cardiovascular mortality in heavy smokers". *JACC: Cardiovascular Imaging* (2019). In press.

Abstract

OBJECTIVES The aim of this study was to investigate sex differences in prevalence, extent and association with cardiovascular mortality of coronary artery calcium (CAC) and thoracic aorta calcium (TAC) scores in the population eligible for lung screening.

BACKGROUND CAC and TAC scores derived from chest computed tomography (CT) may be useful biomarkers for individualized cardiovascular disease prevention and could be especially relevant in high risk populations such as heavy smokers. It is therefore important to know the prevalence of arterial calcifications in male and female heavy smokers, and if there are differences in the predictive value they carry.

METHODS We performed a nested case-control study with 5718 participants of the CT arm of the National Lung Screening Trial (NLST). Prevalence and extent of CAC and TAC were resampled to the full cohort to provide unbiased estimates of the typical calcium burden of male and female heavy smokers. Weighted Cox proportional hazards regression was used to assess differences in the association of CAC and TAC scores with all-cause and cardiovascular mortality.

RESULTS CAC was substantially more common and more severe in men (prevalence 81 % vs. 60 %, median volume 104 mm^3 vs. 12 mm^3). Women had CAC comparable to 10 year younger men. TAC was equally common in men and women with a tendency to be more pronounced in women (prevalence 92 % vs. 93 %, median volume 388 mm^3 vs. 404 mm^3). Both types of calcification were associated with increased cardiovascular and all-cause mortality. TAC scores improved the prediction of coronary heart disease mortality over CAC in men, but not in women. In both sexes, TAC but not CAC was associated with cardiovascular mortality other than coronary heart disease.

CONCLUSIONS CAC develops later in women while TAC develops equally in both sexes. CAC is strongly associated with coronary heart disease while TAC is especially associated with extracardiac vascular mortality in either sex.

3.1 Introduction

Although cardiovascular disease is one of the major causes of death for both men and women, clear differences exist between the sexes in incidence rates of cardiovascular events.[73,74] A strong predictor of cardiovascular events and mortality is the amount of arterial calcification that can be detected and quantified in cardiac or chest computed tomography (CT) scans.[16] CT-based lung cancer screening of heavy smokers therefore provides an opportunity for assessment of cardiovascular risk in addition to lung screening.[15,29,30,75] However, in the lung cancer screening eligible population, the prevalence and extent of vascular calcification is likely higher than in the general population due to smoking history and advanced age. Reference values for coronary artery calcium (CAC) and thoracic aorta calcium (TAC) scores obtained from ungated lung screening CT and the associated risk for cardiovascular events and mortality are not well established. Furthermore, while coronary atherosclerosis and coronary heart disease are known to affect men more often than women,[74] differences in aortic atherosclerosis are less well established. The sex difference has been reported to be smaller for TAC than for CAC,[76] which would be important information for correct interpretation of lung cancer screening derived calcium scores and sex specific management of these prevalent findings. Moreover, a limited number of studies suggest that TAC might be more relevant than CAC for prediction of non-cardiac events and all-cause mortality.[17,77] The purpose of this study was to investigate differences in prevalence, extent and association with cardiovascular and all-cause mortality of CAC and TAC between male and female participants of a lung cancer screening trial.

3.2 Methods

3.2.1 Study design

We performed a retrospective nested case-control study with data collected in the National Lung Screening Trial (NLST).[4] The design of the NLST including eligibility criteria, imaging protocols and participant characteristics has been described in detail elsewhere.[23] Briefly, the NLST compared screening for lung cancer with chest radiography against screening with low-dose chest CT. Eligibility criteria were a smoking history of at least 30 pack-years, active smoking within the previous 15 years and an age of 55–74 years at enrollment. A total of 26 722 participants were randomized to the CT arm of the trial. These participants underwent three rounds of low-dose, unenhanced, ungated chest CT imaging at 120 kVp in 1-year intervals in one of 33 participating screening centers in the United States. Enrollment began in August 2002 and participants were followed until the end of 2009, with a median follow-up time of 6.5 years.

In this case-control study, we included all participants in the CT arm of the NLST

who died and a control group of about twice as many randomly chosen participants from the CT arm who were either censored or still alive at the end of the follow-up period. In total, 5718 participants were included, of which 3.553 (62 %) were men and 1724 (30 %) died. Details of the selection are shown in Figure 3.1. In our statistical analysis, cases and controls were resampled to the full cohort size to remove the bias of the case-control design and make the results interpretable for the lung cancer screening eligible population.[78]

Endpoints for the survival analysis were defined based on the cause of death as reported on the death certificates. These were coded using the International Classification of Disease, 10th edition, and a single underlying cause of death was determined. We performed separate analyses for multiple endpoints. Next to all-cause mortality, these were death from cardiovascular disease (CVD), which was defined as death from any disease of the circulatory system (I00–I99), death from coronary heart disease (CHD), which was any death from ischemic heart diseases (I20–I25), cardiac arrest (I46) or heart failure (I50), and extracardiac vascular deaths, which were deaths from CVD excluding CHD.

3.2.2 Image analysis

Calcifications of the coronary arteries and the thoracic aorta were scored in the baseline scans using automatic, in-house developed software. This software has previously been validated with NLST scans.[79] Briefly, the software uses a deep learning approach based on two consecutive convolutional neural networks to detect calcifications and to label them according to the affected vascular bed. In this study, we calculated calcium scores for the total amount of calcification detected in either the coronary arteries (CAC score) or the thoracic aorta as far as visible, including the aortic arch (TAC score). Only voxels with an intensity of at least 130 Hounsfield units were taken into consideration, which is the standard threshold for calcium scoring.[27] Because calcium scoring is normally performed in images with a slice thickness of 3 mm, all scans were resampled prior to the analysis to a slice thickness of 3 mm with 1.5 mm increment. The detected calcifications were quantified by calculating the total calcium volume in mm^3 and a modified Agatston score that accounts for overlapping slices.[60]

3.2.3 Statistical analysis

Differences in the prevalence and extent of calcification of the coronary arteries and the aorta were tested in 5-year age groups (55–59, 60–64, 65–70, 70–74). Because of the non-normal distribution of calcium scores, scores stratified by age group and sex were summarized with medians and interquartile ranges. We calculated weighted medians and prevalences to account for the bias introduced by the case-control design of our study. Because our study population included all deaths, but only a subset of the remaining participants of the full NLST low-dose CT cohort, controls were

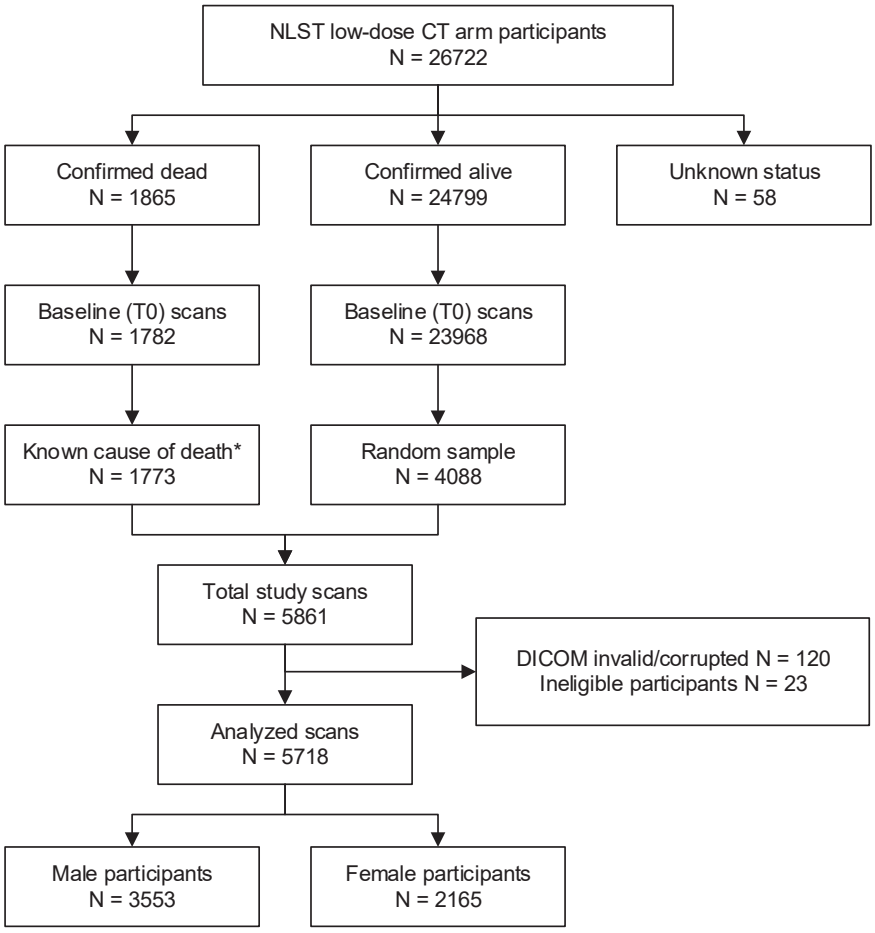


Figure 3.1: Flowchart summarizing the selection of the studied subgroup of NLST participants.

* Unknown cause of death was defined as missing death certificate or death coded as R96–R99 (ICD-10).

weighted with the inverse of their sampling probability. Prevalence of CAC and TAC was defined as an Agatston score ≥1 and was compared using chi-square tests with Rao-Scott second-order corrections to account for the sample weights.[80] The extent of calcification in terms of calcium volume was compared using Mann-Whitney U-tests.[81]

We used weighted Cox proportional-hazards regression[82] to examine the association between CAC and TAC scores and all-cause or cause-specific mortality. In addition to CVD death, we examined the association with cardiac, i.e., CHD, and extracardiac cardiovascular causes of death to investigate the hypothesis that CAC is more closely associated with cardiac and TAC with extracardiac causes of death. In the models for time to a specific cause of death, participants who died of another cause were censored at the time of death. In all cases, separate models were fit for male and female participants.

All models were adjusted for cardiovascular risk factors based on questionnaires filled in by the participants at study entry. The variables included in the models were age, smoking status, pack years, body mass index (BMI), race, education, marital status, and self-reported history of heart disease or stroke, diabetes and hypertension. CAC and TAC scores were examined as stratified variables with Agatston score cutoff points for CAC: 0, 1–100, 101–1000, >1000 and for TAC: 0–1000, 1001–4000, >4000. The cutoff points for CAC were higher than commonly used cutoff points because of the large number of elevated CAC scores in the study population.[15,29] Both CAC and TAC were included in all models because we were interested in their complementary rather than their individual predictive value.

We tested for differences in baseline characteristics between male and female cases and controls using Kruskal-Wallis-tests for continuous variables and chi-square tests, or Fisher's exact test if frequencies were low, for categorical variables. Participants with missing survival status or unknown cause of death and participants with missing calcium scores, owing to missing or unreadable images, were excluded from the analysis (Figure 3.1). Missing demographic and disease history data were imputed using regression techniques.[83] P-values < 0.05 were considered statistically significant. All statistical analyses were performed using R 3.4.

3.3 Results

3.3.1 Study population

Baseline characteristics of the study population are shown for cases and controls and according to sex in Table 3.1. In the study population, men compared with women were more likely married and better educated, and had on average accumulated more pack years. Men were also more likely to have had a diagnosis of heart disease or diabetes, while previous diagnoses of hypertension or stroke were equally likely in men and women.

Table 3.1: Baseline characteristics for cases and control subjects according to sex

	Men				Women			
	Deceased (cases) (n = 1206)	Selected control subjects (n = 2347)	NLST CT-arm control subjects (n = 14 422)	p-value	Deceased (cases) (n = 518)	Selected control subjects (n = 1647)	NLST CT-arm control subjects (n = 10 377)	p-value
Age (years)	64 ± 9	60 ± 7	60 ± 8	< 0.001	63 ± 9	60 ± 7	60 ± 7	< 0.001
BMI (kg/m^2)	27.3 ± 6.1	27.4 ± 5.6	27.6 ± 5.6	0.008	26.0 ± 8.7	26.6 ± 6.8	26.6 ± 6.9	0.143
Pack-years	60 ± 36	50 ± 30	51 ± 30	< 0.001	50 ± 26	44 ± 22	44 ± 22	< 0.001
Current smoker	695 (58)	1053 (45)	6557 (45)	< 0.001	327 (63)	808 (49)	5150 (50)	< 0.001
Married	809 (67)	1808 (77)	11 071 (77)	< 0.001	239 (46)	921 (56)	5587 (54)	< 0.001
Race				0.146				0.135
White	1090 (90)	2129 (91)	13 081 (91)		471 (91)	1530 (93)	9487 (91)	
Black	61 (5)	83 (3)	573 (4)		32 (6)	69 (4)	504 (5)	
Other/unknown	55 (5)	135 (6)	768 (5)		15 (3)	48 (3)	386 (4)	
Education				< 0.001				0.042
Up to high school diploma/GED	417 (35)	604 (26)	3796 (26)		202 (39)	544 (33)	3423 (33)	
Up to bachelor degree	612 (51)	1304 (56)	8025 (56)		250 (48)	894 (54)	5454 (53)	
Graduate school	148 (12)	399 (17)	2330 (16)		53 (48)	164 (10)	1223 (12)	
Other/unknown	29 (2)	40 (2)	271 (2)		13 (3)	45 (3)	277 (3)	
Disease history								
Heart disease	289 (24)	352 (15)	2293 (16)	< 0.001	88 (17)	114 (7)	730 (7)	< 0.001
Stroke	70 (6)	57 (2)	383 (3)	< 0.001	31 (6)	39 (2)	256 (2)	< 0.001
Diabetes	211 (18)	257 (11)	1548 (11)	< 0.001	69 (13)	102 (6)	724 (7)	< 0.001
Hypertension	510 (42)	802 (34)	5132 (36)	< 0.001	219 (42)	534 (32)	3438 (33)	< 0.001
CAC volume (mm^3)	288 ± 857	96 ± 397	—	< 0.001	64 ± 269	10 ± 90	—	< 0.001
TAC volume (mm^3)	1078 ± 2660	351 ± 1054	—	< 0.001	1088 ± 2761	384 ± 1193	—	< 0.001

Values are mean ± standard deviation or n (%). Disease history was self-reported at trial entry.

p-values correspond to comparisons of the distribution of characteristics among deaths (cases) and selected controls.

Deceased study participants were more likely male than female (1206 vs. 518 deaths, 13.3 vs. 8.1 deaths/1000 person years at risk in the NLST CT arm cohort, $p < 0.001$). Death from cardiovascular disease, and specifically coronary heart disease, was also more frequent in men (3.6 vs. 1.9 CVD deaths and 2.4 vs. 1.1 CHD deaths per 1000 person years, $p < 0.001$ in both cases). There was, however, no significant difference in death rates between men and women for extracardiac vascular causes of death (1.1 vs. 0.9 deaths/1000 person years, $p = 0.153$).

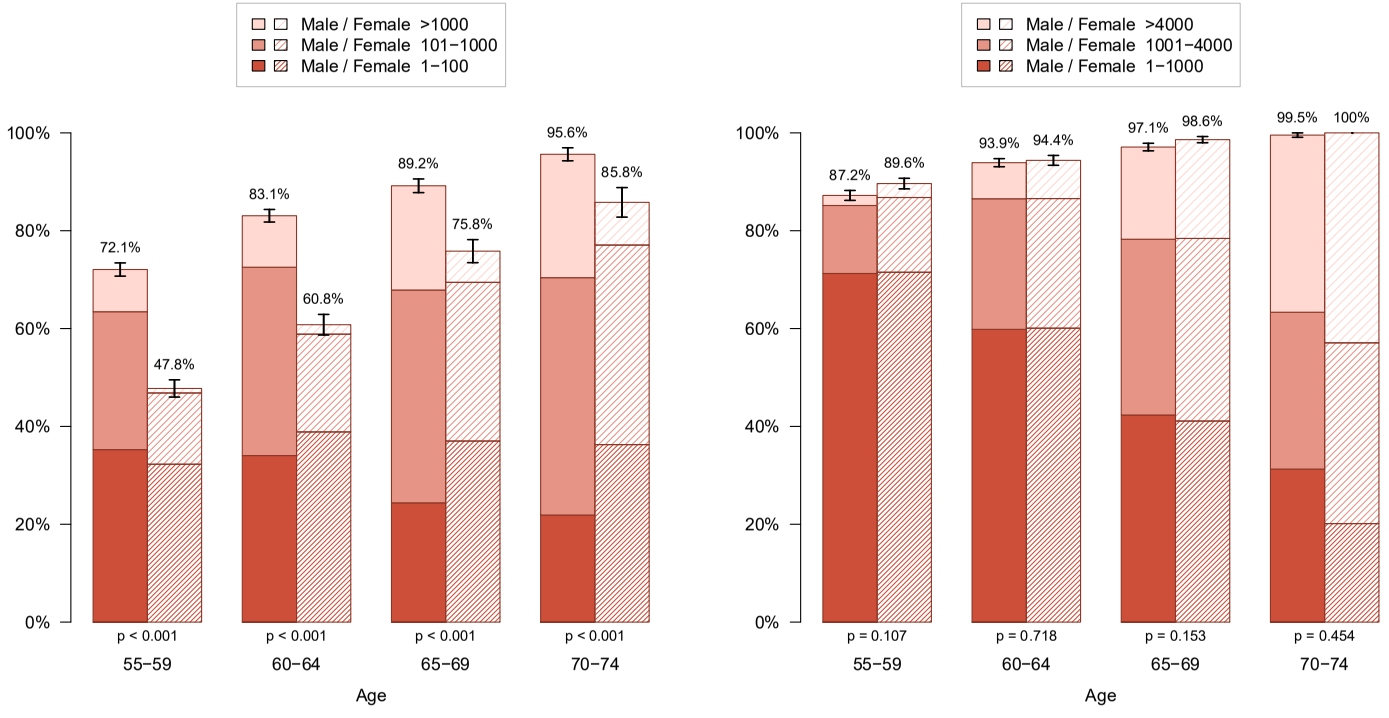
3.3.2 Prevalence and extent of arterial calcification

Arterial calcification was present in a substantial number of participants: 72 % had calcified coronary arteries and 93 % had thoracic aorta calcification (after resampling to the NLST CT arm cohort). Both prevalence and extent of both CAC and TAC increased with age. The prevalence of CAC ranged from 62 % in participants aged 55–59 to 92 % in participants aged 70–74. The prevalence of TAC was higher across all age groups with presence of TAC in 88 % of the participants at age 55–59 and in virtually all participants at age 70–74 (99 %).

CAC was significantly more often prevalent in men than in women of similar age. The extent of CAC was also greater in men throughout all age groups. Women had on average as much CAC as about 10 year younger men (Figures 3.2a and 3.3a). Conversely, TAC was equally likely present in male and female participants of similar age (Figure 3.2b). The extent of TAC was comparable for men and women of similar age, apart from the oldest participants aged 70–74, among which women had more TAC than men (Figure 3.3b).

3.3.3 Zero CAC scores

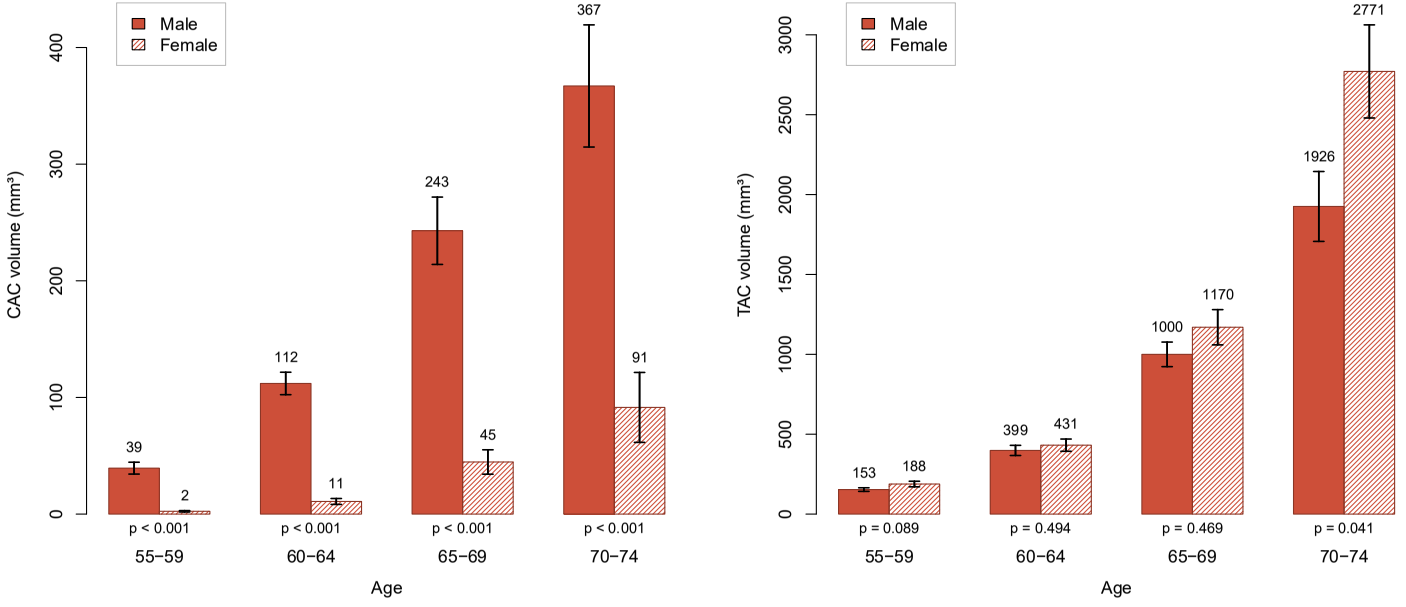
TAC was present in almost the entire study population, however, 19 % of the male and 40 % of the female participants (after resampling to the NLST CT arm cohort) had no visible CAC despite their advanced age and extensive smoking history. Subjects with a zero CAC score were more likely those who did not have TAC (prevalence of TAC of 77 % vs. 96 % in men and 88 % vs. 97 % in women). In both sexes, a zero CAC score was associated with younger age and fewer pack years, but not with smoking cessation. Marital status played a role only in female participants, who had more often a zero CAC score when they were married. Overall, participants with a zero CAC score were healthier than those who developed CAC (Table 3.2). Accordingly, death rates estimated for the NLST CT arm cohort were consistently lower for participants with CAC = 0 compared to those with CAC > 0 at baseline (all $p < 0.001$): there were approximately 7.2 (95 % CI: 5.9–8.6) vs. 15.2 (95 % CI: 14.1–16.2) deaths/1000 person years at risk among male participants and 4.5 (95 % CI: 3.6–5.3) vs. 10.5 (95 % CI: 9.4–11.6) deaths/1000 person years at risk among female participants. Specifically, death from CVD occurred



(a) Coronary artery calcification

(b) Thoracic aorta calcification

Figure 3.2: Prevalence of calcification in the coronary arteries and thoracic aorta per age group and sex. Data were measured in the study sub-cohort, but resampled to the full NLST CT screening cohort. Error bars correspond to standard errors, p-values to comparisons per age group among men and women. Bars are subdivided into risk categories, i.e., indicate the fraction of subjects in certain Agatston score ranges. The remaining percentage per bar corresponds to subjects without prevalent calcification, i.e., (a) CAC = 0 or (b) TAC = 0.



(a) Coronary artery calcification

(b) Thoracic aorta calcification

Figure 3.3: Median calcium volume per age group and sex. Data were measured in the study sub-cohort, but resampled to the full NLST CT screening cohort. Error bars correspond to standard errors, p-values to comparisons per age group among men and women.

Table 3.2: Baseline characteristics according to presence of coronary calcification and sex

	Men			Women		
	CAC = 0 (n = 598)	CAC > 0 (n = 2955)	p-value	CAC = 0 (n = 802)	CAC > 0 (n = 1363)	p-value
Age (years)	58 ± 6	62 ± 9	< 0.001	59 ± 6	62 ± 8	< 0.001
BMI (kg/m^2)	27.3 ± 5.4	27.4 ± 5.7	0.463	26.5 ± 7.2	26.6 ± 7.2	0.772
Pack-years	47 ± 27	55 ± 35	< 0.001	43 ± 22	47 ± 24	< 0.001
Current smoker	304 (51)	1444 (49)	0.380	419 (52)	716 (53)	0.897
Married	437 (73)	2180 (74)	0.723	460 (57)	700 (51)	0.007
Race			< 0.001			0.194
White	523 (88)	2696 (91)		733 (91)	1268 (93)	
Black	43 (7)	101 (3)		46 (6)	55 (4)	
Other/unknown	32 (5)	158 (5)		23 (3)	40 (3)	
Education			0.078			0.002
Up to high school diploma/GED	149 (25)	872 (30)		240 (30)	506 (37)	
Up to bachelor degree	350 (59)	1566 (53)		440 (55)	704 (52)	
Graduate school	87 (15)	460 (16)		96 (12)	121 (9)	
Other/unknown	12 (2)	57 (2)		26 (3)	32 (2)	
Disease history						
Heart disease	20 (3)	621 (21)	< 0.001	37 (5)	165 (12)	< 0.001
Stroke	10 (2)	117 (4)	0.006	18 (2)	52 (4)	0.046
Diabetes	45 (8)	423 (14)	< 0.001	34 (4)	137 (10)	< 0.001
Hypertension	151 (25)	1161 (39)	< 0.001	199 (25)	554 (41)	< 0.001
TAC volume (mm^3)	81 ± 266	709 ± 1864	< 0.001	161 ± 476	932 ± 2179	< 0.001
Cause of death						
CHD	14 (2)	202 (7)	< 0.001	8 (1)	60 (4)	< 0.001
CVD other than CHD	6 (1)	96 (3)	0.003	12 (2)	45 (3)	0.011
Respiratory illness	10 (2)	103 (4)	0.021	10 (1)	42 (3)	0.007
Cancer	72 (12)	446 (15)	0.054	66 (8)	187 (14)	< 0.001
All causes	126 (21)	1080 (36)	< 0.001	118 (15)	400 (29)	< 0.001

Values are mean ± standard deviation or n (%). Disease history was self-reported at trial entry.

at rates of 1.3 (95 % CI: 0.7–1.9) vs. 4.5 (95 % CI: 4.0–5.1) and 0.8 (95 % CI: 0.5–1.2) vs. 2.9 (95 % CI: 2.3–3.4) deaths/1000 person years in male and female participants, respectively.

3.3.4 Survival analysis

Inclusion of CAC and TAC significantly improved the c-statistic for prediction of all outcomes over models without these variables in both sexes (Table 3.3). The improvement, however, was not significantly greater in one of the sexes for any outcome (all $p > 0.05$). Increased levels of both CAC and TAC were associated with increased risk of all-cause and CVD mortality in men and women. CAC was more strongly associated with CVD mortality in women, while large amounts of TAC were more strongly associated with CVD mortality in men than in women. However, these differences did not reach statistical significance (all $p > 0.05$). The association with CHD mortality was statistically significant for CAC in both sexes and for TAC only in men ($p = 0.007$), but not in women ($p = 0.286$). Conversely, increased levels of TAC were associated with increased risk of extracardiac CVD mortality in both sexes, but increased levels of CAC were not.

3.4 Discussion

We investigated sex differences in arterial calcification visible on chest CT and in the associated risks independent of cardiovascular risk factors in a population of heavy smokers eligible for lung cancer screening. Men had more often and more severely calcified coronary arteries. Calcification of the thoracic aorta was vastly prevalent in both sexes, with a tendency to be more pronounced in women. Survival analysis revealed that high CAC scores correspond to higher hazard ratios in women for all outcomes. TAC scores demonstrated incremental predictive value over CAC and other risk factors, especially for prediction of mortality from cardiovascular causes other than CHD and for all-cause mortality.

3.4.1 Comparison with previous studies

It is known that men have a higher CAC burden than women.[84–87] The present study confirms that this also holds for lung cancer screening participants, in which a high calcification burden can be expected in both sexes since eligibility for screening requires the presence of two major risk factors for atherosclerosis, advanced age and extensive smoking history. In other publications, lower age stratified median CAC scores were measured, e.g., in the Multi-Ethnic Study of Atherosclerosis (MESA)[86] as well as in a study in a large population of self-referred asymptomatic subjects.[84] The incidence of CAC and the increase of CAC scores with age has been reported to lag behind in women by 10–20 years.[84,86,88,89] We observed in this population of heavy smokers a

Table 3.3: Adjusted hazard ratios and 95 % confidence intervals for different outcomes for different levels of CAC and TAC. Data are selection-weighted and adjusted for age, smoking status, pack-years, BMI, race, education, marital status, and self-reported history of heart disease or stroke, diabetes, and hypertension. The models contain both CAC and TAC as variables. Extracardiac CVD mortality is defined as death from a cardiovascular cause other than CHD. Cutpoints for CAC and TAC categories refer to Agatston scores. Improvements in C-statistic refers to comparison with a baseline model with risk factors, but without CAC and TAC.

Men	CVD mortality	CHD mortality	Extracardiac CVD mortality	All-cause mortality
No. of deaths	318	216	102	1206
CAC				
0	1.00	1.00	1.00	1.00
1–100	1.51 (0.88–2.59)	1.43 (0.73–2.81)	1.58 (0.64–3.93)	1.16 (0.93–1.45)
101–1000	2.13 (1.26–3.62)	2.17 (1.13–4.19)	1.93 (0.78–4.75)	1.30 (1.04–1.61)
>1000	2.76 (1.56–4.88)	3.66 (1.82–7.34)	1.31 (0.48–3.58)	1.72 (1.34–2.21)
p-value for trend	< 0.001	< 0.001	0.812	< 0.001
TAC				
0–1000	1.00	1.00	1.00	1.00
1001–4000	1.58 (1.18–2.11)	1.23 (0.86–1.75)	2.58 (1.54–4.32)	1.42 (1.22–1.64)
>4000	2.47 (1.75–3.49)	1.77 (1.17–2.68)	4.58 (2.45–8.54)	1.79 (1.49–2.15)
p-value for trend	< 0.001	0.007	< 0.001	< 0.001
C-statistic (%)	73.4 (72.5–74.3)	75.3 (74.3–76.3)	71.8 (70.1–73.5)	70.4 (69.9–70.9)
Baseline improvement	+2.9 (1.6–4.2)	+2.6 (1.1–4.1)	+5.2 (2.8–7.6)	+1.3 (0.5–2.1)
p-value	< 0.001	< 0.001	< 0.001	< 0.001

Women	CVD mortality	CHD mortality	Extracardiac CVD mortality	All-cause mortality
No. of deaths	125	68	57	518
CAC				
0	1.00	1.00	1.00	1.00
1–100	1.87 (1.04–3.36)	2.30 (0.91–5.80)	1.62 (0.75–3.47)	1.43 (1.11–1.83)
101–1000	2.43 (1.29–4.56)	4.52 (1.75–11.68)	1.18 (0.48–2.90)	1.56 (1.18–2.07)
>1000	3.74 (1.62–8.63)	5.81 (1.76–19.14)	2.46 (0.73–8.36)	2.20 (1.44–3.36)
p-value for trend	0.001	< 0.001	0.442	< 0.001
TAC				
0–1000	1.00	1.00	1.00	1.00
1001–4000	1.62 (1.02–2.56)	1.44 (0.77–2.67)	1.83 (0.94–3.59)	1.47 (1.18–1.84)
>4000	1.85 (1.05–3.26)	1.49 (0.69–3.19)	2.31 (1.00–5.35)	1.93 (1.46–2.56)
p-value for trend	0.028	0.286	0.042	< 0.001
C-statistic (%)	77.2 (76.0–78.4)	81.5 (80.0–83.0)	75.4 (73.7–77.1)	71.9 (71.1–72.7)
Baseline improvement	+2.7 (1.8–3.6)	+2.6 (1.5–3.7)	+2.7 (1.4–4.0)	+1.7 (1.1–2.3)
p-value	< 0.001	0.002	0.003	< 0.001

similar age difference of about 10 years between male and female individuals with similar CAC scores.

There is currently no consensus whether a similar sex difference exists for calcification of the aorta. Results of previous studies are incoherent as some studies reported higher prevalence in women,[76,90] some reported higher prevalence in men,[91,92] and some reported no significant difference between the sexes.[88,93] We observed no significant difference in most age groups for both prevalence and extent of TAC. The oldest female participants, however, were more likely to have larger amounts of TAC than male participants of similar age (70–74 years), even though women were less excessive smokers. Remarkably, TAC was present in almost all individuals of this age.

A large number of studies have demonstrated that CAC is a strong independent predictor of cardiovascular events and mortality in both sexes.[27] Studies that examined sex differences in the ability of CAC to stratify cardiovascular risk are scarce, but commonly reported equal power in both sexes.[94,95] However, higher event rates have been reported for women with high CAC scores compared to men with similar scores.[96,97] Our results are in line with these findings inasmuch as CAC was associated with cardiovascular mortality in both sexes independent from other risk factors and as hazard ratios were consistently higher in women.

In contrast to many other studies,[27] a zero CAC score was not associated with better survival than a low CAC score. Motion artifacts may inhibit the distinction of true zero CAC scores from low CAC scores in ungated chest CT. Furthermore, while CVD mortality was significantly lower in subjects who had no CAC at baseline, it was still clearly present regardless of sex. This suggests that there are both CAC and non-CAC related causes of cardiovascular mortality in heavy smokers, which limits the value of a zero CAC score as predictor of the absence of CVD in this population.[98] However, it is remarkable that a sizeable proportion of the lung cancer screening population, who is at least 55 years old and has 30 or more pack years of cigarette smoking history, has not developed CAC.

The value of TAC in the prevention of cardiovascular events is currently not well defined. Most studies that investigated the predictive value of TAC for cardiovascular events and mortality derived TAC scores from cardiac CT scans. Unlike chest CT, these visualize only part of the thoracic aorta so that near the carotid arteries lesions may have been missed that are potentially relevant for cerebrovascular events. A study in chest CT scans from the Dutch-Belgian lung cancer screening trial NELSON showed a stronger association with extracardiac events for TAC than for CAC.[17] Likewise, TAC but not CAC was associated with extracardiac CVD mortality in both sexes in our study. For prediction of CHD, the MESA showed incremental value of TAC over CAC, but only in women.[99] Conversely, our study showed an association of TAC with CHD mortality in addition to CAC only in men. NLST participants, however, were on average substantially older and heavier smokers than MESA participants.

3.4.2 Strengths and limitations of this study

The present study is based on a sub-cohort of the NLST, which provides standardized imaging data for a population at increased cardiovascular risk consisting of a large number of male and female subjects. A particular strength of this study is the substantial number of deaths that occurred in this population, which facilitated robust survival analysis even after stratification by sex. Another strength is the use of chest CT scans for evaluation of calcifications of the thoracic aorta, which ensured that calcifications of the aortic arch are included. This was often not the case in studies based on cardiac CT.

This study has several limitations. First, the findings of this study are limited to heavy smokers. In the general population, prevalence and severity of CAC and TAC are likely to be lower. Second, we derived calcium scores from ungated CT, which may have introduced an overestimation of the calcium burden compared to gated scans.[100] Automatic calcium scoring may have additionally resulted in more false positive than false negative detections, which may have contributed to an overestimation especially of the prevalence of arterial calcification.[79] Third, only fatal events were recorded in the NLST, but sex differences in prediction of non-fatal cardiovascular events would also be highly relevant to better inform treatment decisions. Fourth, because the NLST was primarily targeting lung cancer, lipid measurements were not performed, and intake of lipid-lowering medication or other preventive measures were not recorded. Fifth, the findings of this study are of retrospective nature and do therefore not provide evidence for benefits from sex specific preventive efforts in the lung cancer screening population.

3.4.3 Clinical implications

Neither CAC nor TAC are currently part of guidelines for lung cancer screening with low-dose chest CT. However, reporting of CAC on ungated chest CT scans is increasingly recommended.[101] The relevance of these findings is debated since virtually all heavy smokers eligible for lung cancer screening are at least in the intermediate risk group for which preventive efforts are recommended regardless of their calcium score.[102] The findings of this study, however, indicate that both CAC and TAC scores can support the identification of subjects who will experience an adverse cardiovascular event. Especially if healthcare providers also take sex differences into account, calcium scores can be useful biomarkers for individualized preventive efforts in males and females. In the high risk population of heavy smokers, calcium scores can potentially help to identify individuals at highest risk for whom more aggressive and expensive preventive treatment beyond lipid management, such as anti-inflammatory drugs,[103] might be mandated. Additionally, calcium scores could also play a role in lowering the burden of medication for low risk heavy smokers who would otherwise be statin-eligible. Healthcare providers involved in lung cancer screening need to be aware that

high CAC scores occur less frequently in women than in men, but also need to be aware that this difference does not exist for calcification of the aorta.

3.5 Perspectives

COMPETENCY IN MEDICAL KNOWLEDGE The amount of coronary artery and thoracic aorta calcium visible on lung screening CT scans provides incremental prognostic value over demographic cardiovascular risk factors. Coronary artery calcium is more common in men, but is also present in a substantial number of female heavy smokers. Calcification of the thoracic aorta can be expected in nearly the entire lung cancer screening eligible population, with no significant differences between men and women, and provides a more powerful predictor of extracardiac CVD mortality than coronary artery calcium.

TRANSLATIONAL OUTLOOK The utilization of calcium scores for cardiovascular risk assessment, especially in addition to lung cancer screening, could greatly benefit from standardized reporting. Prospective studies are needed to develop evidence-based guidelines for management of prevalent cardiovascular findings in the high risk population of heavy smokers.

Acknowledgments

We are grateful to the United States National Cancer Institute (NCI) for providing access to NCI's data collected by the National Lung Screening Trial. The statements contained herein are solely ours and do not represent or imply concurrence or endorsement by NCI.

CHAPTER 4

Automatic vertebra segmentation and identification

BASED ON: N. Lessmann, B. van Ginneken, P. A. de Jong, and I. Išgum. "Iterative fully convolutional neural networks for automatic vertebra segmentation and identification", *Medical Image Analysis*, vol. 53 (2019), pp. 142–155.

Abstract

Precise segmentation and anatomical identification of the vertebrae provides the basis for automatic analysis of the spine, such as detection of vertebral compression fractures or other abnormalities. Most dedicated spine CT and MR scans as well as scans of the chest, abdomen or neck cover only part of the spine. Segmentation and identification should therefore not rely on the visibility of certain vertebrae or a certain number of vertebrae. We propose an iterative instance segmentation approach that uses a fully convolutional neural network to segment and label vertebrae one after the other, independently of the number of visible vertebrae. This instance-by-instance segmentation is enabled by combining the network with a memory component that retains information about already segmented vertebrae. The network iteratively analyzes image patches, using information from both image and memory to search for the next vertebra. To efficiently traverse the image, we include the prior knowledge that the vertebrae are always located next to each other, which is used to follow the vertebral column. The network concurrently performs multiple tasks, which are segmentation of a vertebra, regression of its anatomical label and prediction whether the vertebra is completely visible in the image, which allows to exclude incompletely visible vertebrae from further analyses. The predicted anatomical labels of the individual vertebrae are additionally refined with a maximum likelihood approach, choosing the overall most likely labeling if all detected vertebrae are taken into account. This method was evaluated with five diverse datasets, including multiple modalities (CT and MR), various fields of view and coverages of different sections of the spine, and a particularly challenging set of low-dose chest CT scans. For vertebra segmentation, the average Dice score was 94.9 ± 2.1 % with an average absolute symmetric surface distance of 0.2 ± 0.1 mm. The anatomical identification had an accuracy of 93 %, corresponding to a single case with mislabeled vertebrae. Vertebrae were classified as completely or incompletely visible with an accuracy of 97 %. The proposed iterative segmentation method compares favorably with state-of-the-art methods and is fast, flexible and generalizable.

4.1 Introduction

Segmentation and identification of the vertebrae is often a prerequisite for automatic analysis of the spine, such as detection of vertebral fractures,[104] assessment of spinal deformities,[105] or computer-assisted surgical interventions.[106] Automatic spine analysis can be performed with a large variety of tomographic scans, including dedicated spine scans but also scans of the neck, chest or abdomen that incidentally cover part of the spine. A generic vertebra segmentation algorithm therefore needs to be robust with respect to different image resolutions and different coverages of the spine. This especially means that no assumptions should be made about the number of visible vertebrae and their anatomical identity, i.e., to which section of the spine they belong. Vertebra segmentation is therefore essentially an instance segmentation problem with an *a priori* unknown number of instances (*i.e.* vertebrae). However, in contrast to generic instance segmentation the individual instances are not independent of each other. The instances are known to be located in close proximity to each other in the image, forming together the vertebral column. We propose to approach vertebra segmentation with an instance segmentation algorithm that explicitly incorporates this prior knowledge to locate instances, but that makes no further assumptions.

Approaching vertebra segmentation as an instance segmentation problem entails treating all vertebrae as instances of the same class of objects. However, an anatomical identification of the segmented vertebrae is often also needed, for instance, for further analysis steps or for reporting purposes. Especially in images originally not intended for spine imaging, anatomical labeling of the vertebrae can be challenging due to variations in the field of view. These variations lead to variable coverage of the spine and also of structures that provide anatomical cues for identification of the vertebrae, such as the ribs or the sacrum. Additionally, neighboring vertebrae often have similar shape and appearance so that independent labeling of each vertebra may result in mistakes. Vertebra identification therefore requires a global rather than a per-instance approach to ensure an overall plausible, anatomically correct labeling.

Another challenge inherent to an instance segmentation approach is the identification of partially visible instances. While occlusion is a typical problem in two-dimensional but not in three-dimensional images, some vertebrae may be only partially visible due to the limited field of view of the scan. If these incompletely visible vertebrae are included in subsequent analyses that are based on the obtained vertebra segmentations, such as measurement of vertebral heights for detection and classification of vertebral compression fractures,[107] their results may be unreliable. Therefore, incompletely visible instances need to be either ignored or explicitly identified as incomplete so that they can be excluded from subsequent analyses.

In this chapter, we propose an iterative instance-by-instance segmentation approach for vertebra segmentation based on a fully convolutional neural network. This network performs vertebra detection, segmentation, anatomical identification and

classification of their completeness concurrently and therefore presents an entirely supervised approach that can be trained end-to-end. While we propose to attempt a per-instance identification of the individual vertebrae together with the segmentation, the labeling is subsequently adjusted taking all segmented vertebrae into account. In contrast to previous approaches, the presented method can be used for any imaging modality, any field of view and any number and type (cervical, thoracic, lumbar) of visible vertebrae because it avoids explicit modeling of shape and appearance of the vertebrae and the vertebral column. We evaluate these claims using a diverse selection of datasets, including scans from different modalities (CT and MR), various fields of view, cases with severe compression fractures and a particularly challenging set of low-dose chest CT.

4.2 Related work

While a few other methods have been published that address both vertebra segmentation and identification,[108–112] the majority of methods in the literature focused on one of these problems. The existing literature is therefore reviewed separately for vertebra segmentation and for vertebra identification. We also briefly review literature on general instance segmentation.

4.2.1 Vertebra segmentation

Vertebra segmentation has been approached predominantly as a model-fitting problem using statistical shape models and its variants, most often active shape models and shape-constrained deformable models.[108,111,113–126] Other approaches have been based on atlases,[127] level-sets with shape priors[126,128] and active contours.[129,130]

More recently, machine learning has been increasingly used for vertebra segmentation. Kelm et al.[109] used an iterative variant of marginal space learning to find bounding boxes for the intervertebral discs, which were used to initialize and guide vertebra segmentations based on Markov random fields and graph cuts. Zukić et al.[131] applied the Viola-Jones object detection framework based on Adaboost to find bounding boxes for the vertebral bodies, which were subsequently segmented by inflating a mesh from the center of each vertebral body. Chu et al.[110] used random forest regression to detect the center of the vertebral bodies and used these to define regions of interest in which vertebrae were segmented using random forest voxel classification. A similar method was proposed by Suzani et al.,[111] who used a multilayer perceptron to regress the distance to the nearest center of a vertebral body. The detected locations were used to initialize an adaptive shape model for segmentation of the vertebral bodies. Mirzaalian et al.[121] also combined machine learning and shape models by using a probabilistic boosting-tree classifier for boundary detection, which was used to adapt a surface mesh to the vertebrae in combination with a statistical shape model. The shape model

was used for initialization of the mesh and to impose shape constraints. Korez et al.[119] used a convolutional neural network (CNN) to generate probability maps for the vertebral bodies and used these maps to guide a deformable surface model to segment the vertebral bodies.

Even though the aforementioned methods contain a machine learning component beyond statistical modeling, machine learning was primarily used for vertebra detection and thus merely for initialization of the segmentation, which itself was performed with other methods. Many of the most recently published vertebra segmentation methods, however, are based on deep learning and have replaced explicit modeling of the vertebral shape and appearance with convolutional and recurrent neural networks. For instance, Sekuboyina et al.[112] segmented the lumbar vertebrae in 2D sagittal slices using a multiclass CNN for pixel labeling. As a prior step, a simple multilayer perceptron estimated a bounding box of the lumbar region to identify the region of interest in the image. In subsequent work, Sekuboyina et al.[132] used a patch-based 3D network for voxel classification in the entire image and additionally a 2D network to predict a low-resolution mask for the vertebral column, which was used to remove false positives outside the spinal region. Similarly, Janssens et al.[133] relied on two consecutive networks, first using a regression CNN to estimate a bounding box of the lumbar region, followed by a classification CNN to perform voxel labeling within that bounding box to segment the lumbar vertebrae. In our preliminary work,[134] we also applied a two-stage approach in which vertebrae were first segmented in downsampled images using an iterative strategy. The image was repeatedly analyzed by a CNN to segment the vertebrae one after the other. A second network analyzed the full resolution images to refine the low-resolution segmentations. Even though all of these approaches relied on CNNs for segmentation of the vertebrae, they retained the separation into a detection and a segmentation task, and consequently used two dedicated networks.

4.2.2 Vertebra identification

Anatomical identification of individual vertebrae has been mainly approached with one of three strategies: with appearance and shape models,[108,135,136] with machine learning based on hand-crafted features[109–111,135,137–139] and with deep neural networks.[112,125,133,134,140–144] Most of these approaches combined a rough labeling of the vertebrae, typically by performing voxel classification or regression of vertebral centroids or bounding boxes, with a global model to refine the individual predictions, to discard outliers and to find an overall plausible solution. These models have often been graphical models, such as hidden Markov models[110,135] and Markov random fields,[138] statistical shape models[111,139,141] or recurrent neural networks.[143,144] Several methods relied on detection of a reference vertebra, such as the fifth lumbar vertebra (L5), and labeled the detected vertebrae relative to this reference vertebra.[134,140,142] Multiclass CNNs for combined segmentation and identification through voxel classification were

used in scans with a fixed field of view and a limited number of vertebrae, e.g., lumbar spine CT scans.[112,133] Furthermore, probabilistic modeling has been used to calculate a likelihood score for each possible labeling configuration based on shape or appearance similarities or spatial relationships of the vertebrae.[108,109,137,141]

4.2.3 Instance segmentation

Generic instance segmentation frameworks based on convolutional neural networks, such as Mask R-CNN,[145] typically split the task into a detection and a segmentation task. Many of the above discussed publications on vertebra segmentation have used this approach as well, but have usually imposed constraints on the number of instances or other features. Recurrent networks have been used in a similar fashion by localizing individual instances based on attention and memory mechanisms, often also based on instance detection through region proposals and subsequent segmentation.[146–149] Other approaches have relied on clustering instead of explicit instance detection, transforming images into abstract feature representations in which individual instances were detected as individual clusters.[150–153] An iterative approach without region proposals or recurrent connections has been proposed by Li et al.,[148] who repeatedly fed the same image through a convolutional semantic segmentation network together with the previous prediction map.

4.3 Methods

We propose a vertebra segmentation and identification method based on a single fully convolutional neural network (FCN) that performs multiple tasks concurrently. In contrast to existing methods, this avoids a multi-stage process with successive instance detection and segmentation, or segmentation and instance separation steps. Other existing generic instance segmentation methods with 2D deep neural networks often do not generalize well to 3D image volumes because they analyze the entire image at once, which is currently not feasible with typical CT or MR volumes. We therefore apply a patch-based vertebra-by-vertebra segmentation approach in which the image is analyzed in patches large enough to contain at least one vertebra. The network segments a single vertebra in this patch and the anatomical knowledge that the following vertebra must be located in close proximity is used to reposition the patch for segmentation of the following vertebra.

Next to this iterative inference strategy, our approach consists of four major components. The central component is a *segmentation network* that segments voxels from a 3D image patch by binary classification of all voxels in the patch. To enable this network to segment only voxels belonging to a specific instance rather than all vertebrae visible in the patch, we augment the network with an *instance memory* that informs the network about already segmented vertebrae. The network uses this information

to always segment only the following not yet segmented vertebra. Once a vertebra is fully segmented, the instance memory is updated, which triggers the network in the next iteration to ignore this vertebra and focus on the next vertebra instead. The third component is an *identification sub-network* that predicts the anatomical label of each detected vertebra. The fourth component is a *completeness classification sub-network* that is added to the network to distinguish between completely visible and partially visible vertebrae. The full network architecture is illustrated in Figure 4.1. Please note that the aforementioned components, referred to as networks and sub-networks, together form a single network.

In the following, we will first describe the segmentation network (Section 4.3.1) and the instance memory (Section 4.3.2). These two components form the basis for the proposed iterative instance segmentation strategy (Section 4.3.3). Additional network parts without influence on the iterative segmentation are described afterwards: regression of the anatomical label of each vertebra (Section 4.3.4) and classification of its completeness (Section 4.3.5). Finally, the loss function to optimize the network for all three tasks and the training process are detailed (Section 4.3.6).

4.3.1 Segmentation network

The segmentation component of the network is a FCN that predicts a binary label for each voxel in an image patch. This label indicates whether the voxel belongs to the current instance or not. We used a patch size of 128 × 128 × 128 voxels, which is large enough to cover an entire vertebra. To ensure that all patches have the same resolution, even though the analyzed images may have different resolutions, we resample all input images to an isotropic resolution of 1 mm × 1 mm × 1 mm prior to the segmentation. The obtained segmentation masks are resampled back to the original image resolution using nearest neighbor interpolation.

The architecture of the segmentation network is inspired by the U-net architecture,[68,154] i.e., the network consists of a compression and an expansion path with intermediate skip connections. We use a constant number of filters in all layers and added batch normalization as well as additional padding before all convolutional layers to obtain segmentation masks of the same size as the image patches. The segmentation masks are additionally refined in CT scans by removing voxels below 200 HU from the surface of each vertebra.

4.3.2 Instance memory

In the proposed iterative segmentation scheme, which is described in detail in the following section, the network segments one vertebra after the other, one at a time. The purpose of the instance memory is to remind the network of vertebrae that were already segmented in previous iterations so that it can target the next not yet segmented vertebra. This memory is a binary flag for each voxel of the input image that indicates

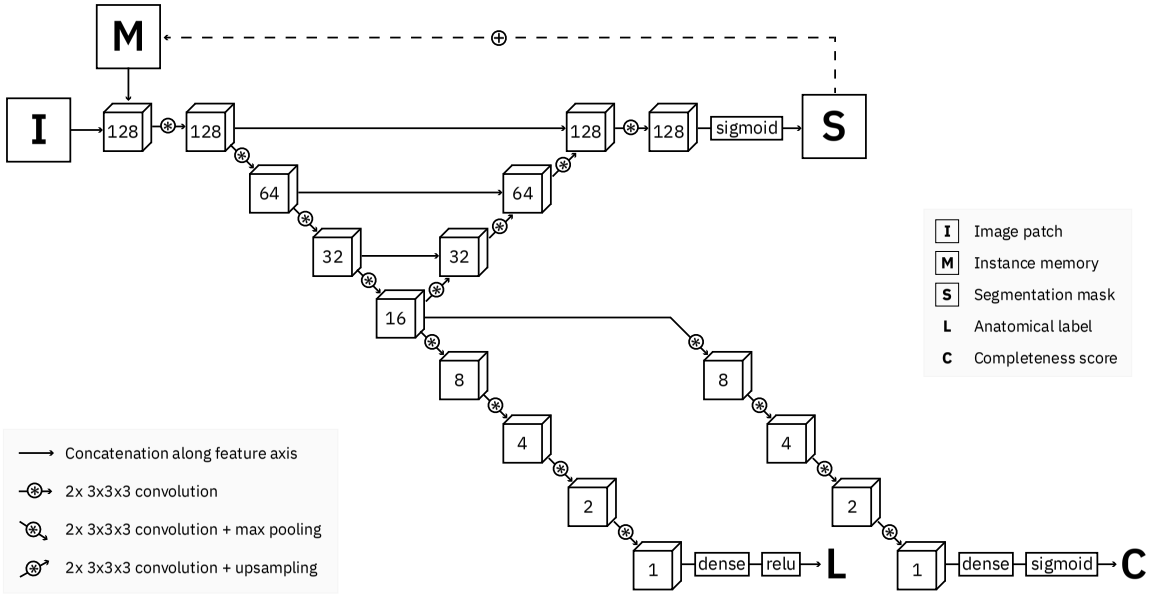


Figure 4.1: Schematic drawing of the network architecture. I, M and S are 3D volumes, L is the predicted label in form of a single value, and C is the predicted probability for complete visibility. Cubes represent 3D feature maps with 84 channels in the path from I and M to S and 48 features maps in the two additional compression paths to L and C. Exceptions are the first cube after I and M, which has two channels as a result of the concatenation of I and M, and the cube before S, which has only one channel. Both dense layers map 48 features to a single value. The number on each cube indicates the size of the feature map (a cube "128" corresponds to a feature map of 128 × 128 × 128 voxels).

whether the voxel has been labeled as vertebra in any of the previous iterations. Binary rather than probabilistic flags are used for simplicity, but probabilistic flags could prove useful in future extensions that allow voxels labeled as part of a vertebra to be unlabeled or relabeled in later iterations. Together with an image patch, the network receives a corresponding memory patch as input. These are fed to the network as a two-channel input volume.

4.3.3 Iterative instance segmentation

The iterative segmentation process is illustrated in Figure 4.2. This process follows either a top-down or bottom-up scheme, i.e., the vertebrae are not segmented in random order but successively from top to bottom, or vice versa. The network learns to infer from the memory patch which vertebra to segment in the current patch. If the memory is empty, i.e., no vertebra has been detected yet, the network segments the top-most or bottom-most vertebra that is visible in the patch, depending on the chosen direction of traversal. Otherwise, the network segments the first following not yet segmented vertebra, even if multiple unsegmented vertebrae are visible. Other vertebrae that are the second or third not yet segmented vertebra in the direction of traversal are disregarded until they become the first not yet segmented vertebra themselves in a later iteration. Each instance of the network therefore has a fixed direction of traversal.

The patch size is chosen large enough to always contain part of the following vertebra when a vertebra is in the center of the patch. This enables utilizing prior knowledge about the spatial arrangement of the individual instances to move from vertebra to vertebra. The FCN iteratively analyzes a single patch centered at $\mathbf{x}_t$, where t denotes the iteration step. Initially, the patch is moved over the image in a sliding window fashion with constant step size $\Delta\mathbf{x}$, searching for the top-most vertebra when using a top-down approach, or the bottom-most vertebra when using a bottom-up approach. As soon as the network detects a large enough fragment of vertebral bone, at least $v_{min} = 1000$ voxels $\hat{=} 10\,\text{mm}^3$ in our experiments, the patch is moved toward this fragment. The center of the bounding box of the detected fragment, referred to as $\mathbf{b}_t$, becomes the center of the next patch:

$$\mathbf{x}_{t+1} = \begin{cases} \mathbf{x}_t + \Delta\mathbf{x}, & \text{if } v_t < v_{min} \\ [\,\mathbf{b}_t\,], & \text{otherwise} \end{cases}$$

Even if initially only a small part of the vertebra is visible and detected in the patch, centering the following patch at the detected fragment ensures that a larger part of the vertebra becomes visible in the next iteration. Eventually, the entire vertebra becomes visible, in which case the patch position converges because no additional voxels are detected anymore that would affect $\mathbf{b}_t$, and hence $\mathbf{x}_{t+1}$. We detect convergence by comparing the current patch position $\mathbf{x}_t$ and the previous patch position $\mathbf{x}_{t-1}$, testing whether they still differ by more than δ_{max} on any axis. Occasionally, the patch position

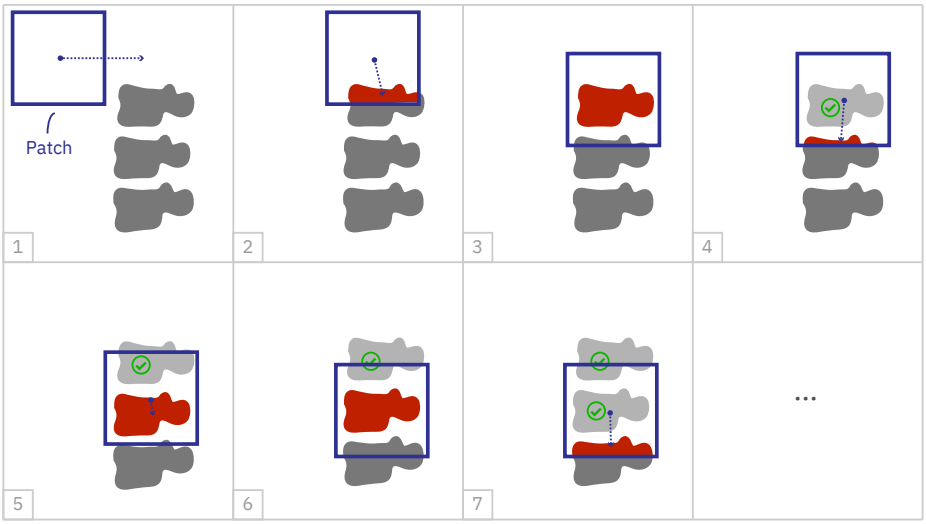


Figure 4.2: Illustration of the iterative instance segmentation and traversal strategy. The patch is first moving in a sliding window fashion over the image (1), until a fragment of vertebral bone is detected (2). The patch is then moved to the center of the detected fragment. This process is repeated until the entire vertebra becomes visible and the patch thus stops moving (3). The segmented vertebra is added to the instance memory and the same patch is analyzed again, now yielding a fragment of the following vertebra because the updated memory forces the network to ignore the previous vertebra (4). The patch is centered now at the detected fragment of the following vertebra and the process repeats (5-7). Animations of the iterative segmentation process are available as supplementary material online.

does not converge but keeps alternating between positions that differ slightly more than the threshold δ_{max}. We therefore limit the number of iterations per vertebra. When this limit is reached, we assume that the patch has converged to the position between the two previous patch positions and we accordingly move the patch to $\mathbf{x}_{t+1} = [\ (\mathbf{x}_t + \mathbf{x}_{t-1}) / 2\ ]$. In our experiments, δ_{max} was set to 2, $\Delta\mathbf{x}$ to half the patch size and the maximum number of iterations to 10.

Once the position has converged, the segmented vertebra is added to the output mask using a unique instance label and the segmentation mask obtained at this final position. Furthermore, the instance memory is updated. In the following iteration, the network analyzes the same patch again. The updated memory prompts the network to detect a fragment of the following vertebra and the patch is moved to the center of the detected new fragment, repeating the segmentation process for the next vertebra. Should no fragment of the following vertebra be immediately visible, traversal reverts to a sliding window motion until the next vertebra is found. The entire process continues until no further fragments are found, i.e., until all visible vertebrae are segmented.

4.3.4 Anatomical identification

For each detected vertebra, the network additionally predicts the anatomical label using an additional identification component that is added to the segmentation network. The U-net architecture, which provides the basis of the segmentation network, consists of a compression and a expansion path. These are commonly understood as recognition and segmentation paths, respectively. Since vertebra identification is a recognition task, the identification component is appended to the compression path as further compression steps. Essentially, it shares features with the segmentation network but further compresses the input patch into a single value. The vertebrae C1 to L5 are represented by integers 1 to 24 and the value 0 is used during training as ground truth value for patches that do not contain any vertebral bone. In the network, the label output unit is a rectified linear unit $f(x) = \max(0, x)$, which enables the network to produce any negative activation for patches without vertebral bone. Regression of the label as continuous value instead of classification with 25 output units and a softmax activation function ensures that the loss function penalizes predictions more strongly the further they deviate from the true label.

In order to find a plausible sequence of labels for the detected vertebrae, without duplicates or gaps, the predicted labels are interpreted as probabilities in a simple maximum likelihood estimation similar to Klinder et al..[108] Depending on how many vertebrae are detected, several labeling sequences are theoretically possible – only if all 24 vertebrae are visible in a scan there is only a single possible labeling sequence. For each possible sequence of labels, the average likelihood is calculated by interpreting the predicted labels as probabilities. For example, a regression output value of 22.8 is interpreted as 80 % probability for label 23 (L4) and 20 % probability for label 22 (L3), and as 0 % probability for any other label. In sequences in which this vertebra would be labeled as L3, it therefore contributes a probability of 20 % to the average. The sequence with maximal likelihood is finally used to labeled the detected vertebrae. This step is performed as a post-processing step when all vertebrae have been segmented from the image.

4.3.5 Classification of vertebra completeness

An obvious strategy for disregarding incompletely visible vertebrae in the segmentation process would be to train the network only with examples of fully visible vertebrae. However, the traversal scheme requires detection of vertebral fragments in the patches. We therefore choose to include partially visible vertebrae in the training data but to add a classification component to the segmentation network that classifies each segmented vertebra as complete or incomplete. Similar to the identification component, this completeness classification is a recognition and not a segmentation task. In the network, the classification path is therefore also a continuation of the compression path and has the same architecture as the identification path, with the difference that

the output is a sigmoid unit. The output value is thus a single value in $[0, 1]$, which indicates the probability that the vertebra is completely visible in the image patch. During traversal, vertebrae classified as incomplete are not added to the segmentation mask, but are added to the instance memory to facilitate further traversal because not only the last but also the first vertebrae are often incompletely visible. This additionally ensures that traversal can continue if a vertebra is mistakenly classified as incompletely visible. Depending on the application, these vertebrae could also be added to the segmentation mask, merely flagging them as incomplete but making them available to further analysis steps.

4.3.6 Training the network

During training, we derived the status of the instance memory from the reference segmentation masks. The patches used to train the network were forced to contain vertebral bone by randomly selecting in each iteration a scan and a vertebra visible in that scan, followed by random patch sampling within the bounding box of that vertebra. However, 25 % of the patches were selected randomly from anywhere in the training images. If these patches contained vertebral bone, it was added to the instance memory so that the network could also learn to produce empty segmentation masks for patches without vertebral bone or without unsegmented vertebral bone.

Due to the size of the input patches, the Nvidia Titan X GPUs with 12 GB memory that we used for training allowed processing of only single patches instead of minibatches of multiple patches. We therefore used Adam[62] for optimization with a fixed learning rate of 0.001 and an increased momentum of 0.99, which stabilizes the gradients. Furthermore, the network predicts labels for all voxels in the input patch and the loss term is accordingly not based on a single output value, but on more than two million output values per patch. The loss term $\mathscr{L}$ combines terms for segmentation, anatomical labeling and classification errors:

$$\mathscr{L} = \underbrace{\lambda \cdot \mathrm{FP}_{\mathrm{soft}} + \mathrm{FN}_{\mathrm{soft}}}_{\text{Segmentation error}} + \underbrace{\left| p_L - t_L \right|}_{\text{Labeling error}} + \underbrace{(-t_C \log p_C - (1 - t_C) \log(1 - p_C))}_{\text{Completeness classification error}}$$

We propose to optimize the segmentation by directly minimizing the number of incorrectly labeled voxels, i.e., the number of false positives and false negatives. This is similar to loss terms based on the Dice score,[155] which can be expressed as $2\,\mathrm{TP}/(2\,\mathrm{TP} + \mathrm{FP} + \mathrm{FN})$, but because the number of true positives is not part of the term values are more consistent across empty (TP = 0) and non-empty (TP $\gg$ 0) patches. Given an input patch I and for all voxels i binary reference labels t_i and probabilistic predictions p_i, differentiable expressions for the number of false positive and false negative predictions are:

$$\mathrm{FP}_{\mathrm{soft}} = \sum_{i \in I} \omega_i \cdot (1 - t_i)\, p_i \qquad \mathrm{FN}_{\mathrm{soft}} = \sum_{i \in I} \omega_i \cdot t_i\, (1 - p_i)$$

Here, w_i are weights used to assign more importance to the voxels near the surface of the vertebra. This aims at improving the separation of neighboring vertebrae.[68] The weights are derived from the distance d_i of voxel i to the closest point on the surface of the targeted vertebra: $\omega_i = \gamma \cdot \exp\left(-d_i^2/\sigma^2\right) + 1$. We used $\gamma = 8$ and $\sigma = 6$ in all experiments.

Additionally, the factor λ weights the cost of a false positive error relative to a false negative error, which is useful to counteract an imbalance between background and foreground voxels. In most segmentation problems, the number of background voxels is substantially larger than the number of foreground voxels and consequently a systematic false negative mistake is more favorable than a systematic false positive mistake. This can prevent the network from learning anything other than predicting an empty segmentation mask for any input. We therefore used an increasing value for λ starting from $\lambda_{\mathrm{min}} = 0.1$ and converging eventually at 1, following a sigmoidal shape:

$$\lambda(n) = \lambda_{\mathrm{min}} + \frac{1 - \lambda_{\mathrm{min}}}{1 + e^{-\vartheta(n)}} \quad \text{where} \quad \vartheta(n) = \frac{n - n_{\mathrm{max}}/2}{n_{\mathrm{max}}/10}$$

Here, n is the number of training iterations, i.e., backward passes, that have been performed. The definition of ϑ ensures that false positive and false negative errors are equally weighted after about half of the maximum number of training iterations n_{max}. The network is therefore initially biased towards making false positive errors rather than false negative errors, but this bias is reduced over time to ensure that the network will not converge in a state where it tends to oversegment the vertebra. We found that assigning weights to the individual error components that form the loss $\mathscr{L}$ was not necessary.

The labeling error is defined as the $\ell 1$ norm of the difference between predicted label and true label. The predicted labels are real numbers ≥ 0 while the true labels are integers 0 to 24. Using the absolute difference of these two values as loss function penalizes large errors more strongly than small errors. This is especially beneficial because small deviations can be more likely corrected in the label refinement step.

The completeness classification error is defined as the binary cross entropy between the true label t_c, which is a binary value that indicates whether the vertebra is completely visible in the patch, and the predicted probability for complete visibility p_c. In our experiments, vertebrae were considered completely visible in a patch when they were manually marked as completely visible in the scan and when not more than 2 % of the volume of the reference segmentation of the vertebra was not contained in the patch. This allows for some tolerance as manual identification of incompletely visible vertebrae can be ambiguous in scans with low resolution or low-dose artifacts.

We used random elastic deformations, random Gaussian noise, random Gaussian

smoothing as well as random cropping along the z-axis to augment the training data. Rectified linear units were used as activation function in all layers except the output layers of the segmentation and the classification paths, in which sigmoid functions were used. The network was implemented using the PyTorch framework[1]. Training on an Nvidia Titan X GPU took about 4–5 days when training for 100 000 iterations.

4.4 Evaluation

4.4.1 Datasets

We trained and evaluated the method with five sets of CT and MR scans that visualize the spine. Reference segmentation masks for four of these datasets are publicly available, which allowed for a comparison with other publications that used the same data. Examples of images from the datasets are shown in Figure 4.3.

The *thoracolumbar spine CT* dataset consists of 15 dedicated spine CT scans that visualize all thoracic and lumbar vertebrae. It was originally used for the spine segmentation challenge held in conjunction with the Computational Spine Imaging (CSI) workshop at MICCAI 2014.[156] All subjects were young adults (20 to 34 years) without vertebral fractures who were scanned with IV-contrast administration. The scans were reconstructed to in-plane resolutions of 0.31 mm to 0.36 mm and slice thicknesses of 0.7 mm to 1.0 mm. Semi-automatically obtained reference segmentations were provided by the challenge organizers. To allow for a comparison with the challenge results, we used the same data split with 5 scans for evaluation and the remaining 10 scans for training and development.

The *xVertSeg.v1* dataset consists of 15 lumbar spine CT scans of subjects with compression fractures of various grades and types.[157] Manual reference segmentations are available for the lumbar vertebrae and were defined through a consensus reading of two observers. The scans were reconstructed to in-plane resolutions of 0.29 mm to 0.80 mm and slice thicknesses of 1.0 mm to 1.9 mm. There are currently two other publications that used the same dataset, but with different evaluation/training separation.[112,133] We therefore used the scans 1 to 5 for evaluation and the remaining 10 scans for training.

The *low-dose chest CT* dataset consists of 55 scans from the National Lung Screening Trial.[4] These scans were acquired for lung imaging and visualize in addition to the lungs a variable section of the thoracic and upper lumbar vertebrae. The scanned subjects were heavy smokers aged 50 to 74 years and therefore at increased risk for vertebral compression fractures due to their advanced age and smoking history. The scans were acquired with low radiation dose and reconstructed to in-plane resolutions of 0.54 mm to 0.82 mm and slice thicknesses of 1.0 mm to 2.5 mm. We created manual and semi-automatic reference segmentations for this dataset: 10 scans were used for evaluation and were therefore fully manually annotated by drawing along the contour

[1] pytorch.org

of each vertebra in sagittal slices using an interactive live wire tool.[158] The contours were converted into segmentation masks, in which inaccuracies and other mistakes were corrected voxel-by-voxel. An additional set of 5 scans was annotated in the same way and was used to train a preliminary version of the network. This network was used to predict rough segmentations in the remaining 40 scans. These rough segmentations were manually inspected and corrected voxel-by-voxel, and were used for training of the final network. This strategy enabled us to create a large training set with substantially less manual annotation effort compared to fully manual segmentation, which is not necessarily needed for training data. Additionally, a second observer fully manually annotated two scans from the evaluation set for an estimation of the interobserver agreement. All fully manual and semi-automatic segmentations were performed in sagittal views by observers who received detailed instructions beforehand. Additionally, all segmentations were validated by an experienced radiologist.

The *lumbar spine CT* dataset consists of 10 scans of healthy subjects and corresponding manual reference segmentations of the lumbar vertebrae.[114,118] The scans were reconstructed to in-plane resolutions of 0.28 mm to 0.79 mm and slice thicknesses of 0.7 mm to 1.5 mm. Because this dataset is the smallest of the datasets that we included, it was used for an external evaluation of our supervised approach. Scans from this dataset were therefore only used for evaluation and were not part of the training set.

The *lumbar spine MR* dataset consists of 23 T2-weighted turbo spine echo MR images acquired at 1.5T in sagittal orientation.[110] The scans have a resolution of 2 mm × 1.25 mm × 1.25 mm. Manual reference segmentations are available for 7 vertebrae (T11-L5) in all scans. These reference segmentations contain only the vertebral bodies, not the entire vertebrae.

Some scans in these datasets contained vertebrae or fragments of vertebrae that were not included in the reference segmentation, e.g., scans for which segmentations of only the lumbar vertebrae were available but in which also some thoracic vertebrae were visible. We manually cropped these images to restrict the field of view to the vertebrae that were included in the reference segmentation in order to avoid training the network with vertebra voxels incorrectly labeled as background voxels. In all scans, the segmented vertebrae were also anatomically labeled and marked as either completely or incompletely contained in the scan. None of the datasets contained multiple scans of the same subject.

4.4.2 Evaluation metrics

The segmentation performance was evaluated with the metrics most commonly reported in publications that used the same datasets. These metrics were the Dice coefficient to measure volume overlap and the average absolute symmetric surface distance (ASSD) to measure segmentation accuracy along the vertebral surface. Both metrics were calculated for individual vertebrae and then averaged across all scans.

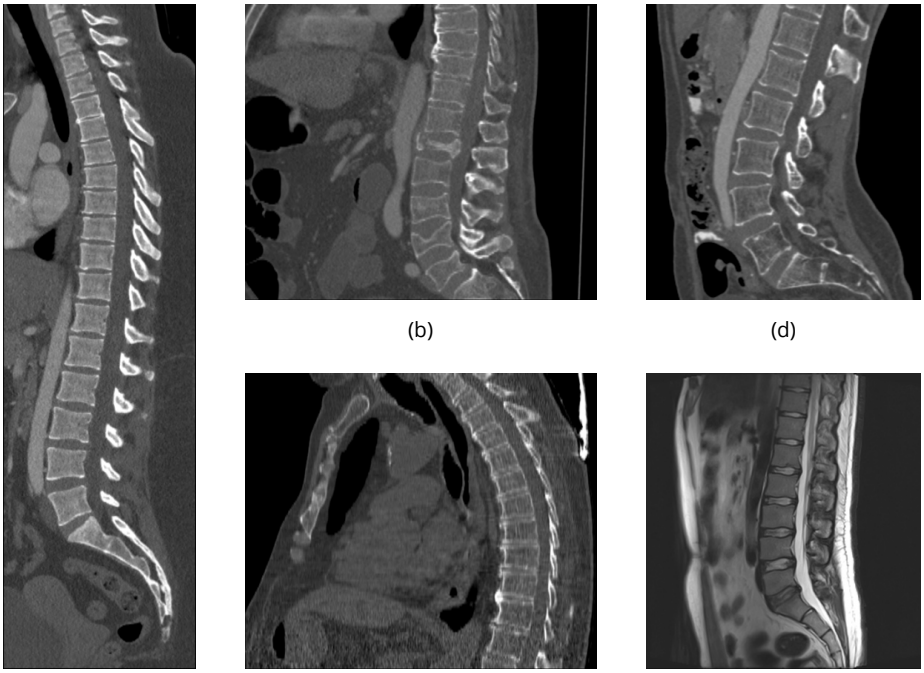


Figure 4.3: Examples of the various image types. Shown are examples of (a) thoracolumbar spine CT, (b) lumbar spine CT with multiple compression fractures, (c) low-dose chest CT, (d) lumbar spine CT, and (e) lumbar spine MR (T2-weighted).

Each vertebra in the reference segmentation was compared with the vertebra in the automatic segmentation mask with which it had the largest overlap.

The identification performance was evaluated using the identification accuracy, i.e., the percentage of vertebrae that were assigned the correct anatomical label, and the linearly weighted kappa coefficient.[159] Through the linear weighting, the kappa coefficient captures also the magnitude of mistakes.

The completeness classification performance was evaluated using the classification accuracy and the average number of false positives and false negatives per scan. False positives were in this case vertebrae that were incompletely visible, but were classified as completely visible.

4.5 Experiments and Results

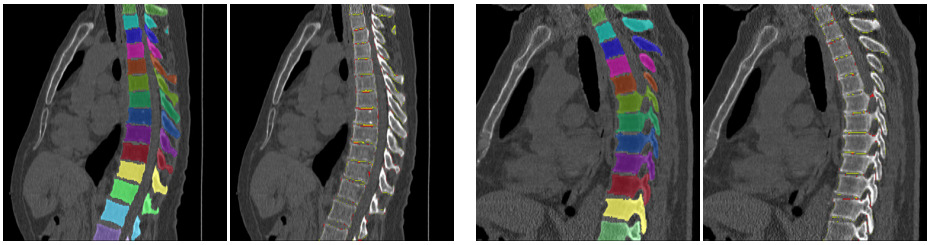
We trained modality-specific instances of the network adjusted to the different ground truths, i.e., to perform vertebra segmentation in CT and vertebral body segmentation in MR. The CT training set consisted of 60 scans, of which 10 were thoracolumbar

Table 4.1: Quantitative results of automatic segmentation, anatomical identification and completeness classification. Data are reported as mean ± standard deviation and were obtained with modality-specific networks, i.e., with a network trained with CT images and another network trained with MR images.

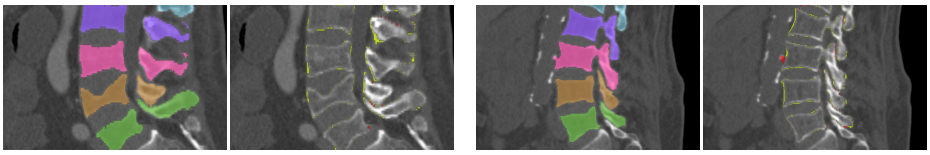
	Segmentation		*Identification*		*Completeness classification*		
Dataset	Dice score (%)	ASSD (mm)	Ac. (%)	κ	Ac. (%)	FP / scan	FN / scan
Thoracolumbar spine CT							
Proposed method	96.3 ± 1.3	0.1 ± 0.1	81	0.97	100	0	0
Lessmann et al.[134]	94.8 ± 1.6	0.3 ± 0.1	-	-	-	-	-
Korez et al.[118]	94.4 ± 2.1	0.3 ± 0.1	-	-	-	-	-
Lumbar spine CT [†]							
Proposed method	94.6 ± 2.2	0.3 ± 0.2	100	1.0	96.0	0	0.2
Janssens et al.[133]	95.7 ± 0.8	0.4 ± 0.1	-	-	-	-	-
Sekuboyina et al.[112]	94.3 ± 2.8	-	-	-	-	-	-
Low-dose chest CT							
Proposed method	93.1 ± 1.5	0.3 ± 0.1	100	1.0	95.7	0.5	0.1
Second observer [‡]	96.4 ± 0.6	0.2 ± 0.0	100	1.0	100	0.0	0.0
Lumbar spine CT							
Proposed method	96.5 ± 0.8	0.2 ± 0.0	100	1.0	100	0.0	0.0
Korez et al.[118]	95.3 ± 1.4	0.3 ± 0.1	-	-	-	-	-
Chu et al.[110]	91.0 ± 7.0	0.9 ± 0.3 [¶]	-	-	-	-	-
Ibragimov et al.[114]	93.6 ± 1.1	0.8 ± 0.1	-	-	-	-	-
Lumbar spine MR [§]							
Proposed method	94.4 ± 3.3	0.4 ± 0.3	100	1.0	100	0	0
Korez et al.[119]	93.4 ± 1.7	0.5 ± 0.1	-	-	-	-	-
Chu et al.[110]	88.7 ± 2.9	1.5 ± 0.2 [¶]	-	-	-	-	-

Ac. = Accuracy [†] xVertSeg.v1 dataset [‡] subset (2/10 scans) [§] only vertebral bodies [¶] ASD (non-symmetric)

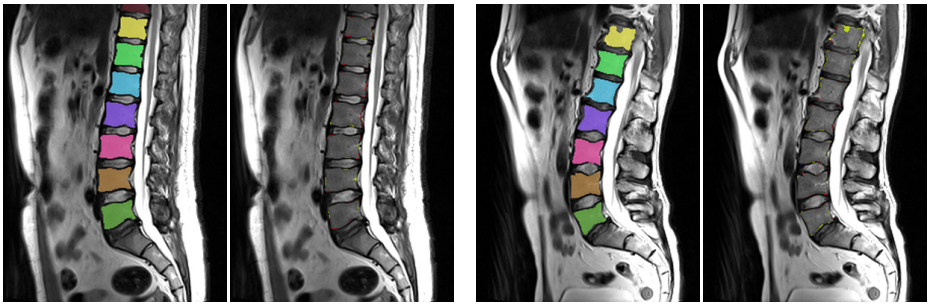
spine CT, 10 lumbar spine CT scans with compression fractures and 40 NLST scans. The CT evaluation set consisted of 30 scans, of which 5 were thoracolumbar spine CT scans, 5 lumbar spine CT with compression fractures, 10 NLST scans, and 10 normal lumbar spine CT scans, which were not represented in the training set. The dataset for vertebral body segmentation in MR consisted of only 23 scans in total and we therefore performed 3-fold cross-validation with evaluation sets of 8, 8 and 7 scans and the remaining scans used for training. The MR images were normalized by clipping off values below the 5^{th} and above the 95^{th} percentile and transforming the values into the range $[-1, 1]$. We compared the performance in vertebra segmentation, identification and completeness classification with that of other methods in the literature that were evaluated on the same datasets (Table 4.1). The networks were trained to traverse the spine upwards.



(a) Low-dose chest CT



(b) Lumbar spine CT (xVertSeg.v1 dataset)



(c) Lumbar spine MR

Figure 4.4: Segmentation results in different types of images. The segmentations are shown both as color overlay with different colors for different instances (left), and as difference maps with oversegmentation errors marked in red and undersegmentation errors in yellow (right). Some images have been cropped to better show the vertebral column.

4.5.1 Segmentation performance

Similar performance was achieved for vertebra segmentation in various CT datasets with an average Dice score of 94.9 ± 2.1 % and for vertebral body segmentation in an MR dataset with an average Dice score of 94.4 ± 3.3 %. Surface distances were lower on CT images compared to MR images (0.2 ± 0.1 mm vs. 0.4 ± 0.3 mm), however, there were also fewer training scans available in the MR dataset. Figure 4.4 illustrates the magnitude of differences of the automatic segmentations from the ground truth segmentations.

In the CT datasets, the segmentation was more accurate on high-resolution dedicated spine scans of healthy subjects compared with low-dose low-resolution chest CT

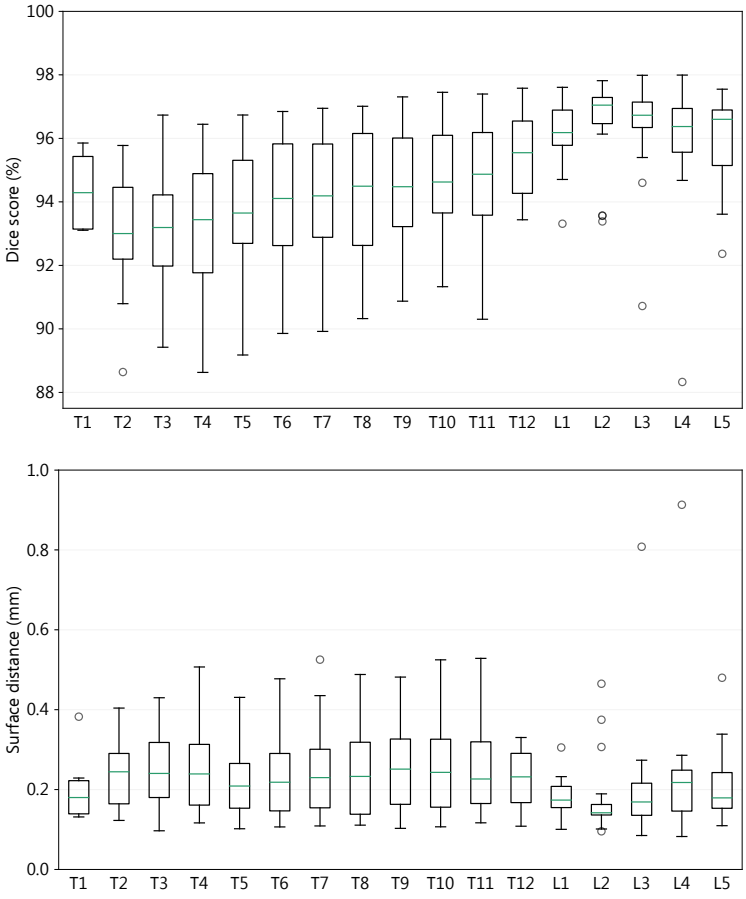


Figure 4.5: Box and whisker plots for per-vertebra Dice scores (top) and average absolute symmetric surface distances (bottom) in the CT evaluation set with 25 scans. Note that not all vertebrae were visible in every scan. Only completely visible vertebrae are included in the evaluation.

scans and scans of subjects with in some cases severe compression fractures. This is also visible in the segmentation performance stratified by vertebra (Figure 4.5). Segmentations were more accurate for the lumbar (L1-L5) than for the thoracic vertebrae (T1-T12), which are covered by the more challenging low-dose chest CT scans. Outliers among the lumbar vertebrae correspond to vertebrae from the xVertSeg.v1 dataset, which features a number of severely deformed lumbar vertebrae that are particularly challenging to segment.

In comparison with other vertebra segmentation methods, our iterative instance segmentation approach outperformed previous methods on the thoracolumbar spine CT dataset as well as on the lumbar spine CT dataset. In both cases, there was a substantial improvement in average Dice score and especially also in the surface

distance (Table 4.1). On the xVertSeg.v1 dataset with various fractured vertebrae, our method performed comparable to the method of Sekuboyina et al.[112] and not as well as the method of Janssens et al..[133] However, both of these publications used a different separation between training and evaluation data and the results are therefore not directly comparable. For vertebral body segmentation in MR, our approach achieved on average higher Dice scores and lower surface distances than previous methods, but with higher variance compared to Korez et al..[119] Although the automatic segmentation was overall accurate on low-dose chest CT, the performance was still slightly below the level of interobserver variation (average difference of 3.3 % in Dice score and 0.1 mm in surface distance).

The segmentation was optimized by minimizing a loss term based on false positive and false negative predictions. We additionally trained instances of the network using the categorical cross-entropy and the Dice coefficient as segmentation loss. However, the network optimized using the categorical cross-entropy converged in a state in which it predicted all voxels of any input patch as background. The network optimized using the Dice score performed substantially worse than the network trained with the proposed loss function, achieving for instance on low-dose chest CT scans an average Dice coefficient of 89.9 ± 6.8 % and an average surface distance of 0.5 ± 0.7 mm vs. an average Dice coefficient of 93.1 ± 1.5 % and an average surface distance of 0.3 ± 0.1 mm achieved by the proposed loss function.

Furthermore, the proposed network architecture predicts from an image patch not only a segmentation mask, but additionally also the anatomical label and a score for the completeness of the vertebra. To investigate whether embedding these additional tasks into the same network influences the segmentation performance, we trained an instance of the network with identical training data and settings, but without the labeling and completeness classification paths. While we observed minor differences, the segmentation performance of the networks with and without additional tasks was overall comparable (Table 4.2).

4.5.2 Identification performance

The anatomical identification was correct for 93 % ($^{206}/_{222}$, $\kappa = 0.99$) of the vertebrae in the CT datasets. The labeling of one thoracolumbar spine CT scan was offset by one vertebra, resulting in incorrect labeling of all 16 visible vertebrae. The range T2-L5 was predicted, but this patient had only four lumbar vertebrae and the correct range was therefore T1-L4. In the MR dataset, anatomical identification succeeded in all cases ($\kappa = 1.0$). Overall, the labeling was correct in 97 % of the scans.

Even though a variety of scans with different sections of the spine were included in our evaluation, most still visualized anatomical landmarks such as the sacrum that potentially simplified the anatomical identification. Even low-dose chest CT scans have a fairly standardized field of view defined by the location and size of the lungs,

Table 4.2: Comparison of the segmentation performance of a network consisting only of the segmentation path (S network) and a network consisting of the segmentation, the anatomical labeling and the completeness classification paths (S-L-C network), as proposed. Both networks traverse the spine upwards. Additionally, the segmentation performance is reported for the full network traversing downwards (S-L-C network ^{down}).

	Segmentation	
Dataset	Dice score (%)	ASSD (mm)
Thoracolumbar spine CT		
S network	96.2 ± 1.3	0.1 ± 0.1
S-L-C network	96.3 ± 1.3	0.1 ± 0.1
S-L-C network ^{down}	96.2 ± 1.4	0.2 ± 0.1
Lumbar spine CT[†]		
S network	95.3 ± 1.2	0.2 ± 0.1
S-L-C network	94.6 ± 2.2	0.3 ± 0.2
S-L-C network ^{down}	95.1 ± 1.1	0.2 ± 0.1
Low-dose chest CT		
S network	92.9 ± 1.7	0.3 ± 0.1
S-L-C network	93.1 ± 1.5	0.3 ± 0.1
S-L-C network ^{down}	91.0 ± 8.2	0.6 ± 1.1
Lumbar spine CT		
S network	96.4 ± 0.9	0.2 ± 0.1
S-L-C network	96.5 ± 0.8	0.2 ± 0.0
S-L-C network ^{down}	96.3 ± 0.9	0.2 ± 0.1

[†] xVertSeg.v1 dataset

which might have simplified vertebra identification. To evaluate the identification performance on arbitrary field of view images, we performed an experiment with randomly cropped images. For each of the 15 evaluation scans in the thoracolumbar spine CT and the chest CT dataset, we created two new images by randomly cropping the original image along the z-axis. These new images were minimally 80 % and 60 % smaller than the original image, but we ensured that they still contained multiple vertebrae by enforcing a minimum size of 150 mm along the z-axis. Of the vertebrae visible in these cropped images, 93 % were correctly identified ($^{174}/_{187}$; $\kappa = 0.98$). In the three images with mistakes, the labeling was offset by ±1, hence the high κ score.

4.5.3 Completeness classification performance

None of the scans in any of the datasets visualized the entire spine, all scans contained vertebrae that were only partially visible due to the limited field of view. In the CT datasets, 97 % of the vertebrae were correctly classified as completely or incompletely visible ($^{268}/_{275}$). In the MR dataset, all vertebrae were correctly classified. Most mistakes were made in low-dose chest CT scans, which was the type of scan with the least standardized field of view. Mistakes occurred only in the first or last visible vertebrae,

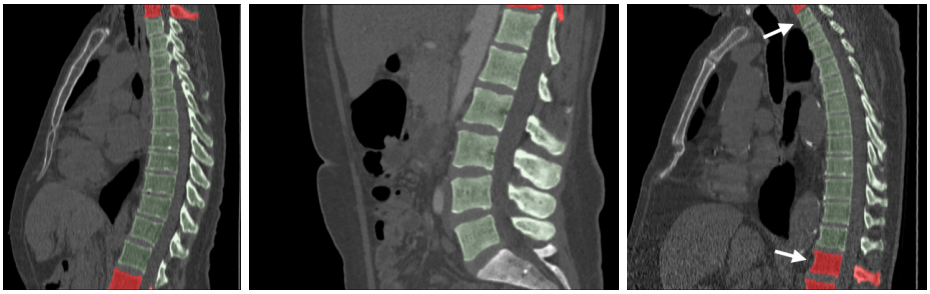


Figure 4.6: Examples of vertebra completeness classification results. Vertebrae classified as completely visible are marked in light green and vertebrae classified as incompletely visible are marked in red. Arrows indicate misclassified vertebrae. Shown are low-dose chest CT scans (left and right) and a lumbar spine CT scan (center).

near the boundary of the field of view, and often in vertebrae of which only a small part was missing from the scan (Figure 4.6). Notably, there were no mistakes where the network predicted an implausible sequence by classifying a vertebra between two completely visible vertebrae as incompletely visible, or vice versa.

4.5.4 Traversal direction

The networks in our experiments were trained to traverse upwards along the spine. To evaluate whether the direction of traversal influences the performance, we trained an additional instance of the network on the CT datasets to traverse downwards along the spine. We found no differences in anatomical identification and completeness classification of the detected vertebrae. However, while the segmentation performance was overall comparable with that of the same network traversing upwards, the segmentation performance deteriorated on the low-dose chest CT dataset (Table 4.2). In multiple low-dose chest CT scans, the top-most vertebrae were less well segmented (Figure 4.7). When traversing downwards, these are the first vertebrae to be segmented. Notably, the traversal process was still set off correctly in all cases.

In an additional experiment, we used two networks that were trained to traverse in opposite directions. The result of the upwards traversal was used to initialize the downwards traversal by starting the downwards traversal from the last vertebra detected during upwards traversal. We hoped that this would result in improved segmentation performance because it relieves the downwards network from detecting the first vertebra without any context information provided by the instance memory. However, this showed not to be beneficial and resulted in virtually identical segmentation performance compared to only upwards traversal.

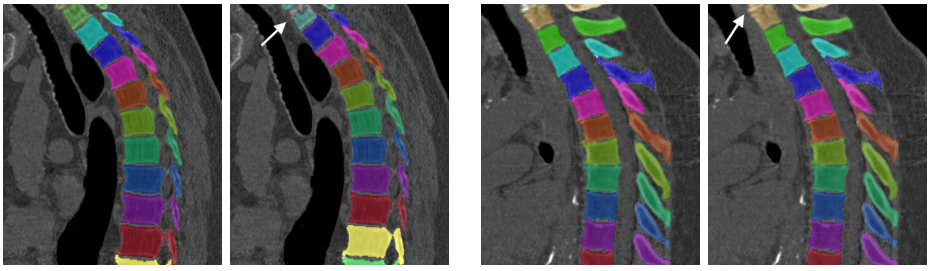


Figure 4.7: Segmentations obtained with the proposed iterative segmentation approach, comparing models trained for either upwards or downwards traversal. Both examples are low-dose chest CT scans (cropped). White arrows indicate segmentation errors in the top-most visible vertebrae, which occur more often when traversing downwards.

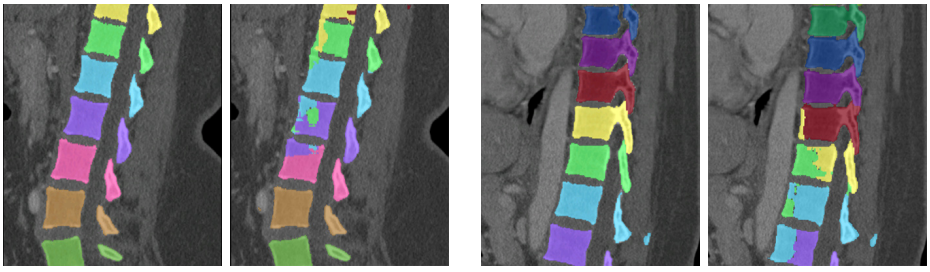


Figure 4.8: Segmentations obtained with a multiclass FCN, thus without using the iterative segmentation strategy, compared with segmentations obtained with the proposed iterative approach. While the segmentations are overall fairly accurate, the individual vertebrae are not well separated.

4.5.5 Comparison with multiclass FCN

To verify that an instance segmentation approach is beneficial, we also trained the segmentation component of our network, which is a U-net like 3D FCN with skip connections (Figure 4.1), to segment and identify the vertebrae using multiclass voxel classification instead of the proposed iterative binary segmentation. This network received an image patch of 128 × 128 × 128 voxels as input, but unlike in our iterative approach not a corresponding patch from the instance memory. The network had 25 output classes, corresponding to the 24 different vertebrae and a background class. At inference time, the patch was moved over the entire image in a sliding window fashion with overlapping windows so that multiple predictions were obtained for each voxel. Each voxel was eventually labeled with the class label that had the highest average probability. Using non-overlapping windows resulted in substantially worse performance. Overall, the multiclass FCN achieved an average Dice score of 78.7 ± 13.5 % and an average ASSD of 8.2 ± 9.4 mm. The anatomical identification was successful in 89 %

($\kappa = 0.98$) of the vertebrae. While the segmentations were overall reasonably accurate with a Dice score of 84.6 ± 6.9 % when the class labels were disregarded, the individual instances were often not well separated from each other (Figure 4.8).

4.5.6 Runtime

The runtime of a single iteration step was about 1 s on standard hardware. The number of required iteration steps and thus the overall runtime per image depends on the size of the image, the number of visible vertebrae and their location within the image, which influences how many steps are initially needed to find the first vertebra. For instance, the low-dose chest CT scans covered on average 13.7 vertebrae and required 55 iteration steps. The thoracolumbar spine CT scans covered on average 17.8 vertebrae, but with a narrower field of view focused on the spine and required a comparable number of iteration steps (56 on average). The average runtime per scan was about one minute in all datasets, excluding time required for loading the image and storing the results.

4.6 Discussion

This chapter demonstrates that fully convolutional neural networks, which have been widely used for semantic segmentation,[160] are also capable of learning a complex instance segmentation task. Vertebra segmentation performed instance-by-instance required the network to learn to infer from an additional memory input which vertebra to segment and to ignore other vertebrae. Additionally, the same network was able to perform multiple tasks concurrently, namely vertebra segmentation, identification and classification to determine whether the vertebra was completely contained in the scan. This approach outperformed all methods that participated in the CSI 2014 spine segmentation challenge[156] and performed better or comparable to state-of-the-art methods on other datasets. In a particularly challenging set of low-dose chest CT scans, the performance was close to the interobserver variability.

The diverse selection of datasets that we used to evaluate the iterative segmentation approach demonstrates that this approach can cope with arbitrary fields of view, with low-resolution and low-dose scans, and with scans with normal as well as severely deformed vertebrae. The approach is entirely supervised, which enables transferring the same approach to other modalities and other segmentation tasks, which we demonstrated by applying the same approach without any modifications to vertebral body segmentation in T2-weighted MR images. Moreover, we demonstrated that the networks did not overfit to the datasets that were represented in the training data, but instead outperformed state-of-the-art methods on an entirely unseen datasets of lumbar spine CT scans. Previous methods for vertebra segmentation were often tailored to specific image types and evaluated on homogeneous datasets. Although our approach performed slightly worse than some of these methods on some of the datasets included

in our evaluation, we demonstrated consistently high performance across multiple datasets.

The proposed iterative vertebra-by-vertebra segmentation performed substantially better than a regular multiclass FCN similar to a 3D U-net.[154] However, the performance of the multiclass FCN exceeded our expectations and might further improve with more training data and with hardware that enables training of larger networks, or by including more context information via, e.g., a multi-scale approach.[161,162] Even though the iterative approach requires the network to combine two inputs to identify a specific vertebra and ignore others, it also simplifies the segmentation problem from a multiclass into a binary voxel labeling task. The direct comparison of these two approaches indicates that individual instances of a target class of objects are better separated by a FCN if the network is trained to focus on individual instances. This strategy could also lead to improvements in other instance segmentation tasks, for instance, in histopathological image analysis.

We combined multiple tasks into a single network, which helps to simplify both training and inference: only a single network needs to be trained, and at inference time, each patch needs to be passed only through one network to obtain multiple predictions, one for each distinct task. Even though the segmentation path and the identification and completeness classification paths shared part of the network, we did not find that this improved the segmentation performance. However, the segmentation performance did also not deteriorate when these additional tasks were added, which indicates that the proposed combination into a single network is useful. The proposed network architecture with the additional output paths uses more GPU memory than task specific networks would and therefore more strongly limits the maximum number of filter per layer and the depth of the network. Future hardware generations with larger memory will enable training of larger networks, which might lead to further performance improvements.

The proposed spine traversal strategy can be applied to segment the vertebrae from top to bottom or vice versa. However, we observed better performance for upwards compared to downwards traversal on low-dose chest CT scans with differences mostly in the upper vertebrae. The size of the vertebrae increases from cervical to lumbar vertebrae, i.e., from top to bottom. Hence, when traversing downwards, the first vertebra that needs to be found, without additional information that could be derived from the memory input, is the smallest vertebra that is visible in the scan. Additionally, the region around the uppermost visible vertebrae is often affected by low-dose artifacts in low-dose chest CT scans, which makes starting the traversal from these vertebrae especially challenging. Because there were no substantial differences between downwards and upwards traversal in the other CT datasets, upwards traversal presents overall the more robust strategy.

Even though we used a variety of datasets in our experiments, some important types of data were not present. These include scans covering the cervical vertebrae

and scans of patients with implants near the spine, such as pedicle screws. However, since our approach is entirely supervised and trained end-to-end, it will likely be able to handle these cases if sufficient training examples are available. Furthermore, there were no scans of patients with irregular numbers of vertebrae present in the training set, which caused mislabeling of the vertebrae in one patient with irregular number of vertebrae in the evaluation set. Ensuring accurate anatomical labeling for such cases could be an interesting direction for further research. However, the high kappa scores indicate that the labeling would be minimally offset in such cases. Depending on the exact clinical application, the impact of such labeling mistakes would therefore be limited.

Segmentation of the vertebrae one after the other, using information about the already segmented vertebrae as a prior, is inherently susceptible to cascading failure. Failure to find or correctly segment a single vertebra may cause failure to find or correctly segment all subsequent vertebrae. There is additionally no element that explicitly ensures that the predicted segmentation mask covers only a single vertebra if multiple are visible. While we did not observe these kind of failures in our evaluation, they are likely to occur in images with extreme anatomical abnormalities or severe imaging artifacts. Refinement of the labeling through a maximum likelihood approach suffers from a similar weakness: Except for cases with irregular number of vertebrae, the labeling can only be entirely correct or entirely offset, even if the correct labels were predicted for some of the vertebrae. This limitation could potentially be addressed in the future by employing a more sophisticated global labeling model, e.g., based on Markov models.

Manual vertebra segmentation is a time-consuming and tedious task, requiring annotation times of about 40 to 60 hours per scan in low-dose chest CT and even longer in scans with higher resolution and coverage of more vertebrae. Semi-automatic segmentation proved to be an effective strategy for generating reference segmentations for network training from only few manual reference segmentations. Especially if little training data is available, additional priors or model fitting steps could help stabilize the performance. These could be statistical knowledge about typical sizes or shapes of vertebra, or additional fitting of a deformable surface mesh model to the segmentation results.[119]

Precise segmentation and identification of the vertebrae from CT and MR scans enables automatic spine analysis, notably also in images that were originally not intended for spine imaging. For instance, our iterative approach achieved a segmentation and identification performance on low-dose chest CT scans that is likely sufficient to analyze the shape of the vertebral bodies for detection of compression fractures. This could enable opportunistic screening for early signs of osteoporosis in lung cancer screening programs, in addition to screening for pulmonary abnormalities.

In conclusion, this chapter presents an iterative instance-by-instance approach to vertebra segmentation and anatomical identification. This approach is fast, flexible and

accurate across a large variety of both dedicated as well as non-dedicated spine scans.

Acknowledgments

We would like to thank the organizers of the CSI 2014 spine segmentation challenge, the Laboratory of Imaging Technologies at the University of Ljubljana and the authors of the MR dataset for making scans and reference segmentations publicly available. Furthermore, we are grateful to the United States National Cancer Institute (NCI) for providing access to NCI's data collected by the National Lung Screening Trial. The statements contained herein are solely ours and do not represent or imply concurrence or endorsement by NCI.

5

CHAPTER 5

Vertebra partitioning with thin-plate splines steered by neural networks

BASED ON: N. Lessmann, J. M. Wolterink, M. Zreik, P. A. de Jong, W. Foppen, S. C. E. Diepstraten, M. A. Viergever, B. van Ginneken, and I. Išgum. "Thin-plate spline surfaces steered by a convolutional neural network: application to vertebra partitioning based on shape information". Submitted.

Abstract

Thin-plate splines are commonly used for interpolation, but they can also be used to represent the boundary that separates two structures as a smooth surface. This chapter presents a method for partitioning vertebra segmentation masks into two substructures, the vertebral body and the posterior elements, using a convolutional neural network that predicts the boundary between the two structures. This boundary is modeled as a thin-plate spline surface defined by a set of control points predicted by the network. The neural network is trained using the reconstruction error of a convolutional autoencoder. This autoencoder is trained with segmentation masks of one of the substructures and is intended to capture a range of plausible shapes of only this substructure. Both networks are trained with binary segmentation masks, containing no image intensities, which enables the combination of datasets from different modalities. The method was trained for vertebra partitioning with a large set of 6247 vertebra segmentation masks from 580 chest CT scans and 1193 vertebral body segmentation masks from 314 lumbar spine MR scans. We evaluated the method with manual reference segmentations of 50 vertebral bodies in chest CT, and the method achieved an average Dice score of 99.2 ± 0.5 % and an average Hausdorff distance of 3.7 ± 1.5 mm. Partitioning the segmentation masks of all visible vertebrae required less than one second per scan. This enables extraction of disease markers from the vertebral body only, such as detection of osteoporotic compression fractures, or assessment of an involvement of the posterior elements in traumatic fractures.

5.1 Introduction

Segmentation is one of the most common tasks in medical image analysis but has by itself no immediate purpose. Segmentations are often an essential component of analysis pipelines, such as dose planning for radiotherapy based on segmentations of the tumor and sensitive neighboring organs.[163,164] Most commonly, however, segmentations are used to extract clinical parameters that further characterize the segmented structures. For instance, parameters such as size or shape of pulmonary nodules to classify them as benign or malignant,[165] parameters such as surface area and thickness of the cortical gray matter to monitor the development of preterm infants,[166] or parameters such as calcium scores[79] or ventricular ejection fractions[167] to predict and diagnose cardiovascular events. Many of these clinical parameters are related to only part of a larger structure, such as the left ventricle that is part of the heart or the cortical gray matter that is part of the brain. Segmentation methods therefore often either segment only the relevant substructure[79,168] or label different substructures during the segmentation. Examples of the latter are approaches based on deformable shape models where the reference model is labeled into substructures,[169,170] atlas-based segmentation approaches where the atlases are labeled into substructures,[171,172] and approaches based on multi-label convolutional neural networks (CNNs) that are trained to directly label substructures.[160,173,174]

Segmentation with supervised CNNs presents an especially attractive approach due to their ability to learn to segment a target structure directly from images and corresponding reference segmentations. The availability of a sufficiently large number of reference segmentations is therefore usually the most limiting factor. Particularly, reference segmentations of anatomical substructures or multi-label reference segmentations may not be available and may be cumbersome and costly to obtain. If only segmentations of the unpartitioned structure are available, post-processing of the segmentation masks to partition them into substructures can be an alternative to segmenting or labeling these substructures directly, especially if direct segmentation or labeling would require creating a new set of reference segmentations.

Partitioning a segmentation mask into smaller substructures requires finding the boundaries between these substructures. The overall shape of these boundaries is highly task-specific since different anatomical structures often have very different shapes. If the boundary between two substructures is sufficiently smooth, it could be represented by a thin-plate spline (TPS). The position, orientation and shape of a TPS surface is defined by an arbitrary number of control points.[175] Finding appropriate control points of a TPS surface is therefore sufficient to partition a segmentation mask into two substructures.

In this paper, we present a partitioning approach based on a CNN that steers a TPS surface (TPS-CNN) to partition vertebra segmentation masks into the vertebral body and the posterior elements (Figure 5.1). The TPS-CNN receives the unpartitioned vertebra

segmentation mask as input. It outputs the coordinates of a fixed number of control points such that the corresponding TPS surface partitions the segmentation mask into the two substructures, namely vertebral body and posterior elements. This approach is intended as a post-processing step after segmentation of the vertebra. We additionally assume that corresponding finer or partitioned segmentations of the substructures are not available for supervised training. Would these be available, it would be trivial to train a network that performs voxel classification to label the substructures, such as a multi-class U-net.[68] We therefore assume that only unpaired data from different subjects is available. While for one set of subjects vertebra masks are available but no segmentations of substructures, in another set of subjects segmentations of one substructure are available, namely the vertebral body. To demonstrate that even data from different modalities can be combined, the TPS-CNN is trained with vertebra segmentation masks from CT scans and vertebral body segmentation masks from MR scans.

It may seem like a far-fetched scenario that segmentations of an anatomical structure and a substructure are available only in images from different modalities and different subjects. However, an approach that combines such data allows for combination of publicly available datasets—for instance from segmentation challenges—which could otherwise not be used together. We facilitate the combination of data from different modalities by limiting analyses to the segmentation masks, which represent the anatomical structures in the shape domain, and by disregarding intermodality intensity differences in the image domain. We train a convolutional autoencoder neural network (CAE) to capture the shape of vertebral bodies (Figure 5.1).[176] This autoencoder is subsequently used to estimate the out-of-distribution likelihood of a sample, i.e., whether a segmentation mask is similar to segmentations of vertebral bodies or not. This likelihood estimate is used to formulate a differentiable loss function for the TPS-CNN that partitions a vertebra segmentation mask by positioning a TPS surface. The presented partitioning method is evaluated using manual reference segmentations of the vertebral bodies in CT scans. Additionally, we demonstrate on a large set of CT scans that the volume of the extracted vertebral bodies correlates well with surrogate height measurements of the vertebral bodies.

5.2 Datasets

We used our previously developed automatic iterative segmentation method[177] to obtain a large set of vertebra and vertebral body segmentation masks to train the proposed partitioning approach. Segmentations of the full vertebrae were obtained from a set of low-dose chest CT scans and segmentations of the vertebral bodies were obtained from a set of lumbar spine MR scans (Figure 5.2). The automatically obtained segmentation masks were manually inspected and vertebrae with segmentation errors were removed from the training set. For evaluation of the proposed partitioning method, the vertebral

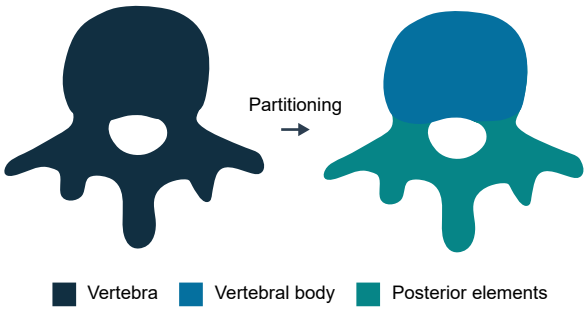


Figure 5.1: Illustration of the vertebra partitioning task. The vertebra (left) is partitioned into the vertebral body and the posterior elements (right). Note that the boundary between body and posterior elements is not a plane. The vertebra is shown from above, corresponding to the axial plane in typical spine CT/MR scans. See also Figure 5.6 for 3D renderings.

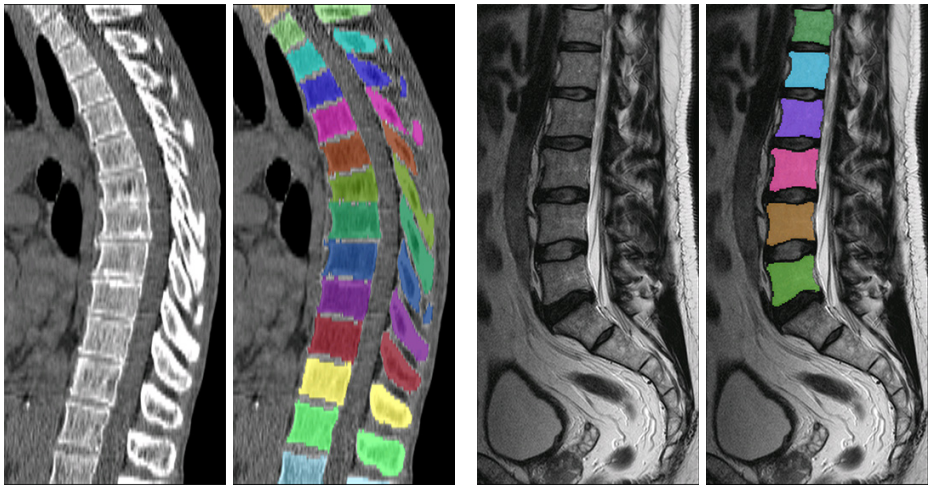
bodies were additionally manually annotated in a set of low-dose chest CT scans.

5.2.1 Chest CT

Segmentation masks of entire vertebrae were obtained from low-dose chest CT scans acquired as part of a large multi-center lung cancer screening trial in the United States, the National Lung Screening Trial (NLST).[23] These chest CT scans visualize a variable section of the spine, typically including all thoracic vertebrae and the upper lumbar vertebrae. We used a step-by-step approach, from manual to semi-automatic to fully automatic segmentation, to obtain a large set of vertebra segmentation masks in these scans.

In 15 scans, the vertebrae were manually segmented by drawing along the contour of the vertebrae in sagittal slices using a live wire tool, followed by pixel-by-pixel correction of inaccuracies. Since this process is very time-consuming (around 40 h per scan), the 15 manual segmentations were used to create additional semi-automatic segmentations in 40 scans by training the automatic segmentation method[177] with manual segmentations and manually correcting the resulting segmentation masks. Using both the manual and the semi-automatic segmentations, the automatic segmentation method was retrained and applied to a set of 600 NLST scans. After visual inspection and removal of vertebra segmentations with segmentation errors, the automatic segmentation results consisted of in total 6247 individual vertebrae from 580 scans. These automatic segmentation results were used to train the partitioning method.

To establish a ground truth for evaluation of the proposed automatic vertebra partitioning approach, the posterior elements were manually removed from the manual vertebra segmentation masks in six scans with 60 vertebrae, leaving only the vertebral bodies. These annotations were only used for evaluation and not for training of the



(a) Vertebrae in chest CT

(b) Vertebral bodies in lumbar spine MR

Figure 5.2: Examples of the scans from which segmentation masks of (a) the full vertebrae and (b) only the vertebral bodies were derived. In the images overlayed with the segmentation masks, colors indicate different vertebrae. Images show the sagittal plane, roughly through the middle of the vertebrae.

partitioning approach. Because this manual annotation process is very time-consuming, making the creation of a large set of reference segmentations hardly feasible, we additionally measured the anterior, middle and posterior heights of the vertebral bodies as a surrogate marker of vertebral body size (Figure 5.3) in 389 chest CT scans. Heights were measured using digital calipers in mid-sagittal slices, taking only a few minutes per scan. The four observers who measured the vertebral body heights were a radiologist, two radiology residents and a specifically trained analyst, and were blinded to the segmentation and partitioning results. In total, heights of 4391 vertebral bodies were measured.

5.2.2 Lumbar spine MR

Segmentation masks of vertebral bodies were obtained from T2-weighted turbo spin echo lumbar spine MR images acquired in sagittal orientation. The automatic segmentation method was trained with 23 publicly available MR images from multiple hospitals and corresponding manual segmentations of the vertebral bodies.[110] The same automatic segmentation method,[177] used before for vertebra segmentation in chest CT, was used for vertebral body segmentation in MR scans. The retrained automatic segmentation method was applied to 497 lumbar spine MR scans with similar acquisition parameters (T2-weighted, turbo spin echo, sagittal) that were acquired as part of routine clinical care at the University Medical Center Utrecht. After visual inspection

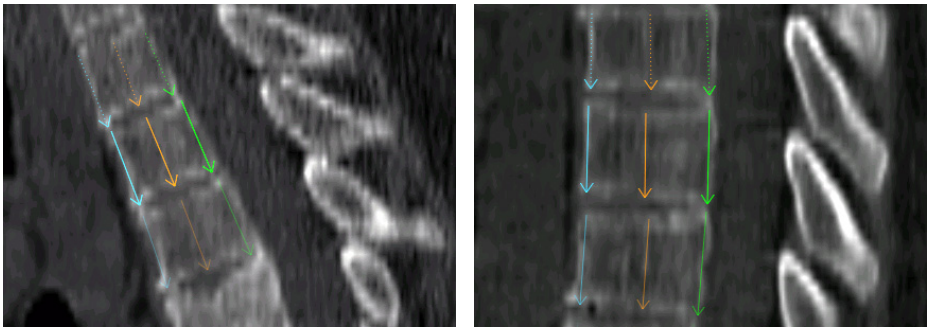


Figure 5.3: Examples of vertebral body height measurements. The observers identified the mid-sagittal slice of each vertebral body and manually drew three 2D vectors, measuring the anterior (blue), middle (orange) and posterior (green) heights. These measurements were averaged and used as estimate of vertebral body size.

and removal of vertebral body segmentations with segmentation errors, the dataset consisted of in total 1193 individual vertebral bodies from 314 scans. These automatic segmentation results were used to train the partitioning method.

5.3 Methods

We propose to partition segmentation masks of vertebrae into vertebral body and posterior elements using a CNN that steers a TPS surface, which represents the boundary between the partitions. This TPS-CNN predicts the locations of a set of control points that define the TPS surface. This surface is used to construct a probabilistic mask to mask out one of the partitions. Because we assume that there are no matching reference segmentations for the remaining structure, we cannot calculate a direct loss function. Instead, we first train a convolutional autoencoder with segmentations of this substructure from different subjects, and in this chapter even from MR instead of CT scans. This autoencoder is used to compress and afterwards decompress the substructure that remains after applying the TPS-mask, assuming that it has learned to reconstruct this particular structure and will fail to reconstruct other structures. The reconstruction error of the autoencoder is therefore used to update the TPS-CNN through backpropagation (Figure 5.4). The autoencoder is not updated anymore during training of the TPS-CNN. At inference time, the autoencoder is not needed and the partitioning TPS surface is directly predicted from the segmentation mask.

5.3.1 Thin-plate spline surfaces

Before discussing how to steer TPS surfaces with CNNs, we first briefly revisit their definition and properties, following the notation introduced by Bookstein.[178] The thin-plate spline is a radial basis function of the form

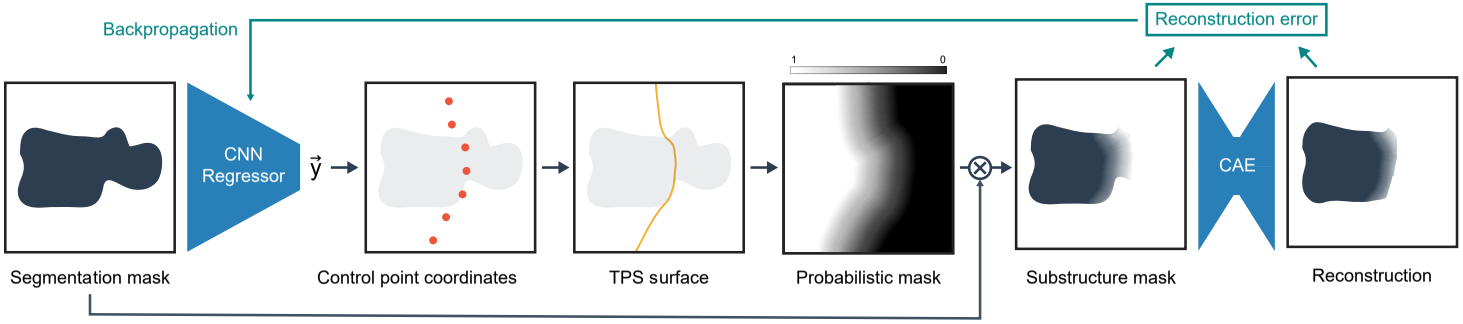


Figure 5.4: Flowchart illustrating the training process. The segmentation mask is fed into a convolutional neural network (CNN Regressor, in this chapter referred to as TPS-CNN) that regresses the y-coordinates of a number of control points. These define a thin-plate spline surface, from which a probabilistic mask is derived which is applied to the input segmentation mask. The resulting substructure mask is encoded and then decoded by a convolutional autoencoder (CAE), which was previously trained with segmentation of the substructure from a different modality. The reconstruction error is used to update the CNN. At test time, the CAE is not used and the probabilistic map is thresholded to obtain a segmentation of the substructure.

$$U(r) = r^2 \log r. \tag{5.1}$$

A weighted combination of thin-plate splines centered at N control points $\mathbf{p}_i = (x_i, y_i, z_i)$ describes the height z of a three-dimensional surface at point (x, y):

$$z(x, y) = a + bx + cy + \sum_{i=1}^{N} w_i \cdot U\left(r_i(x, y)\right) \tag{5.2}$$

where

$$r_i(x, y) = \sqrt{(x - x_i)^2 + (y - y_i)^2}. \tag{5.3}$$

The coefficients a, b, c and $w_1, \ldots, w_N$ can be chosen such that the surface passes exactly through all control points $\mathbf{p}_i$ by forcing $z(x_i, y_i) = z_i$. The resulting surface is minimally bent and smooth between control points. Physically, a TPS surface corresponds to a thin metal sheet that is fixated at the control points and that is minimally bent because it resists bending due to the natural rigidity of the material.

The constraint $z(x_i, y_i) = z_i$ corresponds to a system of linear equations that can be written in matrix form using the matrices

$$\begin{aligned} \mathbf{K} &= \begin{pmatrix} 0 & U(p_1, p_2) & \cdots & U(p_1, p_N) \\ U(p_2, p_1) & 0 & \cdots & U(p_2, p_N) \\ \vdots & \vdots & \ddots & \vdots \\ U(p_N, p_1) & U(p_N, p_2) & \cdots & 0 \end{pmatrix}, \\ \mathbf{P} &= \begin{pmatrix} 1 & x_1 & y_1 \\ 1 & x_2 & y_2 \\ \vdots & \vdots & \vdots \\ 1 & x_N & y_N \end{pmatrix}, \quad \text{and} \quad \mathbf{L} = \begin{pmatrix} \mathbf{K} & \mathbf{P} \\ \mathbf{P}^\mathsf{T} & 0 \end{pmatrix}, \end{aligned} \tag{5.4}$$

and coefficients vector $\mathbf{C} = (w_1, \ldots, w_N, a, b, c)^\mathsf{T}$ and control point heights vector $\mathbf{Z}_p = (z_1, \ldots, z_N, 0, 0, 0)^\mathsf{T}$, resulting in the equation $\mathbf{L} \cdot \mathbf{C} = \mathbf{Z}$. The last three rows of $\mathbf{L}$ complete the system of linear equations by imposing the three additional constraints $\sum w_i = 0$, $\sum x_i w_i = 0$ and $\sum y_i w_i = 0$ so that there are N + 3 equations and N + 3 coefficients. These additional constraints ensure that the influence of the control points is zero at infinity, i.e., the surface is flat at large distance from the control points.

The coefficients $\mathbf{C}$ of the TPS surface can be determined by multiplication with the inverse or pseudo inverse of $\mathbf{L}$:

$$\mathbf{C} = \mathbf{L}^{-1} \cdot \mathbf{Z}_p. \tag{5.5}$$

5.3.2 Steering TPS surfaces with CNNs

For the intended application of segmentation mask partitioning, the TPS surface acts as barrier between the two parts. In the specific application of vertebra partitioning, this corresponds to a barrier between the vertebral body and the posterior elements (Figure 5.1). A CNN that receives the segmentation mask as input could position a TPS surface by predicting the location of suitable control points. However, using this TPS surface to partition the segmentation mask requires calculating the height of the predicted TPS surface at various points across the segmentation mask grid, which requires calculating U(r) for all pairs of mask grid points and control points. Furthermore, to calculate the coefficients $\mathbf{C}$ from a set of predicted control point locations, U(r) also have to be calculated pairwise for all control points. Furthermore, the inverse of the matrix $\mathbf{L}$, which grows quadratically with the number of control points, has to be computed, which will usually involve matrix decomposition to calculate a pseudo inverse. This latter step will in many cases not be differentiable and training the CNN with backpropagation would therefore not be possible.

However, we can simplify this process by using a fixed grid of control points, i.e., by keeping the x_i and y_i coordinates of the control points constant and having the CNN predict only the corresponding heights z_i. If x_i and y_i are constant, the matrices $\mathbf{K}$ and $\mathbf{P}$, and consequently $\mathbf{L}$, become constant as well (eq. 5.4). If $\mathbf{L}$ is constant, also its inverse $\mathbf{L}^{-1}$ is constant and therefore the coefficients of the TPS surface can be calculated with a single matrix-vector multiplication (eq. 5.5), which is differentiable.

If we additionally define a fixed mask grid $q_j = (x_j, y_j)$ of M points at which we probe the height of the TPS surface, e.g., at each discrete x and y of the segmentation mask, the corresponding calculations of $z(x_j, y_j)$, based on eq. 5.2, can be written using the matrix

$$\mathbf{G} = \begin{pmatrix} U(q_1, p_1) & \cdots & U(q_1, p_N) & 1 & x_1 & y_1 \\ \vdots & & \vdots & \vdots & \vdots & \vdots \\ U(q_M, p_1) & \cdots & U(q_M, p_N) & 1 & x_M & y_M \end{pmatrix}, \tag{5.6}$$

which enables calculating the heights $\mathbf{Z}_q$ of the TPS surface across the mask grid as

$$\mathbf{Z}_q = \mathbf{G} \cdot \mathbf{C} \overset{(5.5)}{=} \mathbf{G} \cdot \left(\mathbf{L}^{-1} \cdot \mathbf{Z}_p\right). \tag{5.7}$$

With a fixed control point grid and a fixed mask point grid, the matrices G and $\mathbf{L}^{-1}$ become constants that can be precomputed before training the CNN. With the control point heights $\mathbf{Z}_p$ predicted by the CNN, the mask grid heights $\mathbf{Z}_q$ can be calculated with simple matrix-vector multiplications. These mask grid heights are subsequently transformed into a probabilistic mask by calculating for each voxel in the segmentation mask the signed distance to the TPS surface and applying the sigmoid function so that the resulting mask has values in the range $[0, 1]$. We approximate the distance of voxel

(x_k, y_k, z_k) to the TPS surface simply by the difference with the predicted height of the surface at (x_k, y_k):

$$d(x_k, y_k, z_k) = z(x_k, y_k) - z_k. \tag{5.8}$$

The resulting probabilistic mask is multiplied voxel-by-voxel with the input segmentation mask to mask out one of the substructures. In our application of vertebra partitioning into vertebral body and posterior elements, the posterior elements are masked out and the vertebral body remains. Note that for this application, the TPS surface is positioned in the coronal plane, i.e., the height of the surface corresponds to the y coordinate of the segmentation mask, and the x and z coordinates are fixed.

Using the resulting probabilistic segmentation mask of the remaining substructure (the vertebral body), the CNN could be trained in a conventional supervised fashion if corresponding reference segmentations would be available. These would enable training the CNN with loss functions commonly used for segmentation tasks, such as the voxelwise cross entropy or the Dice score. However, we assume here that reference segmentations of the substructure are not available for the same scans and therefore train the CNN using an alternative loss function in the shape domain.

5.3.3 Shape-encoding loss function

To enable the use of reference segmentations of the substructure stemming from different scans, different modalities and even different subjects than those of the superstructure, we propose to use a loss function based on shape encodings. Image information is not used, assuming that the shape of an anatomical structure is roughly the same across modalities while the appearance in the image is usually not. A convolutional autoencoder is trained with segmentation masks of the substructure, i.e., vertebral bodies. This autoencoder learns to represent segmentation masks of vertebral bodies in a feature space, encoding the inter- and intra-patient variations present in the training data. The loss function for training the TPS-CNN is based on the assumption that the trained autoencoder is good at encoding and reconstructing segmentation masks of the substructure but will ideally produce worse reconstructions of other objects, such as the unpartitioned structure or the result of inaccurate partitioning attempts. We therefore use the reconstruction error of the autoencoder to train the CNN. The parameters of the autoencoder itself are not changed anymore when training the CNN, unlike in related approaches for energy-based adversarial training with autoencoders.[179] In our experiments, the reconstruction error was defined as the squared error between input and output of the autoencoder.

In addition to the reconstruction loss, two regularizing terms are added to the loss function which are intended to stabilize the learning behavior. With the first term, we penalize high curvature of the TPS surface by minimizing the absolute value of the Hessian determinant. The Hessian matrix is approximated with finite differences of

neighboring control point heights. With the second term, we penalize partitioning solutions in which the remaining substructure is much smaller than the unpartitioned structure. We observed that an empty mask is usually perfectly "reconstructed" by the autoencoder, even though it was never trained with empty input masks, and the CNN would therefore often converge to placing the partitioning TPS surface such that the entire structure is masked out. The volume penalty therefore encourages the network to make the substructure for which the reconstruction loss is computed as large as possible. We defined this loss term as the difference of the number of voxels in the binary segmentation mask of the unpartitioned structure M_{super} and in the probabilistic mask of the substructure M_{sub}, where the sum of all voxel values corresponds to a probabilistic volume score. The volume penalty can then be written as:

$$L_v = \sum M_{super} - \sum M_{sub}. \tag{5.9}$$

The loss function L used to optimize the TPS-CNN is a linear combination of the reconstruction loss L_r, the curvature penalty L_c and the volume penalty L_v:

$$L = \alpha_r L_r + \alpha_c L_c + \alpha_v L_v. \tag{5.10}$$

We used $\alpha_r = 1$, $\alpha_c = 1000$ and $\alpha_v = 0.5$ in our experiments based on preliminary results on the validation set. Note that the reconstruction loss is the sum of the squared errors, not the mean, and is therefore a much larger value than the other two terms. When using the mean squared error as reconstruction loss, the weights for the curvature and volume penalties might need to be reduced.

5.3.4 Network architectures

For both the CAE and the TPS-CNN, we used a 3D architecture based on repeated blocks consisting of two 3 × 3 × 3 convolutions with padding, batch normalization and the rectifier activation function, followed by 2 × 2 × 2 max-pooling.

The CAE consists of five such blocks that encode 128 × 128 × 128 segmentation masks into vectors of 3072 features, corresponding to a 99.9 % compression rate. The decoding path of the CAE consists of five such blocks in which the max-pooling is replaced by 3 × 3 × 3 transposed convolution with a 2 × 2 × 2 stride. The last convolutional layer produces a mask of the same dimensions as the input mask and is followed by a sigmoid activation function to ensure that the values are in $[0, 1]$. The input mask is provided as a binary mask with discrete values 0 and 1 for background and foreground, respectively.

The TPS-CNN outputs a vector of control point positions (y-coordinates) and therefore consists only of an encoding path with five encoding blocks without corresponding decoding path. For regularization, we use dropout after each block with increasing dropout probability, from 10 % to 50 %. The last layer is a dense (fully connected) layer

with N units, each representing the y-coordinate of one of the control points. The sigmoid function is applied to this vector to ensure that the predicted heights are in $[0, 1]$, corresponding to the range from lowest to highest possible value for the control point height such that the control point is placed inside the mask. The x and z coordinates of the control points are fixed.

The exact architectures of both networks, including the number and sizes of filters, the position of batch normalization and dropout operations as well as activation functions, are listed in the Appendix at the end of this chapter (Tables 5.2 to 5.4).

5.3.5 Training details

We used Adam[62] for optimization of the network parameters with fixed learning rates of 1×10^{-3} for the CAE and 1×10^{-4} for the TPS-CNN. A weight decay regularization term (L2 norm of the network parameters) was added to the loss function of both networks, weighted by a factor of 1×10^{-5}. The segmentation masks were resampled to a standard resolution of 1 mm × 1 mm × 1 mm and cubes of 128 × 128 × 128 voxels were extracted, centered at the vertebra or vertebral body. One of the CT scans with manual reference segmentation of the vertebral bodies was used for hyperparameter tuning and to determine convergence during training. The remaining five scans were used for evaluation of the method.

The MR and CT datasets cover different sections of the spine (lumbar vs. thoracic region) with differences in size, orientation and shape of the vertebrae. During training of the CAE, we therefore applied various random deformations to the segmentation masks to increase the variability of the data. The masks were subject to transformations with normally distributed magnitudes: random scaling ($\sigma = 10\,\%$), rotation ($\sigma = 20°$), elastic deformation (random deformation field $[-1, 1]$ smoothed with $\sigma = 2$ and scaled by a factor $\alpha = 10$), and shifting (uniformly distributed random shifts of maximally $d_x = 12, d_y = d_z = 24$ pixels). During training of the TPS-CNN, we applied only random rotations ($\sigma_x = 30°, \sigma_y = 15°, \sigma_z = 10°$) to half of the segmentation masks. Whether a training sample was subject to random deformation was randomly decided in each iteration with a 10 % and 50 % chance during training of the CAE and TPS-CNN, respectively, for each segmentation mask.

Moreover, to enforce a competition between the volume loss term L_v, which pushes the TPS surface toward the posterior of the vertebra, and the reconstruction loss term L_r immediately from the beginning of the training, we initialized the bias in the final layer of the TPS-CNN to a constant value of −1. Compared with initialization to 0, this initializes the position of the TPS-surface offset from the center of the image toward the anterior of the vertebra, making the initial size of the vertebral body partition small.

Table 5.1: Accuracy of the automatic vertebra partitioning, expressed as the segmentation accuracy of the vertebral body. Dice and Hausdorff distance (HD): mean ± standard deviation over the test set for the thoracic vertebrae T2–T11, which were visible in all scans, are listed. N denotes the number of control points of the TPS surface, which corresponds to a certain control point spacing Δp_i in pixels. The columns L_r, L_c and L_v specify whether the respective loss term was part of the loss.

N	Δp_i	L_r	L_c	L_v	Dice (%)	HD (mm)
64	= 16	X	X	X	98.9 ± 1.1	3.7 ± 1.6
100	= 12	X	X	X	99.1 ± 1.2	3.5 ± 1.8
256	= 8	X	X	X	99.3 ± 0.5	4.1 ± 2.3
1024	= 4	X	X	X	99.2 ± 0.5	3.7 ± 1.5
100	= 12	X	X		0.0 ± 0.0	∞
100	= 12	X		X	99.1 ± 0.8	3.8 ± 1.6
100	= 12		X	X	71.8 ± 5.8	43.7 ± 3.2

5.4 Experiments and Results

5.4.1 Evaluation

The accuracy of the predicted TPS surface with respect to the boundary between vertebral body and posterior elements was evaluated by comparing the vertebral body segmentation obtained with the partitioning with manual reference segmentations of the vertebral bodies. We used the Dice score for volume overlap and the Hausdorff surface distance to compare the segmentation masks. One of the six available reference segmentation masks of the vertebral bodies in CT scans was used for validation during training and method development. The other five masks were used for evaluation.

Additionally, we compared the volume of the vertebral bodies with the height of the vertebral body measured in 2D. The height of each vertebral body was manually measured in anterior, middle and posterior positions and the overall height was estimated as the average of these three measurements. These measurements can be performed in a few minutes per scan and served as a simple surrogate measurement for the vertebral body size, which we therefore expected to correlate well with the true 3D volume of the vertebral body.

5.4.2 Segmentation results

We compared the performance of TPS-CNNs with different numbers of control points, ranging from 64 to 1024. These were evenly spaced across the segmentation masks of 128 × 128 × 128 voxels and therefore corresponded to a spacing between control points of 16 to 4 voxels. There were no substantial performance differences, only the version with least control points (N = 64) performed slightly worse than versions with more control points (Table 5.1, top rows). Overall, the method achieved Dice scores of

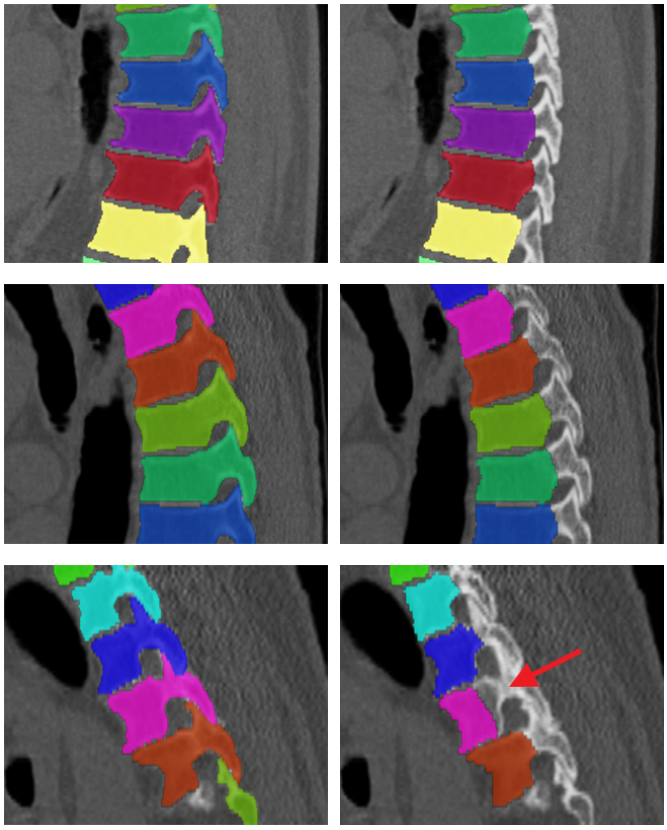
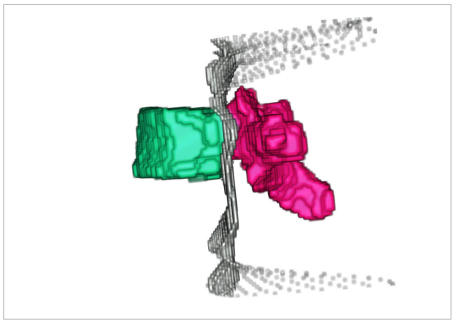


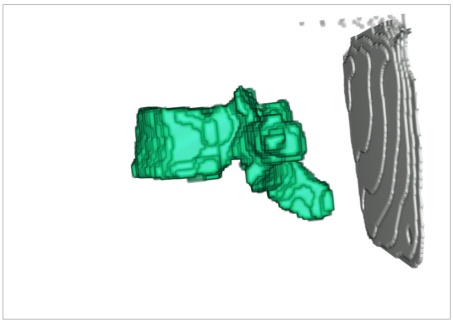
Figure 5.5: Examples of vertebra segmentation masks (left) and corresponding partitioning results, showing only the vertebral body partition (right). In the last row, the red arrow points at a partitioning error. Note that the partitioning was successful for the neighboring vertebrae. These results were generated using a TPS-CNN with N = 100 control points.

around 99 % and Hausdorff distances of 3.5 mm to 4.1 mm. Visual inspection revealed only few vertebrae with larger inaccuracies, but in the vast majority the vertebral body and posterior elements were well separated (Figure 5.5).

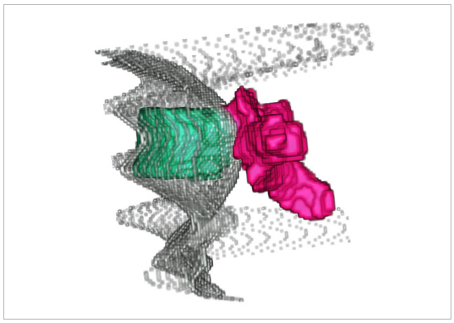
Furthermore, we evaluated the influence of the three loss components by training three TPS-CNNs in which one of the components was removed from the loss function (Table 5.1, bottom rows). Removing the reconstruction loss term L_r left no incentive for the network to place the TPS surface between vertebral body and posterior elements. Because the other two loss components reward a large vertebral body partition and low curvature of the TPS surface, the network learned to always place a flat surface behind the vertebra (Figure 5.6b). Removing the volume loss term L_v had a similar effect: with no incentive to make the vertebral body partition as large as possible, the network learned to always place the surface in front of the vertebra, resulting in an



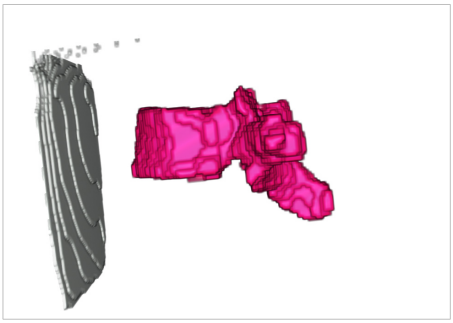
(a) Proposed



(b) Without reconstruction loss L_r



(c) Without curvature loss L_c



(d) Without volume loss L_v

Figure 5.6: Partitioning results as three-dimensional renderings for networks trained with different loss functions. The number of control points was set to N = 100 in all cases. The vertebral body partition is displayed in green, the posterior partition in red and the TPS surface in gray. Note that the predicted surfaces are not smooth in the parts where there is never any vertebral bone and hence no gradient information, i.e., at the very top and bottom of the volume. (a) The proposed loss function with reconstruction, curvature and volume loss terms results in a relatively flat surface that separates vertebral body and posterior elements well. (b) Omitting the reconstruction loss term results in the volume term pushing the surface to the right such that the entire vertebra is the vertebral body partition. (c) Omitting the curvature loss results in a smooth but curved surface that separates body and posterior elements well and that is tightly wrapped around the vertebral body. (d) The network converges to a trivial solution for the mask reconstruction by pushing the surface to the left such that the vertebral body partition is an empty mask when the volume loss is omitted.

empty vertebral body partition which the CAE perfectly reconstructed (Figure 5.6d). Removing the curvature loss term L_c had large impact on the shape of the predicted TPS surface, but did not substantially affect the segmentation performance (Figure 5.6c). The predicted surface approximated the boundary between vertebral body and posterior elements well, but did not continue as a relatively flat surface outside the vertebral mask. Rather, the surface was usually wrapped tightly around the vertebral body, resembling a bowl placed over it.

5.4.3 Correlation results

Simple measurements of the vertebral body height are clinically used to detect and quantify abnormalities, such as compression fractures.[107] Because creating reference segmentations of vertebral bodies or posterior elements, or the boundary between them, is a time-consuming and often difficult task, we used height measurements in 2D sagittal slices as estimates of vertebral body size. The anterior, middle and posterior heights of each vertebra were averaged and cubed to roughly approximate the vertebral body volume. These were then plotted against the volume of the vertebral body extracted with the proposed automatic partitioning method, and the Pearson correlation coefficient was calculated. The manually estimated and the automatically measured volumes were highly correlated (Figure 5.7, $r = 0.843$, $p < 0.001$).

The upper vertebrae are generally smaller than the lower vertebrae, and the upper vertebral bodies are accordingly also smaller than the lower vertebral bodies. The volume of the original unpartitioned vertebra segmentations is therefore also highly correlated with the vertebral body heights measured in 2D sagittal slices ($r = 0.818$, $p < 0.001$). An incorrect partitioning approach that assigns large parts of the vertebra to the vertebral body partition, such as in Figure 5.6b, would therefore also demonstrate high correlation with the height measurements. We therefore tested whether the correlation was significantly greater for the vertebral body volumes obtained with the proposed automatic partitioning method compared with the volume of the unpartitioned vertebra[180] and found that this was the case ($p < 0.001$).

5.4.4 Runtime

All experiments were performed with Nvidia Titan Xp GPUs. Training of the CAE required around 35 h to 40 h for the network to converge after around 30 000 epochs of 4 samples. Training of the TPS-CNN required around 15 h to 20 h for the network to converge after around 25 000 epochs of 2 samples. At inference time, prediction of a TPS surface took on average 44 ms. Processing all vertebrae visible on an entire chest CT scan took on average 630 ms.

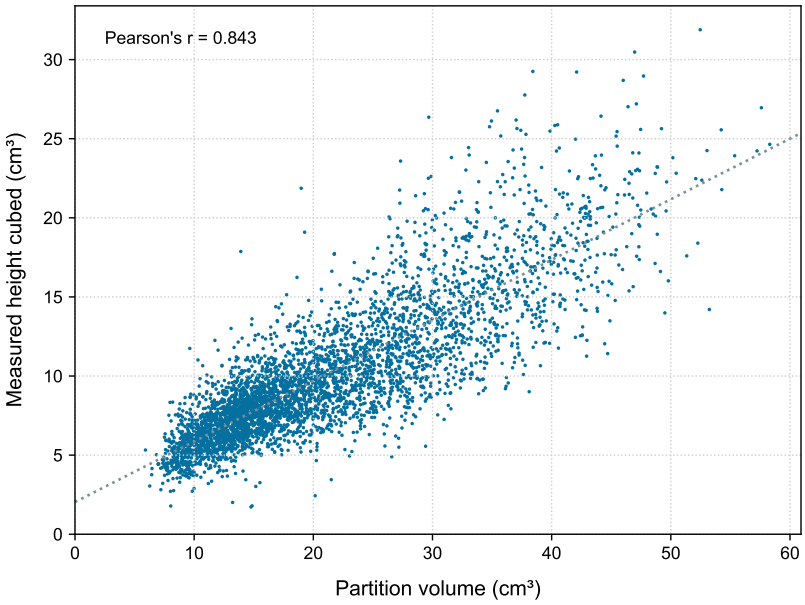


Figure 5.7: Scatter plot showing manually measured vertebral heights (cubed) as estimate of vertebral body size versus corresponding volume of the vertebral body extracted with the proposed partitioning approach. Dotted line represents a linear fit to the data. A TPS-CNN with N = 1024 control points was used to obtain the partitioning into vertebral body and posterior elements.

5.5 Discussion

There are many possible approaches to partitioning a segmentation mask into substructures, such as posing this task as a voxel labeling or surface mesh fitting problem. In this chapter, we presented a novel approach based on a CNN that controls the position and shape of a smooth TPS surface, representing the boundary between two substructures. Thin-plate splines have been used in combination with CNNs before, but mostly for interpolation, for instance in spatial transformer networks.[72] Contrarily, we used them to represent a surface in the image or segmentation mask domain. Additionally, the TPS-steering CNN was trained with the help of a CAE, and both networks were trained with segmentation masks only, enabling the use of cross-modality reference segmentation masks. Applied to partitioning vertebra segmentation masks into segmentations of the vertebral body and posterior elements, the proposed method achieved high accuracy in terms of segmentation error and high correlation with manual measurements of the vertebral body height.

The proposed TPS-CNN learns from three individual loss terms, of which the autoencoder reconstruction loss contributes information about plausible shapes of one of the partitions and therefore directly contributes information about the expected

placement of the partitioning TPS surface. The other two terms serve as regularizers. The volume loss term, which rewards a large partition size, acts adversarially to the reconstruction loss term: the volume term pushes the surface to enlarge the size of one of the partitions, while the reconstruction term pushes back once the partition becomes too large to still represent a plausible shape. We found this interplay of the two loss terms to be crucial for obtaining meaningful partitioning results. Our approach therefore has some similarity with generative adversarial networks[181] and, potentially, the adversarial loss terms could be replaced by adversarial networks, considering the TPS-CNN a generator network and replacing the CAE by a discriminator network. However, in preliminary experiments we found that such an approach was unstable and did not converge well. The second regularizing loss term rewards low curvature of the predicted TPS surface, i.e., enforces a smoother surface. We found that this improves the performance only marginally for vertebra partitioning. While we derived the curvature loss from the Hessian matrix, it could also be formulated in other ways, for instance similar to bending energy terms used for regularization of deformation fields in image registration.[182]

We did not observe large performance differences between TPS surfaces with 64 to 1024 control points. This might be attributed to the relatively simple shape of the boundary between vertebral body and posterior elements, which does not require a large number of control points to be accurately represented. However, in other applications with more complex shapes, models with more control points might outperform models with few control points.

Both the TPS-CNN and the CAE were trained with automatically obtained segmentation masks, which enabled us to create a large training set of both vertebra and vertebral body segmentation masks. However, this also required an automatic segmentation method for vertebrae and vertebral bodies in different modalities, CT and MR in our case, because if they would be available for the same modality, a partitioning approach would not be necessary. The presented method therefore offers a solution for a rather specific situation. However, TPS surfaces steered by CNNs potentially have other applications as well, such as pulmonary lobe segmentation,[183] separation of chest and abdomen by approximation of the thoracic diaphragm or separation of the left and right hemispheres of the brain. Future work could also address the use of different surface models, such as B-spline surfaces. Moreover, future work could investigate whether image information, in addition to shape information, further improves the accuracy of the partitioning.

Manual segmentation of one or both of the substructures in our example application of vertebra partitioning into vertebral body and posterior elements is time-consuming and labour-intensive. While this motivated an approach for partitioning instead of direct segmentation of the substructures, it also limited our evaluation. The partitioning accuracy was evaluated in terms of segmentation accuracy of the vertebral bodies, using manual reference segmentations of 50 vertebral bodies in chest CT scans. While

the additional comparison with height measurements in 2D sagittal slices of 4391 vertebrae itself can only give an indication of the partitioning performance, both high segmentation accuracy in terms of Dice score and Hausdorff distance, and high correlation with height measurements together demonstrate that the partitioning results are accurate and robust.

In clinical practice, partitioning vertebra segmentation masks into the vertebral body and posterior elements has several potential applications, such as automatic detection of vertebral fractures from shape characteristics of the vertebral body,[184,185] or from approximated bone mineral density of the vertebral body.[186] Another potential application could be automatic assessment of the involvement of the posterior elements in traumatic fractures, which is important for deciding whether surgery is needed.[187] Especially with regard to the current practice of visual grading or simplified measurement of vertebral shape characteristics, automatic methods might be able to provide improved robustness and more objective measures.

In conclusion, CNNs can perform segmentation and partitioning tasks not only as voxel classification tasks, but also by steering a TPS surface by predicting the position of its control points. Autoencoders can be used to train these networks even if no matching reference segmentation pairs are available for a partitioning task.

Acknowledgments

We are grateful to the United States National Cancer Institute (NCI) for providing access to NCI's data collected by the National Lung Screening Trial. The statements contained herein are solely ours and do not represent or imply concurrence or endorsement by NCI. We also thank Dr. Wouter Veldhuis (UMC Utrecht) for his help with collecting the set of MR scans.

Appendix

Table 5.2: Architecture of the TPS-steering CNN

Operation	Feature map size
32x 3 × 3 × 3 convolution	128 × 128 × 128
ReLU	
32x 3 × 3 × 3 convolution	
ReLU	
Dropout 10 %	
2 × 2 max-pooling	64 × 64 × 64
32x 3 × 3 × 3 convolution	
ReLU	
64x 3 × 3 × 3 convolution	
ReLU	
Dropout 20 %	
2 × 2 max-pooling	32 × 32 × 32
64x 3 × 3 × 3 convolution	
ReLU	
128x 3 × 3 × 3 convolution	
ReLU	
Dropout 30 %	
2 × 2 max-pooling	16 × 16 × 16
127x 3 × 3 × 3 convolution	
ReLU	
256x 3 × 3 × 3 convolution	
ReLU	
Dropout 40 %	
2 × 2 max-pooling	8 × 8 × 8
256x 3 × 3 × 3 convolution	
ReLU	
512x 3 × 3 × 3 convolution	
ReLU	
Dropout 50 %	
2 × 2 max-pooling	4 × 4 × 4
256x 3 × 3 × 3 convolution	
ReLU	
16x 3 × 3 × 3 convolution	
ReLU	
Dense (fully connected) + bias	N
Sigmoid	

N denotes the number of control points.

Table 5.3: Architecture of the encoder of the convolutional autoencoder

Operation	Feature map size
56x 3 × 3 × 3 convolution	128 × 128 × 128
Batch normalization	
ReLU	
56x 3 × 3 × 3 convolution	
Batch normalization	
ReLU	
2 × 2 max-pooling	64 × 64 × 64
56x 3 × 3 × 3 convolution	
Batch normalization	
ReLU	
56x 3 × 3 × 3 convolution	
Batch normalization	
ReLU	
2 × 2 max-pooling	32 × 32 × 32
56x 3 × 3 × 3 convolution	
Batch normalization	
ReLU	
56x 3 × 3 × 3 convolution	
Batch normalization	
ReLU	
2 × 2 max-pooling	16 × 16 × 16
56x 3 × 3 × 3 convolution	
Batch normalization	
ReLU	
56x 3 × 3 × 3 convolution	
Batch normalization	
ReLU	
2 × 2 max-pooling	8 × 8 × 8
56x 3 × 3 × 3 convolution	
Batch normalization	
ReLU	
56x 3 × 3 × 3 convolution	
Batch normalization	
ReLU	
2 × 2 max-pooling	4 × 4 × 4
28x 3 × 3 × 3 convolution	
Batch normalization	
ReLU	
48x 1 × 1 × 1 convolution + bias	3072
Tanh	

Table 5.4: Architecture of the decoder of the convolutional autoencoder

Operation	Feature map size
28x 1 × 1 × 1 convolution + bias	4 × 4 × 4
ReLU	
56x 3 × 3 × 3 convolution	
Batch normalization	
ReLU	
56x 3 × 3 × 3 up-convolution* + bias	8 × 8 × 8
ReLU	
56x 3 × 3 × 3 convolution	
Batch normalization	
ReLU	
56x 3 × 3 × 3 convolution	
Batch normalization	
ReLU	
56x 3 × 3 × 3 up-convolution* + bias	16 × 16 × 16
ReLU	
56x 3 × 3 × 3 convolution	
Batch normalization	
ReLU	
56x 3 × 3 × 3 convolution	
Batch normalization	
ReLU	
56x 3 × 3 × 3 up-convolution* + bias	32 × 32 × 32
ReLU	
56x 3 × 3 × 3 convolution	
Batch normalization	
ReLU	
56x 3 × 3 × 3 convolution	
Batch normalization	
ReLU	
56x 3 × 3 × 3 up-convolution* + bias	64 × 64 × 64
ReLU	
56x 3 × 3 × 3 convolution	
Batch normalization	
ReLU	
56x 3 × 3 × 3 convolution	
Batch normalization	
ReLU	
1x 3 × 3 × 3 up-convolution* + bias	128 × 128 × 128
Sigmoid	

*Transposed convolution with 2 × 2 × 2 stride and padding.

6

CHAPTER 6

Summary and discussion

Low-dose chest CT scans allow for quantification of various pulmonary and extrapulmonary diseases. The previous chapters describe methods for automatic quantification of extrapulmonary diseases, namely cardiovascular disease by means of calcium scoring (Chapters 2 and 3) and osteoporosis by means of vertebral shape analysis (Chapters 4 and 5). This chapter summarizes the methods and findings presented in these chapters and discusses implications and possible directions for future research.

6.1 Summary

6.1.1 Calcium scoring in chest CT

Manual quantification of the amount of arterial calcification in CT scans is often a tedious task, especially in chest CT scans, which are not optimized for cardiac imaging and therefore frequently contain cardiac motion artifacts. **CHAPTER 2** describes a method for automatic calcification quantification in chest CT scans using convolutional neural networks. A network using dilated convolutions, a special type of convolution that enables the network to include information from a large area of the image, searches the CT image for potential calcifications and determines their anatomical location, i.e., labels them according to the calcified artery or valve. A second network with conventional convolutions, and hence focused on a smaller area of the image, subsequently analyzes the detected calcification candidates to discard false-positive detections of the first network. The approach was trained and evaluated with a large and deliberately diverse set of manually annotated low-dose chest CT scans, focusing especially on performance differences in scans reconstructed with different filtered backprojection filter kernels. Several kernels that are commonly used for reconstruction of chest CT scans result in low signal to noise ratio, which is disadvantageous for calcification quantification. We showed that our method can reliably detect calcifications of the coronary arteries, the thoracic aorta and the cardiac valves regardless of CT reconstruction settings.

This method was used in **CHAPTER 3** to perform a retrospective case-control study with data from the low-dose chest CT arm of a randomized controlled lung cancer screening trial. We investigated sex differences in the prevalence and severity of arterial calcification, and whether there are differences in the association of calcium scores and cardiovascular mortality between male and female heavy smokers. Automatic calcification detection and calcium scores computation allowed for inclusion of a large study population. The study confirmed that calcification of the coronary arteries is much more common and severe in men, but revealed that no such prominent difference exists between men and women for calcification of the thoracic aorta. Furthermore, calcification of the coronary arteries was strongly associated with coronary heart disease mortality, while calcification of the thoracic aorta was mainly associated with mortality from other cardiovascular diseases, such as stroke.

6.1.2 Vertebral analysis in chest CT

First steps toward analysis of the vertebrae are their detection and segmentation in the image, and their anatomical identification. **CHAPTER 4** introduces an automatic method based on an iteratively applied convolutional neural network. This method searches the image in a sliding-window fashion until it finds the spine, then traverses along the spine to segment the vertebrae one after the other. The vertebrae are additionally anatomically labeled and examined for completeness, since in many types of scans several vertebrae are only partially visible. Using a global maximum likelihood model, the anatomical labels of all vertebrae in a scan are used to find the most plausible anatomical labeling. We trained and evaluated this method with low-dose chest CT scans, but also with dedicated spine CT and MR scans from four publicly available datasets. The method performed comparable or better than previous methods on these datasets, completing vertebra segmentation in CT and vertebral body segmentation in MR in only a few seconds.

Markers for osteoporosis and osteoporotic fracture risk, such as height loss or reduced bone density, are measurements of only part of the vertebra, the vertebral body. Therefore, **CHAPTER 5** presents a method for partitioning of segmentation masks into two substructures, applied to partitioning of vertebra segmentations into the vertebral body and the posterior elements. The method uses thin-plate splines to model the boundary between the two structures. A convolutional neural network analyzes the segmentation mask and predicts the location of control points defining the thin-plate spline surface. The loss function used to train this network is based on the reconstruction error of the extracted vertebral body segmentations when they are encoded and reconstructed by a convolutional autoencoder network trained with segmentation masks of vertebral bodies. This additional network enables the use of non-matching vertebra and vertebral body segmentation masks for training. We demonstrated that this method accurately predicts the boundary between vertebral body and posterior elements, and that the volume of the extracted vertebral bodies correlates well with height measurements.

6.2 Discussion and future directions

6.2.1 Validation

The research presented in this thesis explores the idea of using chest CT scans obtained for pulmonary imaging, in particular lung cancer screening, for early detection of extrapulmonary diseases and quantification of related biomarkers. We developed automatic image analysis algorithms, which are only useful for routine analysis of chest CT scans and retrospective analysis of existing datasets if they are carefully validated. In Chapters 2, 4 and 5, we therefore ascribed special importance to validation of the developed algorithms and assembled large and deliberately diverse validation

datasets. The results demonstrate that the presented methods allow for robust and precise calcium scoring as well as vertebra segmentation and partitioning in (low-dose) chest CT.

These methods are predominantly based on supervised machine learning. A large part of the manually annotated data was therefore used as training data. With the exception of a small external dataset in Chapter 4, we performed only an internal validation of the methods, i.e., we used a subset of the available data for training and the remaining data for evaluation. A next step toward routine use of these methods would therefore be an external validation using data from different populations and from hospitals that did not contribute to the training data.

Preliminary external validation results with non-screening chest CT scans indicate that the automatic calcium scoring method from Chapter 2 generalizes well to data with acute pulmonary pathologies.[188] However, the method made mistakes especially in abnormal cases, for instance in scans of patients who had pneumonectomy or lobectomy of the lung before the scan. Examples of such cases were not part of the training data, the NLST explicitly excluded these subjects from participation in the trial.[23] Further external validation will therefore likely confirm this limitation of our supervised approaches and will find that they produce unpredictable results for cases, or parts of scans, with very different characteristics than the training data.

However, a strength of supervised approaches is their versatility. The methods presented in this thesis can be adapted to different types of data by retraining with representative data, without requiring any changes to the method itself. In Chapter 4, we also demonstrated that the vertebra segmentation and identification method can learn from a mixed dataset with various types of images, including dedicated spine CT scans but also low-dose chest CT scans. The method was able to outperform previous state-of-the-art methods on most datasets without the need for a separate method for each dataset, or dataset-specific adaptations or retraining. Other publications have confirmed that deep learning methods are capable of handling a variety of data in a single model.[174] Given that calcium scoring can be performed in many different types of CT scans as well, for instance radiotherapy planning CT scans of breast cancer patients,[189] it would be desirable to retrain and validate the calcium scoring method from Chapter 2 with a broad training set consisting of all kinds of CT scans visualizing the heart. For the vertebra segmentation and partitioning methods (Chapters 4 and 5), a broader training set could include acute trauma cases with fractures as well as cases with pedicle screws and other implants. Ideally, this would result in a single calcium scoring method and a single vertebra segmentation method that work regardless of the image type.

The methods for calcium scoring (Chapter 2) and vertebra segmentation (Chapter 4) reached an accuracy close to the level of interobserver agreement, i.e., when using the annotations of one observer as the reference standard, a second observer performed only marginally better than the automatic method. However, there were still cases

where the automatic results contained mistakes that a human observer would hardly ever make (e.g., Figure 2.6k and Figure 4.4c). At the same time, perfect reference annotations are virtually impossible to obtain. The datasets used for training and evaluation certainly contain many small and possibly some larger mistakes, caused by distracted observers, incorrect use of the annotation software or simply incorrect interpretation of the image. The quality of reference annotations is important for training of supervised machine learning algorithms[190] and for reliable evaluation of their performance, but can be challenging to improve. Future work could address these issues, for instance by trying to automatically identify incorrectly labeled voxels in the reference annotations[191] or by including indications of uncertainty in the output.[38,192]

6.2.2 Application in clinical care and research

Routine reporting of calcium scores on ungated chest CT scans, especially non-screening scans, is increasingly advocated and has become part of expert statements and guidelines.[101] Automatic calcium scoring might therefore become an important assistive technology that enables reporting of exact calcium scores not only in screening programs but also in routine care. However, there is currently insufficient evidence for the effectiveness of opportunistic screening for cardiovascular disease and osteoporosis with chest CT. Large prospective clinical trials like the ROBINSCA[193] and the BRAGATSTON[194] trial have the potential to provide evidence for screening for cardiovascular disease with risk scores that include CT-derived calcium scores, in the BRAGATSTON trial even derived from non-cardiac scans. While screening for osteoporosis is already being recommended for women using bone mineral density measurements from dual-energy X-ray absorptiometry,[21] it is currently not well established whether disease markers derived from chest CT are equally suited predictors. However, for both diseases, the methods developed in this thesis have the potential to facilitate clinical research in large chest CT datasets. Such studies can help to better understand the predictive value of disease markers and their epidemiology, similar to our study of sex differences in arterial calcification (Chapter 3).

The study in Chapter 3 used the automatic calcium scoring method described in Chapter 2 to extract calcium scores from a large set of images for clinical research. Similarly, the vertebra segmentations and partitioning results that can be obtained with the methods described in Chapters 4 and 5 could be used to investigate vertebral characteristics. Relevant characteristics could be density, volume and shape. Segmentation masks and image information could also be combined input for a fracture prediction model. In current clinical practice, osteoporotic vertebral compression fractures are often diagnosed and classified using Genant's semi-quantitative method, which is based on the height loss of the vertebral body in comparison with a neighboring vertebra.[107] The height loss is either visually estimated or measured with digital calipers on 2D image slices. There is inevitably an unnecessary inaccuracy in extracting information

from only a few local 2D measurements when a 3D scan is available: measurements derived from 3D segmentations have the potential to be more accurate and reproducible.

Simple visual estimates of the calcification burden, often referred to as ordinal calcium scoring, have also been advocated and have been shown to be useful predictors of cardiovascular events.[29] This scoring method could be a viable alternative to manual calcium scoring with much lower manual effort. However, especially in a screening context in which participants typically return regularly for another CT scan, an exact quantification could be advantageous for continuous monitoring of the calcification burden. For this purpose, the reliability of automatically determined calcium scores in ungated CT scans could be further improved by combining the method in Chapter 2 with methods that address partial volume effects and motion artifacts.[195,196]

6.2.3 Methodological contributions

With the exception of the epidemiological study in Chapter 3, the methods introduced in this thesis rely predominantly on machine learning, particularly deep learning and convolutional neural networks. In contrast to most other machine learning methods, convolutional neural networks can be directly applied to images, or parts of images, and do not require the prior extraction of a set of features from the image. This affects the method development and shifts the attention from feature engineering to engineering of the network architecture, the inputs and outputs of the network, and the loss function that controls how the network is optimized.

In Chapter 2, we cascaded two networks to combine a focus on context information with a focus on local details. A focus on context information was achieved by constructing one of the convolutional networks with dilated convolutions, a special type of convolution that allows for a large receptive field with low number of parameters, which helps to prevent overfitting to the training data. The two networks were trained independently, and the second network mainly contributed by reducing the false positive rate of the first network. We therefore expect that more sophisticated network architectures or training strategies, for example adversarial training,[197] can eventually replace the two-network design.

In Chapter 4, we introduced a new strategy for instance segmentation with convolutional neural networks. A single network was used to produce various outputs, including a segmentation mask, a regression output and a probability for binary classification. This design greatly simplifies both training and inference, and we expect that similar multi-task networks will be more often chosen over cascaded networks in the future. Furthermore, deep learning instance segmentation methods are currently not used much for medical image analysis, even though there are many medical instance segmentation problems, e.g., all kinds of lesion detection tasks, where they could be useful.

In Chapter 5, we demonstrated that convolutional neural networks can perform

image-to-mask or mask-to-mask prediction tasks not only by performing voxel classification but also by predicting the position of a smooth surface that represents the boundary between two objects. We applied this approach to segmentation masks of vertebrae to partition them into substructures, but it can also be directly applied to images to predict a boundary in the image.

6.2.4 Future directions

Heavy smokers who become eligible for lung cancer screening by reaching the minimum age or number of pack years will typically stay eligible for a number of years. These people are invited to return regularly while they are eligible, usually annually, to receive another low-dose chest CT scan, even if no suspicious lesions were found on the last exam. Thus, temporal imaging data is available for most screening participants. This aspect has not been considered in this thesis, but is a promising research direction. Temporal data can both contribute to the understanding of disease development and to risk prediction aimed at identifying subjects who would benefit from preventive measures.

The current hype around deep learning and artificial intelligence, paired with the advent of methods that perform comparable to or even better than human experts in specific (tedious) tasks, has the potential to lead to an increased use of automatic image analysis as part of routine reading of chest CT scans. Many other disease markers can be quantified on chest CT, such as other extrapulmonary diseases like hypertension, which can be quantified by measuring the size of the pulmonary artery,[198] and various pulmonary diseases.[199] This creates opportunities for personalized medicine based on a whole array of measurements automatically extracted from a single chest CT scan. Combined consideration of all of this information could enable identification of those conditions that need therapeutic or preventive measures, or that need closer monitoring. Given the amount of data that can be extracted from a single scan in addition to data related to its original purpose, and the even larger amount of data that a time series of CT scans contains, computer-based analysis of this data is another important direction of future research.

Nederlandse samenvatting

Longkanker is het type kanker waar wereldwijd de meeste mensen aan overlijden.[1] Omdat longkanker meestal veroorzaakt wordt door roken, zijn er in een aantal landen screeningsprogramma's opgezet voor zware rokers en worden in veel andere landen dergelijke screeningsprogramma's overwogen of momenteel onderzocht. De longkankerscreening maakt gebruik van thorax computertomografie (CT) met lage stralingsdoses. Op deze beelden kan longkanker worden opgespoord terwijl deze zich nog in een vroeg stadium bevindt, zodat er nog meer en betere behandelingsopties beschikbaar zijn.

Roken is echter ook een belangrijke risicofactor voor vele andere chronische ziekten, zoals COPD[12] en hart- en vaatziekten.[13] Veel van deze ziekten kunnen ook aan de hand van thorax CT beelden geïdentificeerd en gekwantificeerd worden. De CT beelden die worden gebruikt voor longkankerscreening bieden daarom de mogelijkheid om gelijktijdig deze andere ziekten te evalueren.

Er zijn echter miljoenen huidige en voormalige zware rokers die voldoen aan de criteria voor longkankerscreening.[22] Het zou voor radiologen een grote additionele belasting betekenen om naast longkanker ook naar andere ziekten te kijken en deze te kwantificeren. Het doel van dit proefschrift was daarom om hier automatische methoden voor te ontwikkelen. We hebben ons daarbij gericht op twee soorten extrapulmonaire ziekten, namelijk hart- en vaatziekten en osteoporose.

Automatische kwantificering van hart- en vaatziekten

Hart- en vaatziekten behoren tot de meest voorkomende doodsoorzaken van zware rokers.[4] In thorax CT beelden kan het risico op hart- en vaatziekten worden voorspeld door de hoeveelheid arteriële verkalking te meten. In **HOOFDSTUK 2** hebben we daarom een softwarealgoritme ontwikkeld dat automatisch een CT scan analyseert om de hoeveelheid verkalking (de zogeheten calciumscore) te meten. Dit algoritme is, net als alle andere algoritmes die in dit proefschrift worden gepresenteerd, gebaseerd op *machine learning*. Twee opeenvolgende convolutionele neurale netwerken analyseren het beeld. Het eerste netwerk is gericht op het detecteren van mogelijke verkalkingen en het bepalen van hun anatomische locaties. Het netwerk bepaalt of een van de kransslagaders, de aorta of een van de hartkleppen verkalkt is en gebruikt daarvoor een speciaal type convoluties, de zogenaamde gedilateerde convoluties. Hiermee kan het netwerk informatie uit een groot deel van het CT beeld combineren. Een tweede netwerk, gebaseerd op conventionele convoluties en dus gericht op informatie uit een

kleiner deel van het beeld, analyseert vervolgens de gedetecteerde kandidaten om foutpositieve detecties te herkennen en te verwerpen. Dit algoritme werd geëvalueerd met een groot aantal thorax CT beelden van een longkankerscreening-onderzoek waarin verkalkingen handmatig werden geannoteerd.

In **HOOFDSTUK 3** hebben we dit algoritme gebruikt om een retrospectieve case-control studie uit te voeren met data van een grootschalig Amerikaans longkanker-screeningsonderzoek. Dankzij de automatische beeldanalyse in plaats van handmatige beeldinterpretatie, was het mogelijk om een groot aantal CT beelden te analyseren. We hebben onderzocht of de mate van arteriële verkalking en de associatie tussen calciumscores en het risico om aan hart- en vaatziekten te overlijden verschillen tussen mannen en vrouwen. Deze studie toonde aan dat mannen vaker en meer verkalking in hun kransslagaders hebben. Daarentegen hebben we geen verschil gevonden tussen mannen en vrouwen wat betreft de frequentie en hoeveelheid verkalking in de aorta. Verder was in deze studiepopulatie verkalking van de kransslagaders sterk geassocieerd met coronaire hartaandoeningen, terwijl calcificatie van de thoracale aorta vooral geassocieerd was met sterfte door andere hart- en vaatziekten, zoals beroerte.

Automatische kwantificering van osteoporose

Roken is tevens een risicofactor voor osteoporose en daardoor veroorzaakte wervelfacturen.[18] Als dit soort wervelfracturen borstwervels betreft, zijn deze te zien op thorax CT beelden. Verder kunnen op de CT beelden verschillende markers voor osteoporose en een verhoogd risico op fracturen worden gekwantificeerd, zoals bijvoorbeeld botmineraaldichtheid. Een eerste stap naar automatische analyse van de wervels is het detecteren, segmenteren en anatomisch labelen van de wervels in de CT scan. In **HOOFDSTUK 4** hebben we daarom een softwarealgoritme ontwikkeld dat dit automatisch doet door gebruik te maken van iteratieve convolutionele neurale netwerken. Het algoritme doorzoekt de CT scan stukje bij beetje tot het de wervelkolom vindt en loopt dan langs de wervelkolom om de wervels een voor een te segmenteren. De voorspelde anatomische labels van alle gedetecteerde wervels worden gecombineerd in een globaal *maximum-likelihood* model om de meest plausibele anatomische labeling te vinden. Dit algoritme werd geëvalueerd met zowel een dataset van thorax CT scans als vier andere datasets bestaande uit CT of MRI scans van verschillende delen van de wervelkolom, en presteerde vergelijkbaar met of beter dan andere methoden.

Biomarkers voor osteoporose zijn vaak metingen van slechts een deel van een wervel, bijvoorbeeld het wervellichaam. In **HOOFDSTUK 5** hebben we daarom een softwarealgoritme ontwikkeld dat een wervelsegmentatie in twee delen splitst, namelijk het wervellichaam en de overige delen erachter. Dit algoritme is gebaseerd op het representeren van de grens tussen de twee structuren als een glad oppervlak. Daarvoor worden *thin-plate splines* gebruikt. Deze worden geplaatst door een convolutioneel neuraal netwerk dat de segmentatie van een wervel analyseert en de locaties voorspelt

van een set controlepunten die het oppervlak definiëren. Een tweede netwerk, een *autoencoder* netwerk, wordt gebruikt om dit netwerk met handmatige segmentaties van wervellichamen te trainen. Door gebruik te maken van een *autoencoder* kan het netwerk getraind worden met segmentaties van wervels en wervellichamen uit verschillende scans.

Conclusies en toekomstperspectieven

Het onderzoek in dit proefschrift was gericht op het ontwikkelen van softwarealgoritmes voor de kwantificering van extrapulmonaire ziekten in thorax CT scans. Het in hoofdstuk 2 beschreven algoritme maakt het routinematige kwantificeren van hart- en vaatziekten in de vorm van calciumscores mogelijk. De in de hoofdstukken 4 en 5 beschreven algoritmes maken het routinematige kwantificeren van botmineraaldichtheid en andere biomarkers voor osteoporose mogelijk. Verder maken deze algoritmes ook retrospectieve analysen van bestaande grote datasets mogelijk, zoals onze studie naar sekseverschillen bij arteriële verkalkingen in hoofdstuk 3. Het kwantificeren van andere ziekten naast longkanker in CT beelden die oorspronkelijk zijn gemaakt voor longkankerscreening kan mogelijk helpen bij het identificeren van aandoeningen die behandeling of nadere controle vereisen en dus bijdragen aan het afstemmen van preventieve maatregelen op elke individuele patiënt.

Bibliography

1. F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal. "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries", *CA: A Cancer Journal for Clinicians*, vol. 68 (2018), pp. 394–424 (cited on pp. 8, 115).

2. The International Early Lung Cancer Action Program Investigators. "Survival of patients with stage I lung cancer detected on CT screening", *New England Journal of Medicine*, vol. 355 (2006), pp. 1763–71 (cited on p. 8).

3. V. P. Doria-Rose and E. Szabo. "Screening and prevention of lung cancer", in *Lung Cancer: A Multidisciplinary Approach to Diagnosis and Management*, edited by K. H. Kernstine and K. L. Reckamp. New York: Demos Medical Publishing, 2010, pp. 53–72 (cited on p. 8).

4. The National Lung Screening Trial Research Team. "Reduced lung-cancer mortality with low-dose computed tomographic screening", *New England Journal of Medicine*, vol. 365 (2011), pp. 395–409 (cited on pp. 8, 15, 17, 39, 68, 115).

5. P. F. Pinsky. "Lung cancer screening with low-dose CT: a world-wide view", *Translational Lung Cancer Research*, vol. 7 (2018), pp. 234–242 (cited on p. 8).

6. V. A. Moyer. "Screening for lung cancer: U.S. Preventive Services Task Force recommendation statement", *Annals of Internal Medicine*, vol. 160 (2014), pp. 330–338 (cited on p. 8).

7. Q. Zhou, Y. Fan, H. Bu, Y. Wang, N. Wu, Y. Huang, G. Wang, X. Wang, and Y. Qiao. "China national lung cancer screening guideline with low-dose computed tomography (2015 version)", *Thoracic Cancer*, vol. 6 (2015), pp. 812–818 (cited on p. 8).

8. J. H. Pedersen and H. Ashraf. "Implementation and organization of lung cancer screening", *Annals of Translational Medicine*, vol. 4 (2016), pp. 152–152 (cited on p. 8).

9. A. J. Alberg and J. M. Samet. "Epidemiology of lung cancer", *Chest*, vol. 123 (2003), pp. 21S–49S (cited on p. 8).

10. M. Ezzati and A. D. Lopez. "Estimates of global mortality attributable to smoking in 2000", *The Lancet*, vol. 362 (2003), pp. 847–852 (cited on p. 8).

11. D. R. Youlden, S. M. Cramb, and P. D. Baade. "The international epidemiology of lung cancer: geographical distribution and secular trends", *Journal of Thoracic Oncology*, vol. 3 (2008), pp. 819–831 (cited on p. 8).

12. C. B. Sherman. "The health consequences of cigarette smoking: pulmonary diseases", *Medical Clinics of North America*, vol. 76 (1992), pp. 355–375 (cited on pp. 8, 115).

13. P. E. McBride. "The health consequences of smoking: cardiovascular diseases", *Medical Clinics of North America*, vol. 76 (1992), pp. 333–353 (cited on pp. 8, 115).

14. World Health Organization. "Global status report on noncommunicable diseases", 2014 (cited on p. 8).

15. P. C. Jacobs, M. J. A. Gondrie, Y. van der Graaf, H. J. de Koning, I. Išgum, B. van Ginneken, and W. P. Th. M. Mali. "Coronary artery calcium can predict all-cause mortality and cardiovascular events on low-dose CT screening for lung cancer", *American Journal of Roentgenology*, vol. 198 (2012), pp. 505–511 (cited on pp. 8, 15, 39, 42).

16. R. Detrano, A. D. Guerci, J. J. Carr, D. E. Bild, G. Burke, A. R. Folsom, K. Liu, S. Shea, et al. "Coronary calcium as a predictor of coronary events in four racial and ethnic groups", *New England Journal of Medicine*, vol. 358 (2008), pp. 1336–45 (cited on pp. 9, 39).

17. P. C. Jacobs, M. Prokop, Y. van der Graaf, M. J. Gondrie, K. J. Janssen, H. J. de Koning, I. Išgum, R. J. van Klaveren, M. Oudkerk, B. van Ginneken, et al. "Comparing coronary artery calcium and thoracic aorta calcium for prediction of all-cause mortality and cardiovascular events on low-dose non-gated computed tomography in a high-risk population of heavy smokers", *Atherosclerosis*, vol. 209 (2010), pp. 455–462 (cited on pp. 9, 16, 39, 50).

18. K. D. Ward and R. C. Klesges. "A meta-analysis of the effects of cigarette smoking on bone mineral density", *Calcified Tissue International*, vol. 68 (2001), pp. 259–270 (cited on pp. 9, 116).

19. C. F. Buckens, Y. van der Graaf, H. M. Verkooijen, W. P. Mali, I. Išgum, C. P. Mol, H. J. Verhaar, R. Vliegenthart, et al. "Osteoporosis markers on low-dose lung cancer screening chest computed tomography scans predict all-cause mortality", *European Radiology*, vol. 25 (2015), pp. 132–139 (cited on p. 9).

20. E. A. P. M. Romme, J. T. Murchison, K. F. Phang, F. H. Jansen, E. P. A. Rutten, E. F. M. Wouters, F. W. J. M. Smeenk, E. J. R. van Beek, and W. MacNee. "Bone attenuation on routine chest CT correlates with bone mineral density on DXA in patients with COPD", *Journal of Bone and Mineral Research*, vol. 27 (2012), pp. 2338–43 (cited on p. 9).

21. US Preventive Services Task Force. "Screening for osteoporosis to prevent fractures", *JAMA*, vol. 319 (2018), pp. 2521–31 (cited on pp. 9, 111).

22. P. F. Pinsky and C. D. Berg. "Applying the National Lung Screening Trial eligibility criteria to the US population: what percent of the population and of incident lung cancers would be covered?", *Journal of Medical Screening*, vol. 19 (2012), pp. 154–156 (cited on pp. 10, 115).

23. The National Lung Screening Trial Research Team. "The National Lung Screening Trial: overview and study design", *Radiology*, vol. 258 (2011), pp. 243–253 (cited on pp. 10, 39, 87, 110).

24. C. A. van Iersel, H. J. de Koning, G. Draisma, W. P. Mali, E. T. Scholten, K. Nackaerts, M. Prokop, J. D. F. Habbema, M. Oudkerk, and R. J. van Klaveren. "Risk-based selection from the general population in a screening trial: selection criteria, recruitment and power for the Dutch-Belgian randomised lung cancer multi-slice CT screening trial (NELSON)", *International Journal of Cancer*, vol. 120 (2006), pp. 868–874 (cited on p. 10).

25. E. A. Regan, J. E. Hokanson, J. R. Murphy, B. Make, D. A. Lynch, T. H. Beaty, D. Curran-Everett, E. K. Silverman, and J. D. Crapo. "Genetic epidemiology of COPD (COPDGene) study design", *COPD: Journal of Chronic Obstructive Pulmonary Disease*, vol. 7 (2010), pp. 32–43 (cited on p. 10).

26. S. Vollset, A. Tverdal, and H. Gjessing. "Smoking and deaths between 40 and 70 years of age in women and men", *Annals of Internal Medicine*, vol. 144 (2006), pp. 381–389 (cited on p. 15).

27. H. S. Hecht. "Coronary artery calcium scanning: past, present, and future", *JACC: Cardiovascular Imaging*, vol. 8 (2015), pp. 579–596 (cited on pp. 15, 40, 50).

28. J. Shemesh, C. I. Henschke, D. Shaham, R. Yip, A. O. Farooqi, M. D. Cham, D. I. McCauley, M. Chen, et al. "Ordinal scoring of coronary artery calcifications on low-dose CT scans of the chest is predictive of death from cardiovascular disease", *Radiology*, vol. 257 (2010), pp. 541–548 (cited on p. 15).

29. C. Chiles, F. Duan, G. W. Gladish, J. G. Ravenel, S. G. Baginski, B. S. Snyder, S. DeMello, S. S. Desjardins, and R. F. Munden. "Association of coronary artery calcification and mortality in the National Lung Screening Trial: a comparison of three scoring methods", *Radiology*, vol. 276 (2015), pp. 82–90 (cited on pp. 15, 39, 42, 112).

30. O. M. Mets, R. Vliegenthart, M. J. Gondrie, M. A. Viergever, M. Oudkerk, H. J. de Koning, W. P. Mali, M. Prokop, et al. "Lung cancer screening CT-based prediction of cardiovascular events", *JACC: Cardiovascular Imaging*, vol. 6 (2013), pp. 899–907 (cited on pp. 15, 39).

31. H. S. Hecht, C. Henschke, D. Yankelevitz, V. Fuster, and J. Narula. "Combined detection of coronary artery disease and lung cancer", *European Heart Journal*, vol. 35 (2014), pp. 2792–96 (cited on p. 15).

32. I. Išgum, A. Rutten, M. Prokop, and B. van Ginneken. "Detection of coronary calcifications from computed tomography scans for automated risk assessment of coronary artery disease", *Medical Physics*, vol. 34 (2007), pp. 1450–61 (cited on p. 15).

33. U. Kurkure, D. R. Chittajallu, G. Brunner, Y. H. Le, and I. A. Kakadiaris. "A supervised classification-based method for coronary calcium detection in non-contrast CT", *International Journal of Cardiovascular Imaging*, vol. 26 (2010), pp. 817–828 (cited on p. 15).

34. G. Brunner, D. R. Chittajallu, U. Kurkure, and I. A. Kakadiaris. "Toward the automatic detection of coronary artery calcification in non-contrast computed tomography data", *International Journal of Cardiovascular Imaging*, vol. 26 (2010), pp. 829–838 (cited on p. 15).

35. J. Wu, G. Ferns, J. Giles, and E. Lewis. "A fully automated multi-modal computer aided diagnosis approach to coronary calcium scoring of msct images", *Medical Imaging*, vol. 8314 of *Proceedings of SPIE* (2012), 83142Y (cited on p. 15).

36. X. Ding, P. J. Slomka, M. Diaz-Zamudio, G. Germano, D. S. Berman, D. Terzopoulos, and D. Dey. "Automated coronary artery calcium scoring from non-contrast CT using a patient-specific algorithm", *Medical Imaging*, vol. 9413 of *Proceedings of SPIE* (2015), 94132U (cited on p. 15).

37. R. Shahzad, T. van Walsum, M. Schaap, A. Rossi, S. Klein, A. C. Weustink, P. J. de Feyter, L. J. van Vliet, and W. J. Niessen. "Vessel specific coronary artery calcium scoring: an automatic system", *Academic Radiology*, vol. 20 (2013), pp. 1–9 (cited on p. 15).

38. J. M. Wolterink, T. Leiner, R. A. P. Takx, M. A. Viergever, and I. Išgum. "Automatic coronary calcium scoring in non-contrast-enhanced ECG-triggered cardiac CT with ambiguity detection", *IEEE Transactions on Medical Imaging*, vol. 34 (2015), pp. 1867–78 (cited on pp. 15, 111).

39. D. Dey, V. Y. Cheng, P. J. Slomka, R. Nakazato, A. Ramesh, S. Gurudevan, G. Germano, and D. S. Berman. "Automated 3-dimensional quantification of noncalcified and calcified coronary plaque from coronary CT angiography", *Journal of Cardiovascular Computer Tomography*, vol. 3 (2009), pp. 372–382 (cited on p. 15).

40. S. Wesarg, M. F. Khan, and E. A. Firle. "Localizing calcifications in cardiac CT data sets using a new vessel segmentation approach", *Journal of Digital Imaging*, vol. 19 (2006), pp. 249–257 (cited on p. 15).

41. D. Eilot and R. Goldenberg. "Fully automatic model-based calcium segmentation and scoring in coronary CT angiography", *International Journal of Computer Assisted Radiology and Surgery*, vol. 9 (2014), pp. 595–608 (cited on p. 15).

42. W. Ahmed, M. A. de Graaf, A. Broersen, P. H. Kitslaar, E. Oost, J. Dijkstra, J. J. Bax, J. H. Reiber, and A. J. Scholte. "Automatic detection and quantification of the agatston coronary artery calcium score on contrast computed tomography angiography", *International Journal of Cardiovascular Imaging*, vol. 31 (2015), pp. 151–161 (cited on p. 15).

43. S. Mittal, Y. Zheng, B. Georgescu, F. Vega-Higuera, S. K. Zhou, P. Meer, and D. Comaniciu. "Fast automatic detection of calcified coronary lesions in 3d cardiac CT images", *International Workshop on Machine Learning in Medical Imaging*, vol. 6357 of *LNCS* (Springer, 2010), pp. 1–9 (cited on p. 15).

44. J. M. Wolterink, T. Leiner, M. A. Viergever, and I. Išgum. "Automatic coronary calcium scoring in cardiac CT angiography using convolutional neural networks", *International Conference on Medical Image Computing and Computer-Assisted Intervention*, vol. 9349 of *LNCS* (Springer, 2015), pp. 589–596 (cited on p. 15).

45. J. M. Wolterink, T. Leiner, B. D. de Vos, R. W. van Hamersvelt, M. A. Viergever, and I. Išgum. "Automatic coronary artery calcium scoring in cardiac CT angiography using paired convolutional neural networks", *Medical Image Analysis*, vol. 34 (2016), pp. 123–136 (cited on pp. 15–17, 20).

46. I. Išgum, M. Prokop, M. Niemeijer, M. A. Viergever, and B. van Ginneken. "Automatic coronary calcium scoring in low-dose chest computed tomography", *IEEE Transactions on Medical Imaging*, vol. 31 (2012), pp. 2322–34 (cited on pp. 15, 16, 31).

47. Y. Xie, M. D. Cham, C. Henschke, D. Yankelevitz, and A. P. Reeves. "Automated coronary artery calcification detection on low-dose chest CT images", *Medical Imaging*, vol. 9035 of *Proceedings of SPIE* (2014), 90350F (cited on pp. 15, 16, 31).

48. N. Lessmann, I. Išgum, A. A. A. Setio, B. D. de Vos, F. Ciompi, P. A. de Jong, M. Oudkerk, W. P. Th. M. Mali, M. A. Viergever, and B. van Ginneken. "Deep convolutional neural networks for automatic coronary calcium scoring in a screening study with low-dose chest CT", *Medical Imaging*, vol. 9785 of *Proceedings of SPIE* (2016), 978511 (cited on pp. 15, 16, 21, 31).

49. G. González, G. R. Washko, and R. S. J. Estépar. "Automated Agatston score computation in a large dataset of non ECG-gated chest computed tomography", *IEEE 13th International Symposium on Biomedical Imaging (ISBI)*, 2016, pp. 53–57 (cited on pp. 15, 16, 31).

50. G. H. Tison, M. Guo, M. J. Blaha, R. L. McClelland, M. A. Allison, M. Szklo, N. D. Wong, R. S. Blumenthal, M. J. Budoff, and K. Nasir. "Multisite extracoronary calcification indicates increased risk of coronary heart disease and all-cause mortality: The Multi-Ethnic Study of Atherosclerosis", *Journal of Cardiovascular Computer Tomography*, vol. 9 (2015), pp. 406–414 (cited on p. 16).

51. M. J. Willemink, R. A. P. Takx, I. Išgum, H. J. de Koning, M. Oudkerk, W. P. Th. M. Mali, R. P. J. Budde, T. Leiner, R. Vliegenthart, and P. A. de Jong. "Prognostic value of heart valve calcifications for cardiovascular events in a lung cancer screening population", *International Journal of Cardiovascular Imaging*, vol. 31 (2015), pp. 1243–49 (cited on p. 16).

52. S. Kurugol, R. S. J. Estépar, J. Ross, and G. R. Washko. "Aorta segmentation with a 3D level set approach and quantification of aortic calcifications in non-contrast chest CT", *34th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2012, pp. 2343–46 (cited on pp. 16, 31).

53. Y. Xie, Y. M. Htwe, J. Padgett, C. Henschke, D. Yankelevitz, and A. P. Reeves. "Automated aortic calcification detection in low-dose chest CT images", *Medical Imaging*, vol. 9035 of *Proceedings of SPIE* (2014), 90350P (cited on pp. 16, 31).

54. I. Išgum, A. Rutten, M. Prokop, M. Staring, S. Klein, J. P. W. Pluim, M. A. Viergever, and B. van Ginneken. "Automated aortic calcium scoring on low-dose chest computed tomography", *Medical Physics*, vol. 37 (2010), pp. 714–723 (cited on pp. 16, 31).

55. J. Long, E. Shelhamer, and T. Darrell. "Fully convolutional networks for semantic segmentation", *Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3431–40 (cited on p. 19).

56. J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller. "Striving for simplicity: the all convolutional net", *Proceedings of ICLR*, 2015 (cited on p. 19).

57. F. Yu and V. Koltun. "Multi-scale context aggregation by dilated convolutions", *Proceedings of ICLR*, 2016 (cited on pp. 20, 30).

58. A. Prasoon, K. Petersen, C. Igel, F. Lauze, E. Dam, and M. Nielsen. "Deep feature learning for knee cartilage segmentation using a triplanar convolutional neural network", *International Conference on Medical Image Computing and Computer-Assisted Intervention*, vol. 8150 of *LNCS* (Springer, 2013), pp. 246–253 (cited on p. 20).

59. C.-Y. Lee, S. Xie, P. W. Gallagher, Z. Zhang, and Z. Tu. "Deeply-supervised nets", *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2015 (cited on p. 20).

60. B. Ohnesorge, T. Flohr, R. Fischbach, A. Kopp, A. Knez, S. Schröder, U. Schöpf, A. Crispin, E. Klotz, M. Reiser, et al. "Reproducibility of coronary calcium quantification in repeat examinations with retrospectively ECG-gated multisection spiral CT", *European Radiology*, vol. 12 (2002), pp. 1532–40 (cited on pp. 22, 40).

61. D. Clevert, T. Unterthiner, and S. Hochreiter. "Fast and accurate deep network learning by exponential linear units (ELUs)", *Proceedings of ICLR*, 2016 (cited on p. 26).

62. D. P. Kingma and J. Ba. "Adam: a method for stochastic optimization", *Proceedings of ICLR*, 2015 (cited on pp. 26, 66, 95).

63. N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. "Dropout: a simple way to prevent neural networks from overfitting", *Journal of Machine Learning Research*, vol. 15 (2014), pp. 1929–58 (cited on p. 26).

64. S. Ioffe and C. Szegedy. "Batch normalization: accelerating deep network training by reducing internal covariate shift", *Proceedings of the 32nd International Conference on Machine Learning*, 2015 (cited on p. 26).

65. Theano Development Team. "Theano: a Python framework for fast computation of mathematical expressions", arXiv:1605.02688, 2016 (cited on p. 26).

66. S. Dieleman, J. Schlüter, C. Raffel, E. Olson, S. K. Sønderby, D. Nouri, et al. "Lasagne: first release", Zenodo:27878, 2015. DOI 10.5281/zenodo.27878 (cited on p. 26).

67. R. W. van Hamersvelt, M. J. Willemink, R. A. P. Takx, A. L. M. Eikendal, R. P. J. Budde, T. Leiner, C. P. Mol, I. Išgum, and P. A. de Jong. "Cardiac valve calcifications on low-dose unenhanced ungated chest computed tomography: inter-observer and inter-examination reliability, agreement and variability", *European Radiology*, vol. 24 (2014), pp. 1557–64 (cited on p. 33).

68. O. Ronneberger, P. Fischer, and T. Brox. "U-net: convolutional networks for biomedical image segmentation", *International Conference on Medical Image Computing and Computer-Assisted Intervention*, vol. 9351 of *LNCS* (Springer, 2015), pp. 234–241 (cited on pp. 34, 61, 67, 86).

69. K. He, X. Zhang, S. Ren, and J. Sun. "Deep residual learning for image recognition", *Conference on Computer Vision and Pattern Recognition*, 2016 (cited on p. 34).

70. G. Huang, Z. Liu, K. Q. Weinberger, and L. van der Maaten. "Densely connected convolutional networks", *Conference on Computer Vision and Pattern Recognition*, 2017 (cited on p. 34).

71. F. Milletari, S.-A. Ahmadi, C. Kroll, A. Plate, V. Rozanski, J. Maiostre, J. Levin, O. Dietrich, B. Ertl-Wagner, K. Bötzel, et al. "Hough-CNN: deep learning for segmentation of deep brain regions in MRI and ultrasound", *Computer Vision and Image Understanding*, vol. 164 (2017), pp. 92–102 (cited on p. 34).

72. M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu. "Spatial transformer networks", *Advances in Neural Information Processing Systems*, vol. 28 (2015) (cited on pp. 34, 100).

73. M. Naghavi, A. A. Abajobir, C. Abbafati, K. M. Abbas, F. Abd-Allah, S. F. Abera, V. Aboyans, O. Adetokunboh, et al. "Global, regional, and national age-sex specific mortality for 264 causes of death, 1980–2016: a systematic analysis for the Global Burden of Disease Study 2016", *The Lancet*, vol. 390 (2017), pp. 1151–1210 (cited on p. 39).

74. D. J. Lerner and W. B. Kannel. "Patterns of coronary heart disease morbidity and mortality in the sexes: a 26-year follow-up of the framingham population", *American Heart Journal*, vol. 111 (1986), pp. 383–390 (cited on p. 39).

75. O. M. Mets, P. A. de Jong, and M. Prokop. "Computed tomographic screening for lung cancer: an opportunity to evaluate other diseases", *JAMA*, vol. 308 (2012), pp. 1433–34 (cited on p. 39).

76. K. Nasir, A. Roguin, A. Sarwar, J. A. Rumberger, and R. S. Blumenthal. "Gender differences in coronary arteries and thoracic aorta calcification", *Arteriosclerosis, Thrombosis, and Vascular Biology*, vol. 27 (2007), pp. 1220–22 (cited on pp. 39, 50).

77. D. M. Hermann, N. Lehmann, J. Gronewold, M. Bauer, A. A. Mahabadi, C. Weimar, K. Berger, S. Moebus, et al. "Thoracic aortic calcification is associated with incident stroke in the general population in addition to established risk factors", *European Heart Journal - Cardiovascular Imaging*, vol. 16 (2014), pp. 684–690 (cited on p. 39).

78. D. R. Aberle, A. M. Adams, C. D. Berg, J. D. Clapp, K. L. Clingan, I. F. Gareen, D. A. Lynch, P. M. Marcus, and P. F. Pinsky. "Baseline characteristics of participants in the randomized National Lung Screening Trial", *JNCI: Journal of the National Cancer Institute*, vol. 102 (2010), pp. 1771–79 (cited on p. 40).

79. N. Lessmann, B. van Ginneken, M. Zreik, P. A. de Jong, B. D. de Vos, M. A. Viergever, and I. Išgum. "Automatic calcium scoring in low-dose chest CT using deep neural networks with dilated convolutions", *IEEE Transactions on Medical Imaging*, vol. 37 (2018), pp. 615–625 (cited on pp. 40, 51, 85).

80. J. N. K. Rao and A. J. Scott. "On simple adjustments to chi-square tests with sample survey data", *The Annals of Statistics*, vol. 15 (1987), pp. 385–397 (cited on p. 42).

81. T. Lumley and A. J. Scott. "Two-sample rank tests under complex sampling", *Biometrika*, vol. 100 (2013), pp. 831–842 (cited on p. 42).

82. T. M. Therneau and P. M. Grambsch. "Modeling survival data: extending the cox model", New York: Springer, 2000 (cited on p. 42).

83. A. R. T. Donders, G. J. M. G. van der Heijden, T. Stijnen, and K. G. M. Moons. "Review: a gentle introduction to imputation of missing values", *Journal of Clinical Epidemiology*, vol. 59 (2006), pp. 1087–91 (cited on p. 42).

84. J. A. Hoff, E. V. Chomka, A. J. Krainik, M. Daviglus, S. Rich, and G. T. Kondos. "Age and gender distributions of coronary artery calcium detected by electron beam tomography in 35,246 adults", *The American Journal of Cardiology*, vol. 87 (2001), pp. 1335–39 (cited on p. 48).

85. K. Nasir, P. Raggi, J. A. Rumberger, J. B. Braunstein, W. S. Post, M. J. Budoff, and R. S. Blumenthal. "Coronary artery calcium volume scores on electron beam tomography in 12,936 asymptomatic adults", *The American Journal of Cardiology*, vol. 93 (2004), pp. 1146–49 (cited on p. 48).

86. R. L. McClelland, H. Chung, R. Detrano, W. Post, and R. A. Kronmal. "Distribution of coronary artery calcium by race, gender, and age: results from the Multi-Ethnic Study of Atherosclerosis (MESA)", *Circulation*, vol. 113 (2006), pp. 30–37 (cited on p. 48).

87. U. Hoffmann, J. M. Massaro, C. S. Fox, E. Manders, and C. J. O'Donnell. "Defining normal distributions of coronary artery calcium in women and men (from the Framingham Heart Study)", *The American Journal of Cardiology*, vol. 102 (2008), pp. 1136–41.e1 (cited on p. 48).

88. P. M. Jairam, P. A. de Jong, W. P. Mali, M. J. Gondrie, P. C. Jacobs, Y. van der Graaf, et al. "Age and sex based reference values for incidental coronary artery and thoracic aorta calcifications on routine clinical chest CT: a powerful tool to appreciate available imaging findings", *Atherosclerosis*, vol. 235 (2014), pp. 546–553 (cited on pp. 48, 50).

89. Y. Ohmoto-Sekine, R. Yanagibori, K. Amakawa, M. Ishihara, H. Tsuji, K. Ogawa, R. Ishimura, S. Ishiwata, et al. "Prevalence and distribution of coronary calcium in asymptomatic Japanese subjects in lung cancer screening computed tomography", *Journal of Cardiology*, vol. 67 (2016), pp. 449–454 (cited on p. 48).

90. J. Takasu, R. Katz, K. Nasir, J. J. Carr, N. Wong, R. Detrano, and M. J. Budoff. "Relationships of thoracic aortic wall calcification to cardiovascular risk factors: the Multi-Ethnic Study of Atherosclerosis (MESA)", *American Heart Journal*, vol. 155 (2008), pp. 765–771 (cited on p. 50).

91. N. D. Wong, M. Sciammarella, Y. Arad, R. Miranda-Peats, D. Polk, R. Hachamovich, J. Friedman, S. Hayes, A. Daniell, and D. S. Berman. "Relation of thoracic aortic and aortic valve calcium to coronary artery calcium and risk assessment", *The American Journal of Cardiology*, vol. 92 (2003), pp. 951–955 (cited on p. 50).

92. H. Kälsch, N. Lehmann, S. Möhlenkamp, C. Hammer, A. A. Mahabadi, S. Moebus, A. Schmermund, A. Stang, et al. "Prevalence of thoracic aortic calcification and its relationship to cardiovascular risk factors and coronary calcification in an unselected population-based cohort: the Heinz Nixdorf Recall Study", *The International Journal of Cardiovascular Imaging*, vol. 29 (2012), pp. 207–216 (cited on p. 50).

93. I. Kardys, R. Vliegenthart, M. Oudkerk, A. Hofman, and J. C. M. Witteman. "The female advantage in cardiovascular disease: do vascular beds contribute equally?", *American Journal of Epidemiology*, vol. 166 (2007), pp. 403–412 (cited on p. 50).

94. R. Nakanishi, D. Li, M. J. Blaha, S. P. Whelton, S. Darabian, F. R. Flores, C. Dailing, R. S. Blumenthal, et al. "All-cause mortality by age and gender based on coronary artery calcium scores", *European Heart Journal - Cardiovascular Imaging*, vol. 17 (2015), pp. 1305–14 (cited on p. 50).

95. A. Bellasi, C. Lacey, A. J. Taylor, P. Raggi, P. W. F. Wilson, M. J. Budoff, V. Vaccarino, and L. J. Shaw. "Comparison of prognostic usefulness of coronary artery calcium in men versus women (results from a meta- and pooled analysis estimating all-cause mortality and coronary heart disease death or myocardial infarction)", *The American Journal of Cardiology*, vol. 100 (2007), pp. 409–414 (cited on p. 50).

96. P. Raggi, L. J. Shaw, D. S. Berman, and T. Q. Callister. "Gender-based differences in the prognostic value of coronary calcification", *Journal of Women's Health*, vol. 13 (2004), pp. 273–283 (cited on p. 50).

97. M. J. LaMonte, S. J. FitzGerald, T. S. Church, C. E. Barlow, N. B. Radford, B. D. Levine, J. J. Pippin, L. W. Gibbons, S. N. Blair, and M. Z. Nichaman. "Coronary artery calcium score and coronary heart disease events in a large cohort of asymptomatic men and women", *American Journal of Epidemiology*, vol. 162 (2005), pp. 421–429 (cited on p. 50).

98. J. W. McEvoy, M. J. Blaha, J. J. Rivera, M. J. Budoff, A. N. Khan, L. J. Shaw, D. S. Berman, P. Raggi, et al. "Mortality rates in smokers and nonsmokers in the presence or absence of coronary artery calcification", *JACC: Cardiovascular Imaging*, vol. 5 (2012), pp. 1037–45 (cited on p. 50).

99. M. J. Budoff, K. Nasir, R. Katz, J. Takasu, J. J. Carr, N. D. Wong, M. Allison, J. A. C. Lima, et al. "Thoracic aortic calcification and coronary heart disease events: the Multi-Ethnic Study of Atherosclerosis (MESA)", *Atherosclerosis*, vol. 215 (2011), pp. 196–202 (cited on p. 50).

100. M. J. Budoff, K. Nasir, G. L. Kinney, J. E. Hokanson, R. G. Barr, R. Steiner, H. Nath, C. Lopez-Garcia, J. Black-Shinn, and R. Casaburi. "Coronary artery and thoracic calcium on noncontrast thoracic CT scans: comparison of ungated and gated examinations in patients from the COPD Gene cohort", *Journal of Cardiovascular Computed Tomography*, vol. 5 (2011), pp. 113–118 (cited on p. 51).

101. H. S. Hecht, P. Cronin, M. J. Blaha, M. J. Budoff, E. A. Kazerooni, J. Narula, D. Yankelevitz, and S. Abbara. "2016 SCCT/STR guidelines for coronary artery calcium scoring of noncontrast noncardiac chest CT scans: a report of the Society of Cardiovascular Computed Tomography and Society of Thoracic Radiology", *Journal of Cardiovascular Computed Tomography*, vol. 11 (2017), pp. 74–84 (cited on pp. 51, 111).

102. A. Bernheim, W. F. Auffermann, and A. E. Stillman. "The dubious value of coronary calcium scoring on lung cancer screening CT", *Journal of the American College of Radiology*, vol. 14 (2017), pp. 343–344 (cited on p. 51).

103. P. M. Ridker, B. M. Everett, T. Thuren, J. G. MacFadyen, W. H. Chang, C. Ballantyne, F. Fonseca, J. Nicolau, et al. "Antiinflammatory therapy with canakinumab for atherosclerotic disease", *New England Journal of Medicine*, vol. 377 (2017), pp. 1119–31 (cited on p. 51).

104. J. Yao, J. E. Burns, H. Munoz, and R. M. Summers. "Detection of vertebral body fractures based on cortical shell unwrapping", *International Conference on Medical Image Computing and Computer-Assisted Intervention*, vol. 7512 of *LNCS* (Springer, 2012), pp. 509–516 (cited on p. 57).

105. D. Forsberg, C. Lundström, M. Andersson, L. Vavruch, H. Tropp, and H. Knutsson. "Fully automatic measurements of axial vertebral rotation for assessment of spinal deformity in idiopathic scoliosis", *Physics in Medicine and Biology*, vol. 58 (2013), pp. 1775–87 (cited on p. 57).

106. D. Knez, B. Likar, F. Pernuš, and T. Vrtovec. "Computer-assisted screw size and insertion trajectory planning for pedicle screw placement surgery", *IEEE Transactions on Medical Imaging*, vol. 35 (2016), pp. 1420–30 (cited on p. 57).

107. M. Grigoryan, A. Guermazi, F. W. Roemer, P. D. Delmas, and H. K. Genant. "Recognizing and reporting osteoporotic vertebral fractures", *European Spine Journal*, vol. 12 (2003), pp. S104–S112 (cited on pp. 57, 99, 111).

108. T. Klinder, J. Ostermann, M. Ehm, A. Franz, R. Kneser, and C. Lorenz. "Automated model-based vertebra detection, identification, and segmentation in CT images", *Medical Image Analysis*, vol. 13 (2009), pp. 471–482 (cited on pp. 58–60, 65).

109. B. M. Kelm, M. Wels, S. K. Zhou, S. Seifert, M. Suehling, Y. Zheng, and D. Comaniciu. "Spine detection in CT and MR using iterated marginal space learning", *Medical Image Analysis*, vol. 17 (2013), pp. 1283–92 (cited on pp. 58–60).

110. C. Chu, D. L. Belavỳ, G. Armbrecht, M. Bansmann, D. Felsenberg, and G. Zheng. "Fully automatic localization and segmentation of 3D vertebral bodies from CT/MR images via a learning-based method", *PLOS ONE*, vol. 10 (2015), e0143327 (cited on pp. 58, 59, 69, 71, 88).

111. A. Suzani, A. Rasoulian, A. Seitel, S. Fels, R. N. Rohling, and P. Abolmaesumi. "Deep learning for automatic localization, identification, and segmentation of vertebral bodies in volumetric MR image", *Medical Imaging*, vol. 9415 of *Proceedings of SPIE* (2015), 941514 (cited on pp. 58, 59).

112. A. Sekuboyina, A. Valentinitsch, J. S. Kirschke, and B. H. Menze. "A localisation-segmentation approach for multi-label annotation of lumbar vertebrae using deep nets", arXiv:1703.04347, 2017 (cited on pp. 58–60, 68, 71, 74).

113. I. Castro-Mateos, J. M. Pozo, M. Pereanez, K. Lekadir, A. Lazary, and A. F. Frangi. "Statistical interspace models (SIMs): application to robust 3D spine segmentation", *IEEE Transactions on Medical Imaging*, vol. 34 (2015), pp. 1663–75 (cited on p. 58).

114. B. Ibragimov, B. Likar, F. Pernuš, and T. Vrtovec. "Shape representation for efficient landmark-based segmentation in 3-D", *IEEE Transactions on Medical Imaging*, vol. 33 (2014), pp. 861–874 (cited on pp. 58, 69, 71).

115. B. Ibragimov, R. Korez, B. Likar, F. Pernuš, and T. Vrtovec. "Interpolation-based detection of lumbar vertebrae in CT spine images", *Recent Advances in Computational Methods and Clinical Applications for Spine Imaging*, vol. 20 of *Lecture Notes in Computational Vision and Biomechanics* (Springer, 2015), pp. 73–84 (cited on p. 58).

116. S. Kadoury, H. Labelle, and N. Paragios. "Automatic inference of articulated spine models in CT images using high-order markov random fields", *Medical Image Analysis*, vol. 15 (2011), pp. 426–437 (cited on p. 58).

117. S. Kadoury, H. Labelle, and N. Paragios. "Spine segmentation in medical images using manifold embeddings and higher-order MRFs", *IEEE Transactions on Medical Imaging*, vol. 32 (2013), pp. 1227–38 (cited on p. 58).

118. R. Korez, B. Ibragimov, B. Likar, F. Pernuš, and T. Vrtovec. "A framework for automated spine and vertebrae interpolation-based detection and model-based segmentation", *IEEE Transactions on Medical Imaging*, vol. 34 (2015), pp. 1649–62 (cited on pp. 58, 69, 71).

119. R. Korez, B. Likar, F. Pernuš, and T. Vrtovec. "Model-based segmentation of vertebral bodies from MR images with 3D CNNs", *International Conference on Medical Image Computing and Computer-Assisted Intervention*, vol. 9901 of *LNCS* (Springer, 2016), pp. 433–441 (cited on pp. 58, 59, 71, 74, 80).

120. A. Mastmeyer, K. Engelke, C. Fuchs, and W. A. Kalender. "A hierarchical 3D segmentation method and the definition of vertebral body coordinate systems for QCT of the lumbar spine", *Medical Image Analysis*, vol. 10 (2006), pp. 560–577 (cited on p. 58).

121. H. Mirzaalian, M. Wels, T. Heimann, B. M. Kelm, and M. Suehling. "Fast and robust 3D vertebra segmentation using statistical shape models", *35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2013, pp. 3379–82 (cited on p. 58).

122. M. Pereanez, K. Lekadir, I. Castro-Mateos, J. M. Pozo, A. Lazary, and A. F. Frangi. "Accurate segmentation of vertebral bodies and processes using statistical shape decomposition and conditional models", *IEEE Transactions on Medical Imaging*, vol. 34 (2015), pp. 1627–39 (cited on p. 58).

123. A. Rasoulian, R. Rohling, and P. Abolmaesumi. "Lumbar spine segmentation using a statistical multi-vertebrae anatomical shape+pose model", *IEEE Transactions on Medical Imaging*, vol. 32 (2013), pp. 1890–1900 (cited on p. 58).

124. D. Štern, B. Likar, F. Pernuš, and T. Vrtovec. "Parametric modelling and segmentation of vertebral bodies in 3D CT and MR spine images", *Physics in Medicine and Biology*, vol. 56 (2011), pp. 7505–22 (cited on p. 58).

125. D. Yang, T. Xiong, D. Xu, Q. Huang, D. Liu, S. K. Zhou, Z. Xu, J. Park, M. Chen, T. D. Tran, et al. "Automatic vertebra labeling in large-scale 3D CT using deep image-to-image network with message passing and sparsity regularization", *IPMI*, vol. 10265 of *LNCS* (Springer, 2017), pp. 633–644 (cited on pp. 58, 59).

126. M. E. Leventon, W. E. L. Grimson, and O. Faugeras. "Statistical shape influence in geodesic active contours", *5th IEEE EMBS International Summer School on Biomedical Imaging*, 2002 (cited on p. 58).

127. Y. Wang, J. Yao, H. R. Roth, J. E. Burns, and R. M. Summers. "Multi-atlas segmentation with joint label fusion of osteoporotic vertebral compression fractures on CT", *International Workshop on Computational Methods and Clinical Applications for Spine Imaging*, vol. 9402 of *LNCS* (Springer, 2015), pp. 74–84 (cited on p. 58).

128. P. H. Lim, U. Bagci, and L. Bai. "A robust segmentation framework for spine trauma diagnosis", *Computational Methods and Clinical Applications for Spine Imaging*, vol. 17 of *Lecture Notes in Computational Vision and Biomechanics* (Springer, 2014), pp. 25–33 (cited on p. 58).

129. K. Hammernik, T. Ebner, D. Stern, M. Urschler, and T. Pock. "Vertebrae segmentation in 3D CT images based on a variational framework", *Recent Advances in Computational Methods and Clinical Applications for Spine Imaging*, vol. 20 of *Lecture Notes in Computational Vision and Biomechanics* (Springer, 2015), pp. 227–233 (cited on p. 58).

130. J. S. Athertya and G. S. Kumar. "Automatic segmentation of vertebral contours from CT images using fuzzy corners", *Computers in Biology and Medicine*, vol. 72 (2016), pp. 75–89 (cited on p. 58).

131. D. Zukić, A. Vlasák, J. Egger, D. Hořínek, C. Nimsky, and A. Kolb. "Robust detection and segmentation for diagnosis of vertebral diseases using routine MR images", *Computer Graphics Forum*, vol. 33 (2014), pp. 190–204 (cited on p. 58).

132. A. Sekuboyina, J. Kukačka, J. S. Kirschke, B. H. Menze, and A. Valentinitsch. "Attention-driven deep learning for pathological spine segmentation", *Computational Methods and Clinical Applications in Musculoskeletal Imaging*, vol. 10734 of *LNCS* (Springer, 2018), pp. 108–119 (cited on p. 59).

133. R. Janssens, G. Zeng, and G. Zheng. "Fully automatic segmentation of lumbar vertebrae from CT images using cascaded 3D fully convolutional networks", *IEEE 15th International Symposium on Biomedical Imaging (ISBI)*, 2018, pp. 893–897 (cited on pp. 59, 60, 68, 71, 74).

134. N. Lessmann, B. van Ginneken, and I. Išgum. "Iterative convolutional neural networks for automatic vertebra identification and segmentation in CT images", *Medical Imaging*, vol. 10574 of *Proceedings of SPIE* (2018), 1057408 (cited on pp. 59, 71).

135. B. Glocker, J. Feulner, A. Criminisi, D. Haynor, and E. Konukoglu. "Automatic localization and identification of vertebrae in arbitrary field-of-view CT scans", *International Conference on Medical Image Computing and Computer-Assisted Intervention*, vol. 7512 of *LNCS* (Springer, 2012), pp. 590–598 (cited on p. 59).

136. Y. Cai, S. Osman, M. Sharma, M. Landis, and S. Li. "Multi-modality vertebra recognition in arbitrary views using 3D deformable hierarchical model", *IEEE Transactions on Medical Imaging*, vol. 34 (2015), pp. 1676–93 (cited on p. 59).

137. B. Glocker, D. Zikic, E. Konukoglu, D. R. Haynor, and A. Criminisi. "Vertebrae localization in pathological spine CT via dense classification from sparse annotations", *International Conference on Medical Image Computing and Computer-Assisted Intervention*, vol. 8150 of *LNCS* (Springer, 2013), pp. 262–270 (cited on pp. 59, 60).

138. D. Major, J. Hladůvka, F. Schulze, and K. Bühler. "Automated landmarking and labeling of fully and partially scanned spinal columns in CT images", *Medical Image Analysis*, vol. 17 (2013), pp. 1151–63 (cited on p. 59).

139. P. A. Bromiley, E. P. Kariki, J. E. Adams, and T. F. Cootes. "Fully automatic localisation of vertebrae in CT images using random forest regression voting", *International Workshop on Computational Methods and Clinical Applications for Spine Imaging*, vol. 10182 of *LNCS* (Springer, 2016), pp. 51–63 (cited on p. 59).

140. Y. Cai, M. Landis, D. T. Laidley, A. Kornecki, A. Lum, and S. Li. "Multi-modal vertebrae recognition using transformed deep convolution network", *Computerized Medical Imaging and Graphics*, vol. 51 (2016), pp. 11–19 (cited on p. 59).

141. H. Chen, C. Shen, J. Qin, D. Ni, L. Shi, J. C. Y. Cheng, and P.-A. Heng. "Automatic localization and identification of vertebrae in spine CT via a joint learning model with deep neural networks", *International Conference on Medical Image Computing and Computer-Assisted Intervention*, vol. 9349 of *LNCS* (Springer, 2015), pp. 515–522 (cited on pp. 59, 60).

142. D. Forsberg, E. Sjöblom, and J. L. Sunshine. "Detection and labeling of vertebrae in MR images using deep learning with clinical annotations as training data", *Journal of Digital Imaging*, vol. 30 (2017), pp. 406–412 (cited on p. 59).

143. H. Liao, A. Mesfin, and J. Luo. "Joint vertebrae identification and localization in spinal CT images by combining short- and long-range contextual information", *IEEE Transactions on Medical Imaging*, vol. 37 (2018), pp. 1266–75 (cited on p. 59).

144. D. Yang, T. Xiong, D. Xu, S. K. Zhou, Z. Xu, M. Chen, J. Park, S. Grbic, et al. "Deep image-to-image recurrent network with shape basis learning for automatic vertebra labeling in large-scale 3D CT volumes", *International Conference on Medical Image Computing and Computer-Assisted Intervention*, vol. 10435 of *LNCS* (Springer, 2017), pp. 498–506 (cited on p. 59).

145. K. He, G. Gkioxari, P. Dollár, and R. B. Girshick. "Mask R-CNN", *IEEE International Conference on Computer Vision (ICCV)*, 2017 (cited on p. 60).

146. R. Stewart, M. Andriluka, and A. Y. Ng. "End-to-end people detection in crowded scenes", *Conference on Computer Vision and Pattern Recognition*, 2016 (cited on p. 60).

147. B. Romera-Paredes and P. H. S. Torr. "Recurrent instance segmentation", *European Conference on Computer Vision*, vol. 9910 of *LNCS* (Springer, 2016), pp. 312–329 (cited on p. 60).

148. K. Li, B. Hariharan, and J. Malik. "Iterative instance segmentation", *Conference on Computer Vision and Pattern Recognition*, 2016 (cited on p. 60).

149. M. Ren and R. S. Zemel. "End-to-end instance segmentation with recurrent attention", *Conference on Computer Vision and Pattern Recognition*, 2017 (cited on p. 60).

150. J. Uhrig, M. Cordts, U. Franke, and T. Brox. "Pixel-level encoding and depth layering for instance-level semantic labeling", *German Conference on Pattern Recognition (GCPR)*, vol. 9796 of *LNCS* (Springer, 2016), pp. 14–25 (cited on p. 60).

151. B. De Brabandere, D. Neven, and L. Van Gool. "Semantic instance segmentation for autonomous driving", *Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on*, 2017, pp. 478–480 (cited on p. 60).

152. X. Liang, L. Lin, Y. Wei, X. Shen, J. Yang, and S. Yan. "Proposal-free network for instance-level semantic object segmentation", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40 (2018), pp. 2978–91 (cited on p. 60).

153. D. Novotny, S. Albanie, D. Larlus, and A. Vedaldi. "Semi-convolutional operators for instance segmentation", *European Conference on Computer Vision*, *LNCS* Springer, 2018, pp. 89–105 (cited on p. 60).

154. Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger. "3D U-Net: learning dense volumetric segmentation from sparse annotation", *International Conference on Medical Image Computing and Computer-Assisted Intervention*, vol. 9901 of *LNCS* (Springer, 2016), pp. 234–241 (cited on pp. 61, 79).

155. F. Milletari, N. Navab, and S.-A. Ahmadi. "V-net: fully convolutional neural networks for volumetric medical image segmentation", *International Conference on 3D Vision*, 2016, pp. 565–571 (cited on p. 66).

156. J. Yao, J. E. Burns, D. Forsberg, A. Seitel, A. Rasoulian, P. Abolmaesumi, K. Hammernik, M. Urschler, B. Ibragimov, R. Korez, et al. "A multi-center milestone study of clinical vertebral CT segmentation", *Computerized Medical Imaging and Graphics*, vol. 49 (2016), pp. 16–28 (cited on pp. 68, 78).

157. B. Ibragimov, R. Korez, B. Likar, F. Pernus, L. Xing, and T. Vrtovec. "Segmentation of pathological structures by landmark-assisted deformable models", *IEEE Transactions on Medical Imaging*, vol. 36 (2017), pp. 1457–69 (cited on p. 68).

158. W. A. Barrett and E. N. Mortensen. "Interactive live-wire boundary extraction", *Medical Image Analysis*, vol. 1 (1997), pp. 331–341 (cited on p. 69).

159. J. Cohen. "Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit." *Psychological Bulletin*, vol. 70 (1968), pp. 213–220 (cited on p. 70).

160. G. J. S. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. W. M. van der Laak, B. van Ginneken, and C. I. Sánchez. "A survey on deep learning in medical image analysis", *Medical Image Analysis*, vol. 42 (2017), pp. 60–88 (cited on pp. 78, 85).

161. P. Moeskops, M. A. Viergever, A. M. Mendrik, L. S. de Vries, M. J. Benders, and I. Išgum. "Automatic segmentation of MR brain images with a convolutional neural network", *IEEE Transactions on Medical Imaging*, vol. 35 (2016), pp. 1252–1261 (cited on p. 79).

162. K. Kamnitsas, C. Ledig, V. F. Newcombe, J. P. Simpson, A. D. Kane, D. K. Menon, D. Rueckert, and B. Glocker. "Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation", *Medical Image Analysis*, vol. 36 (2017), pp. 61–78 (cited on p. 79).

163. S. Pereira, A. Pinto, V. Alves, and C. A. Silva. "Brain tumor segmentation using convolutional neural networks in MRI images", *IEEE Transactions on Medical Imaging*, vol. 35 (2016), pp. 1240–51 (cited on p. 85).

164. V. Pekar, T. R. McNutt, and M. R. Kaus. "Automated model-based organ delineation for radiotherapy planning in prostatic region", *International Journal of Radiation Oncology*Biology*Physics*, vol. 60 (2004), pp. 973–980 (cited on p. 85).

165. M. F. McNitt-Gray, E. M. Hart, N. Wyckoff, J. W. Sayre, J. G. Goldin, and D. R. Aberle. "A pattern classification approach to characterizing solitary pulmonary nodules imaged on high resolution CT: preliminary results", *Medical Physics*, vol. 26 (1999), pp. 880–888 (cited on p. 85).

166. P. Moeskops, M. J. N. L. Benders, K. J. Kersbergen, F. Groenendaal, L. S. de Vries, M. A. Viergever, and I. Išgum. "Development of cortical morphology evaluated with longitudinal MR brain images of preterm infants", *PLOS ONE*, vol. 10 (2015), e0131552 (cited on p. 85).

167. O. Bernard, A. Lalande, C. Zotti, F. Cervenansky, X. Yang, P.-A. Heng, I. Cetin, K. Lekadir, et al. "Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: is the problem solved?", *IEEE Transactions on Medical Imaging*, vol. 37 (2018), pp. 2514–25 (cited on p. 85).

168. M. Zreik, N. Lessmann, R. W. van Hamersvelt, J. M. Wolterink, M. Voskuil, M. A. Viergever, T. Leiner, and I. Išgum. "Deep learning analysis of the myocardium in coronary CT angiography for identification of patients with functionally significant coronary artery stenosis", *Medical Image Analysis*, vol. 44 (2018), pp. 72–85 (cited on p. 85).

169. O. Ecabert, J. Peters, H. Schramm, C. Lorenz, J. von Berg, M. J. Walker, M. Vembar, M. E. Olszewski, et al. "Automatic model-based segmentation of the heart in CT images", *IEEE Transactions on Medical Imaging*, vol. 27 (2008), pp. 1189–1201 (cited on p. 85).

170. T. Heimann and H.-P. Meinzer. "Statistical shape models for 3D medical image segmentation: a review", *Medical Image Analysis*, vol. 13 (2009), pp. 543–563 (cited on p. 85).

171. I. Išgum, M. Staring, A. Rutten, M. Prokop, M. A. Viergever, and B. van Ginneken. "Multi-atlas-based segmentation with local decision fusion–application to cardiac and aortic segmentation in CT scans", *IEEE Transactions on Medical Imaging*, vol. 28 (2009), pp. 1000–10 (cited on p. 85).

172. C. Ledig, R. Wolz, P. Aljabar, J. Lotjonen, R. A. Heckemann, A. Hammers, and D. Rueckert. "Multi-class brain segmentation using atlas propagation and EM-based refinement", *IEEE 9th International Symposium on Biomedical Imaging (ISBI)*, 2012, pp. 896–899 (cited on p. 85).

173. X. Zhou, T. Ito, R. Takayama, S. Wang, T. Hara, and H. Fujita. "Three-dimensional CT image segmentation by combining 2D fully convolutional network with 3D majority voting", *Deep Learning and Data Labeling for Medical Applications*, vol. 10008 of *LNCS* (Springer, 2016), pp. 111–120 (cited on p. 85).

174. P. Moeskops, J. M. Wolterink, B. H. M. van der Velden, K. G. A. Gilhuijs, T. Leiner, M. A. Viergever, and I. Išgum. "Deep learning for multi-task medical image segmentation in multiple modalities", *International Conference on Medical Image Computing and Computer-Assisted Intervention*, vol. 9901 of *LNCS* (Springer, 2016), pp. 478–486 (cited on pp. 85, 110).

175. J. Duchon. "Splines minimizing rotation-invariant semi-norms in Sobolev spaces", *Constructive Theory of Functions of Several Variables*, vol. 571 of *LNM* (Springer, 1977), pp. 85–100 (cited on p. 85).

176. O. Oktay, E. Ferrante, K. Kamnitsas, M. Heinrich, W. Bai, J. Caballero, S. A. Cook, A. de Marvao, et al. "Anatomically Constrained Neural Networks (ACNNs): application to cardiac image enhancement and segmentation", *IEEE Transactions on Medical Imaging*, vol. 37 (2018), pp. 384–395 (cited on p. 86).

177. N. Lessmann, B. van Ginneken, P. A. de Jong, and I. Išgum. "Iterative fully convolutional neural networks for automatic vertebra segmentation and identification", *Medical Image Analysis*, vol. 53 (2019), pp. 142–155 (cited on pp. 86–88).

178. F. L. Bookstein. "Principal warps: thin-plate splines and the decomposition of deformations", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 11 (1989), pp. 567–585 (cited on p. 89).

179. J. Zhao, M. Mathieu, and Y. LeCun. "Energy-based generative adversarial networks", *Proceedings of ICLR*, 2017 (cited on p. 93).

180. O. J. Dunn and V. Clark. "Correlation coefficients measured on the same individuals", *Journal of the American Statistical Association*, vol. 64 (1969), pp. 366–377 (cited on p. 99).

181. I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. "Generative adversarial nets", *Advances in Neural Information Processing Systems*, vol. 27 (2014), pp. 2672–80 (cited on p. 101).

182. B. D. de Vos, F. F. Berendsen, M. A. Viergever, H. Sokooti, M. Staring, and I. Išgum. "A deep learning framework for unsupervised affine and deformable image registration", *Medical Image Analysis*, vol. 52 (2019), pp. 128–143 (cited on p. 101).

183. J. C. Ross, R. S. J. Estépar, G. Kindlmann, A. Díaz, C.-F. Westin, E. K. Silverman, and G. R. Washko. "Automatic lung lobe segmentation using particles, thin plate splines, and maximum a posteriori estimation", *International Conference on Medical Image Computing and Computer-Assisted Intervention*, vol. 6363 of *LNCS* (Springer, 2010), pp. 163–171 (cited on p. 101).

184. T. Baum, J. S. Bauer, T. Klinder, M. Dobritz, E. J. Rummeny, P. B. Noël, and C. Lorenz. "Automatic detection of osteoporotic vertebral fractures in routine thoracic and abdominal MDCT", *European Radiology*, vol. 24 (2014), pp. 872–880 (cited on p. 102).

185. C. F. Buckens, G. Dijkhuis, B. de Keizer, H. J. Verhaar, and P. A. de Jong. "Opportunistic screening for osteoporosis on routine computed tomography? An external validation study", *European Radiology*, vol. 25 (2015), pp. 2074–79 (cited on p. 102).

186. J. E. Burns, J. Yao, and R. M. Summers. "Vertebral body compression fractures and bone density: automated detection and classification on CT images", *Radiology*, vol. 284 (2017), pp. 788–797 (cited on p. 102).

187. J. Y. Lee, A. R. Vaccaro, M. R. Lim, F. C. Öner, R. J. Hulbert, R. Hedlund, M. G. Fehlings, P. Arnold, et al. "Thoracolumbar injury classification and severity score: a new paradigm for the treatment of thoracolumbar spine trauma", *Journal of Orthopaedic Science*, vol. 10 (2005), pp. 671–675 (cited on p. 102).

188. N. Lessmann, B. van Ginneken, P. A. de Jong, W. B. Veldhuis, M. A. Viergever, and I. Išgum. "Deep learning analysis for automatic calcium scoring in routine chest CT", *Radiological Society of North America, 103rd Annual Meeting*, 2017 (cited on p. 110).

189. S. A. M. Gernaat, S. G. M. van Velzen, V. Koh, M. J. Emaus, I. Išgum, N. Lessmann, S. Moes, A. Jacobson, et al. "Automatic quantification of calcifications in the coronary arteries and thoracic aorta on radiotherapy planning CT scans of Western and Asian breast cancer patients", *Radiotherapy and Oncology*, vol. 127 (2018), pp. 487–492 (cited on p. 110).

190. B. Frenay and M. Verleysen. "Classification in the presence of label noise: a survey", *IEEE Transactions on Neural Networks and Learning Systems*, vol. 25 (2014), pp. 845–869 (cited on p. 111).

191. C. E. Brodley and M. A. Friedl. "Identifying mislabeled training data", *Journal of Artificial Intelligence Research*, vol. 11 (1999), pp. 131–167 (cited on p. 111).

192. J. Sander, B. D. de Vos, J. M. Wolterink, and I. Išgum. "Towards increased trustworthiness of deep learning segmentation methods on cardiac MRI", *Medical Imaging*, vol. 10949 of *Proceedings of SPIE* (2019), 1094944 (cited on p. 111).

193. M. Vonder, C. M. van der Aalst, R. Vliegenthart, P. M. A. van Ooijen, D. Kuijpers, J. W. Gratama, H. J. de Koning, and M. Oudkerk. "Coronary artery calcium imaging in the ROBINSCA trial", *Academic Radiology*, vol. 25 (2018), pp. 118–128 (cited on p. 111).

194. "Automated quantification of coronary artery calcifications on radiotherapy planning CTs for cardiovascular risk prediction in breast cancer patients: the BRAGATSTON study", ClinicalTrials.gov Identifier: NCT03206333, 2017 (cited on p. 111).

195. J. Šprem, B. D. de Vos, N. Lessmann, P. A. de Jong, M. A. Viergever, and I. Išgum. "Impact of automatically detected motion artifacts on coronary calcium scoring in chest computed tomography", *Journal of Medical Imaging*, vol. 5 (2018), 044007 (cited on p. 112).

196. J. Šprem, B. D. de Vos, N. Lessmann, R. W. van Hamersvelt, M. J. W. Greuter, P. A. de Jong, T. Leiner, M. A. Viergever, and I. Išgum. "Coronary calcium scoring with partial volume correction in anthropomorphic thorax phantom and screening chest CT images", *PLOS ONE*, vol. 13 (2018), e0209318 (cited on p. 112).

197. P. Moeskops, M. Veta, M. W. Lafarge, K. A. J. Eppenhof, and J. P. W. Pluim. "Adversarial training and dilated convolutions for brain MRI segmentation", *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, vol. 10553 of *LNCS* (Springer, 2017), pp. 56–64 (cited on p. 112).

198. P. D. Edwards, R. K. Bull, and R. Coulden. "CT measurement of main pulmonary artery diameter", *The British Journal of Radiology*, vol. 71 (1998), pp. 1018–20 (cited on p. 113).

199. E. Pompe, P. A. de Jong, D. A. Lynch, N. Lessmann, I. Išgum, B. van Ginneken, J.-W. J. Lammers, and F. A. A. Mohamed Hoesein. "Computed tomographic findings in subjects who died from respiratory disease in the National Lung Screening Trial", *European Respiratory Journal*, vol. 49 (2017), 1601814 (cited on p. 113).

Acknowledgments

Completing this thesis is a milestone that I could have never reached without the support of a great number of people. I am grateful to all those who have contributed over the years, in many different forms, and I would like to thank a few people in particular.

Dear Ivana, first and foremost, thank you for your trust in me. I have benefited tremendously from the freedom you gave me, the projects you involved me in, and all other forms of your longstanding support. I could always rely on you having my back. You often worked hard for my benefit, many times until late in the evening and too many times also during your vacation. Thank you for always being available for advise and for teaching me so many things about academia.

Dear Bram, thank you for all your solid advise in the past years. This was something that I could always count on and that helped me a lot with finishing papers or whenever I felt stuck. Thank you also for welcoming me at your group in Nijmegen for pulmo meetings, workshops, etc., and for your generous support in the last months that allowed me to finish the final chapter.

Dear Max, thank you for the detailed feedback that I received from you for each of my articles. I am also very grateful for your compliments on achievements along the way, whether it was an accepted paper, a conference paper draft or just a presentation. These were encouraging, especially during stressful times. Thank you also for supporting the idea of an ISI Dutch course in a heartbeat.

Dear Pim, thank you for bringing in the perspective of a radiologist. It was very helpful that you always replied quickly to emails or came to my office when you had time to answer my questions. Thank you especially for focusing on the clinical details of papers for technical journals, and for immediately being willing to become more involved when we wanted to write a paper for a clinical journal.

I would like to thank the members of the reading committee, Prof. dr. Öner, Prof. dr. de Bruijne, Prof. dr. Mali, Prof. dr. Verkooijen and Dr. Lorenz, for their time and effort spent on evaluating this thesis. And I would like to thank all co-authors for the good collaborations and the helpful feedback over the years.

Several other people have been part of this journey. Majd, we shared an office for most of my time at ISI and I enjoyed and benefited a lot from our discussions and brainstorming sessions. Thank you for being a reliable friend, and for being my paranymph. Nadieh, it has been a lot of fun sharing the office with you. Thank you for the many kilograms of nuts and raisins that you supplied, and for also being my

paranymph. Jurica, roommate #1, thank you for many enjoyable moments, including the trip through Florida with you and Mariëlle, some of your jokes, and many more. Jelmer and Bob, I learned a lot from you, especially when we worked together on projects like the SPIE deep learning workshop or the RSNA educational exhibit. Robbert, we kind of also shared an office since you dropped by all the time for coffee, which were nice and very welcome breaks. The QIA group, thank you all for making it a pleasant group to work in and for fun weekends, dinners, pancakes, etc. And everyone at ISI, for sharing many coffee breaks, cakes, conference visits, ski trips, Ardennen weekends, and a lot more.

Dear Chris, thank you for explaining many things about tulip and bunzing, about the pipelines and the Embarcadero hell in the beginning. Dear Gerard, thank you for your always very quick and reliable help with IT things. Dear Jacqueline, Marjan and Renée, thank you for your help with administrative matters, and especially for your support of the ISI Dutch course.

Even though I was only once in a while in Nijmegen, several people from DIAG have contributed to this thesis as well. Dear Colin and Paul, thank you for your help with transferring the calcium scoring algorithm and for being reliable contacts for all kinds of questions. Dear Arnaud and Francesco, thank you for getting me started with deep learning in my very first year. Furthermore, I always received a lot of useful feedback when I presented at the pulmo meeting, and for this I would like to thank especially Ernst Scholten, Monique Brink and Cornelia Schaefer-Prokop.

Dear Tobias and Rafael, without your enthusiasm for research and the time that you invested in me during my Master research project, I would never have started this journey. Tobias, thank you for your encouraging follow-up emails and for recommending me to Bram in the first place. Rafael, your enthusiastic reaction when you heard about my interview for a PhD position in Utrecht was the moment that decided that I would actually go for it.

Dear Hui Shan, meeting you has certainly been the best result of this PhD. Thank you for sharing the good moments and for distracting and supporting me in stressful moments.

Publications

Journal publications

N. Lessmann, J. M. Wolterink, M. Zreik, P. A. de Jong, W. Foppen, S. C. E. Diepstraten, M. A. Viergever, B. van Ginneken, and I. Išgum. "Thin-plate spline surfaces steered by a convolutional neural network: application to vertebra partitioning based on shape information". Submitted.

N. Lessmann, B. van Ginneken, P. A. de Jong, and I. Išgum. "Iterative fully convolutional neural networks for automatic vertebra segmentation and identification", *Medical Image Analysis*, vol. 53 (2019), pp. 142–155.

B. D. de Vos, J. M. Wolterink, T. Leiner, P. A. de Jong, **N. Lessmann**, and I. Išgum. "Direct automatic coronary calcium scoring in cardiac and chest CT". *IEEE Transactions on Medical Imaging* (2019). In press.

N. Lessmann, P. A. de Jong, C. Celeng, R. A. P. Takx, M. A. Viergever, B. van Ginneken, and I. Išgum. "Sex differences in coronary artery and thoracic aorta calcification and their association with cardiovascular mortality in heavy smokers". *JACC: Cardiovascular Imaging* (2019). In press.

J. Šprem, B. D. de Vos, **N. Lessmann**, R. W. van Hamersvelt, M. J. W. Greuter, P. A. de Jong, T. Leiner, M. A. Viergever, and I. Išgum. "Coronary calcium scoring with partial volume correction in anthropomorphic thorax phantom and screening chest CT images", *PLOS ONE*, vol. 13 (2018), e0209318.

J. Šprem, B. D. de Vos, **N. Lessmann**, P. A. de Jong, M. A. Viergever, and I. Išgum. "Impact of automatically detected motion artifacts on coronary calcium scoring in chest computed tomography", *Journal of Medical Imaging*, vol. 5 (2018), 044007.

S. A. M. Gernaat, S. G. M. van Velzen, V. Koh, M. J. Emaus, I. Išgum, **N. Lessmann**, S. Moes, A. Jacobson, et al. "Automatic quantification of calcifications in the coronary arteries and thoracic aorta on radiotherapy planning CT scans of Western and Asian breast cancer patients", *Radiotherapy and Oncology*, vol. 127 (2018), pp. 487–492.

N. Lessmann, B. van Ginneken, M. Zreik, P. A. de Jong, B. D. de Vos, M. A. Viergever, and I. Išgum. "Automatic calcium scoring in low-dose chest CT using deep neural networks with dilated convolutions", *IEEE Transactions on Medical Imaging*, vol. 37 (2018), pp. 615–625.

M. Zreik, **N. Lessmann**, R. W. van Hamersvelt, J. M. Wolterink, M. Voskuil, M. A. Viergever, T. Leiner, and I. Išgum. "Deep learning analysis of the myocardium in coronary CT angiography for identification of patients with functionally significant coronary artery stenosis", *Medical Image Analysis*, vol. 44 (2018), pp. 72–85.

E. Pompe, P. A. de Jong, D. A. Lynch, **N. Lessmann**, I. Išgum, B. van Ginneken, J.-W. J. Lammers, and F. A. A. Mohamed Hoesein. "Computed tomographic findings in subjects who died from respiratory disease in the National Lung Screening Trial", *European Respiratory Journal*, vol. 49 (2017), 1601814.

Conference proceedings

S. G. M. van Velzen, M. Zreik, **N. Lessmann**, M. A. Viergever, P. A. de Jong, H. M. Verkooijen, and I. Išgum. "Direct prediction of cardiovascular mortality from low-dose chest CT using deep learning", *Medical Imaging*, vol. 10949 of *Proceedings of SPIE* (2019), 1094932.

N. Lessmann, B. van Ginneken, P. A. de Jong, and I. Išgum. "Iterative fully convolutional neural networks for automatic vertebra segmentation", *Medical Imaging with Deep Learning (MIDL)*, 2018.

N. Lessmann, B. van Ginneken, and I. Išgum. "Iterative convolutional neural networks for automatic vertebra identification and segmentation in CT images", *Medical Imaging*, vol. 10574 of *Proceedings of SPIE* (2018), 1057408.

N. Lessmann, I. Išgum, A. A. A. Setio, B. D. de Vos, F. Ciompi, P. A. de Jong, M. Oudkerk, W. P. Th. M. Mali, M. A. Viergever, and B. van Ginneken. "Deep convolutional neural networks for automatic coronary calcium scoring in a screening study with low-dose chest CT", *Medical Imaging*, vol. 9785 of *Proceedings of SPIE* (2016), 978511.

Conference abstracts

N. Khalili, **N. Lessmann**, E. Turk, M. A. Viergever, M. Benders, and I. Išgum. "Brain tissue segmentation in fetal MRI using convolutional neural networks with simulated intensity inhomogeneities", *Annual Meeting of the International Society for Magnetic Resonance in Medicine*, 2019.

A. Schreuder, C. Jacobs, **N. Lessmann**, E. Scholten, I. Išgum, M. Prokop, C. Schaefer-Prokop, and B. V. Ginneken. "MA20.09 improved lung cancer and mortality prediction accuracy using survival models based on semi-automatic CT image measurements", *Journal of Thoracic Oncology*, vol. 13 (2018), p. S428.

R. van Hamersvelt, M. Zreik, **N. Lessmann**, J. Wolterink, M. Voskuil, M. A. Viergever, T. Leiner, and I. Išgum. "Improving specificity of coronary CT angiography for the detection of functionally significant coronary artery disease: a deep learning approach", *Radiological Society of North America, 103rd Annual Meeting*, 2017.

N. Lessmann, B. van Ginneken, P. A. de Jong, W. B. Veldhuis, M. A. Viergever, and I. Išgum. "Deep learning analysis for automatic calcium scoring in routine chest CT", *Radiological Society of North America, 103rd Annual Meeting*, 2017.

B. D. de Vos, **N. Lessmann**, P. A. de Jong, M. A. Viergever, and I. Išgum. "Direct coronary artery calcium scoring in low-dose chest CT using deep learning analysis", *Radiological Society of North America, 103rd Annual Meeting*, 2017.

M. Zreik, **N. Lessmann**, R. van Hamersvelt, J. Wolterink, M. Voskuil, M. A. Viergever, T. Leiner, and I. Išgum. "Deep learning analysis of the left ventricular myocardium in cardiac CT images enables detection of functionally significant coronary artery stenosis regardless of coronary anatomy", *Radiological Society of North America, 103rd Annual Meeting*, 2017.

F. A. A. Mohamed Hoesein, E. Pompe, D. A. Lynch, **N. Lessmann**, J. W. J. Lammers, I. Išgum, and P. A. de Jong. "Computed tomographic findings are associated with respiratory mortality in the National Lung Screening Trial", *Radiological Society of North America, 102nd Annual Meeting*, 2016.

N. Lessmann, I. Išgum, S. Lam, J. Mayo, P. A. de Jong, M. A. Viergever, and B. van Ginneken. "Automatic coronary calcium scoring and cardiovascular risk estimation in the Pan-Canadian lung cancer screening trial", *Radiological Society of North America, 101th Annual Meeting*, 2015.

Biography

Nikolas Leßmann was born on July 5, 1988 in Münster, Germany. After secondary school, his plan was to study either mechanical engineering or chemistry, but nine months of compulsory community service at Münster's university hospital changed his mind. He studied biomedical engineering in Lübeck, Germany, and received a Master of Science degree in 2014. In his final year, he worked on automatic analysis of chest CT scans during an internship at Philips Research in Hamburg, Germany. Afterwards, he continued working on this topic as a PhD candidate at the Image Sciences Institute at the University Medical Center Utrecht. The results are described in this thesis. He continues his research as postdoctoral researcher at the Diagnostic Image Analysis Group at Radboud University Medical Center in Nijmegen.