Automatic Image Quality Parameter Determination for Thorax CT

van Haren, Lisan M.A.A.

Award date:
2022

EINDHOVEN UNIVERSITY OF TECHNOLOGY

Department of Applied Physics

Automatic Image Quality Parameter Determination for Thorax CT

by

L.M.A.A. (Lisan) van Haren

MSC THESIS

Assessment committee

Member 1 (chair):	dr. ir. C. van Pul
Member 2:	dr. ir. I. Zivkovic
Member 3:	prof. dr. ir. E.J.E. Cottaar
Advisory member 1:	dr. C.R.L.P.N. Jeukens

Graduation

Program:	Applied Physics
Capacity group:	SMPE/e
Supervisor:	dr. ir. C. van Pul
Date of defense:	December 5, 2022
Student ID:	1009441
Study load (ECTS):	60
Track:	Fluids, Bio and Soft Matter
External supervisor(s):	dr. C.R.L.P.N. Jeukens

The research of this thesis has been carried out in collaboration with *Máxima Medical Center* and *Maastricht University Medical Center+*.
This thesis is public and Open Access.

Abstract

Purpose: To develop an algorithm that automatically determines noise and contrast-related image quality (IQ) metrics for quality control purposes in pulmonary embolism (PE) CT-scans, and to assess agreement between objective IQ metrics and subjective IQ scoring by radiologists.

Methods: In two hospitals, two datasets of 50 consecutive clinical PE CT-scans were retrospectively collected, including repeated scans having a lower IQ. We developed four different algorithms for lungs and pulmonary vessels segmentation based on: thresholding, U-net + thresholding, U-net + K-means clustering and U-net + Otsu thresholding. In these 3D segmentations five IQ metrics were calculated: noise and signal-to-noise ratio in lungs, mean signal in vessels, and contrast and contrast-to-noise ratio between lungs and vessels. Additionally, five noise IQ metrics reported in literature were calculated. In each hospital, two radiologists scored noise, contrast attenuation of the pulmonary arteries and diagnostic confidence separately using a 5-point Likert scale (1 = poor to 5 = excellent). Additionally, presence of lung pathology (PE, lesions, emphysema, effusion) was recorded, as this may influence the automatic segmentation and calculated IQ-scores. Regression analysis was performed to assess correlation between IQ metrics and Likert scores, reporting adjusted R^2 and significance.

Results: The algorithm was able to automatically calculate all IQ metrics for all scans. Multiple significant correlations were found between IQ metrics and all 3 different radiologist scores. Highest correlations with radiologist scores for contrast enhancement and diagnostic confidence were found for signal vessel and contrast metrics. With radiologist scores for noise highest correlations were found for IQ metrics from literature. Often when correlations were significant the model was also able to discriminate between sufficient and insufficient IQ, although this was not always the case. Using linear regression with backward elimination, highest correlations with diagnostic confidence were found for the segmentation method that used a U-net + thresholding for both hospitals (R^2 = 0.48/0.33). Segmentations were not always accurate when pathology was present in the lungs. This could also be seen in the results as for some IQ metrics significant differences were found between results of patients with and without pathology in the lungs.

Conclusion: It is feasible to determine objective IQ metrics that correlate with subjective IQ, based on an automatic algorithm in PE CT-scans. For the diagnostic confidence of PE scans contrast IQ metrics in pulmonary vessels were most important. This opens possibilities for continuous IQ monitoring in clinical practice.

Contents

1 Introduction

Over-testing for pulmonary embolism (PE) is a major public health problem [1]. In addition, often scans are repeated due to inadequate timing of the contrast agent. A recent study showed a repeat rate for CT pulmonary angiography (CTPA) of 3% [2]. Another study showed a CTPA repeat rate of 1.4% and for large patients the repeat rate increased up to 11.2% [3]. Motion artifacts or bad contrast enhancement are the main reasons for CTPA to fall below diagnostic standards [4]. With an increasing use of CT scans the radiation exposure of patients also increases, which gives a potential risk of cancer. The cancer risk due to radiation exposure increases for young patients, which may be a problem as the occurrence of PE is quite common even for young patients ($<$ 40 years) [5–7]. To prevent that patients receive more dose than necessary, it is important to optimize the scan protocol according to the ALARA principle: to use the lowest dose that still gives acceptable image quality (IQ) to answer the clinical question. This optimization process is however a time consuming task and it is not straightforward, as IQ depends on the patients' anatomy and minimal acceptable IQ varies per clinical task.

Currently to determine IQ, either a physical phantom or a score given by radiologists is used. Disadvantages of these are that a phantom does not resemble patients' anatomy and manual scoring by radiologists is a time consuming task and results are subjective since they may vary per radiologist. Also, with both methods it is not possible to monitor IQ continuously.

In previous studies, methods for automated IQ quantification based on patient CT scans have been investigated [8–14]. In these studies IQ metrics were defined by noise, contrast and/ or spatial resolution, mainly focused on the soft tissue regions.

Particularly for chest CT, Reeves et al. investigated automated IQ assessment [14]. The noise and image intensity were calculated in three different, automatically segmented, regions of interest (ROI): external air, trachea lumen air and descending aorta blood. However, these are not the regions of interest for the radiologists when they have to diagnose PE.
PE is a blockage of a pulmonary artery by a blood clot, which is common and often fatal [15]. For PE detection radiologists check all pulmonary vessels to search for an interruption caused by a blood clot. In CTPA contrast in the pulmonary vessels is enhanced using intravenous contrast material. Optimal timing of contrast material bolus is mostly determined using bolus-tracking or a test bolus. A diagnostic CTPA should have sufficient contrast enhancement in the pulmonary vessels, such that the blood clot can be distinguished from the pulmonary vessels.

Automatically calculating IQ metrics in clinical images is needed to monitor performance in daily practice. In the lungs this would require automated recognition of the lungs and pulmonary vessels. Segmentation of lungs and pulmonary vessels has already been investigated in previous studies using different techniques, however the presence of large abnormalities in the lungs can sometimes cause results to be less accurate [16–18].

The aim of this study is to develop a robust method for segmentation based IQ determination. First, our method was tested on an anthropomorphic phantom scanned with different scanner settings. Next, we tested our method in thorax CT images of patients scanned for PE detection, and evaluated the performance of the IQ metrics to subjective IQ scored by radiologists. We compared our newly developed IQ metrics in the segmented lungs and pulmonary vessels to methods from literature, where IQ is calculated over the whole image or in regions outside the lungs.

Report outline

In Chapter 2, the methods are explained and in Chapter 3 the results are presented. The results will be discussed in Chapter 4 and in Chapter 5 a conclusion is given. Further, in Appendix A.1 the basics behind CT-scans are explained and in Appendix A.2 the theory behind artificial intelligence is described. In Appendix B an additional study is described for thorax CT-scans for noduli detection. In Appendix C more information about the choices of model parameters of our method are given. Also, in Appendix D and E extended results for respectively phantom and patient datasets are presented. In Appendix F an additional statistical method is explored. Lastly, in Appendix G two abstracts that were submitted about this study are presented.

2 Materials and methods

The pipeline of our study is shown in Figure 2.1. In two hospitals a dataset of CT-scans of both an anthropomorphic phantom and patients scanned for PE detection. For all scans IQ is automatically calculated using different methods and for the anthropomorphic phantom IQ is compared with dose and for the patient PE scans IQ is compared with subjective IQ of radiologists.

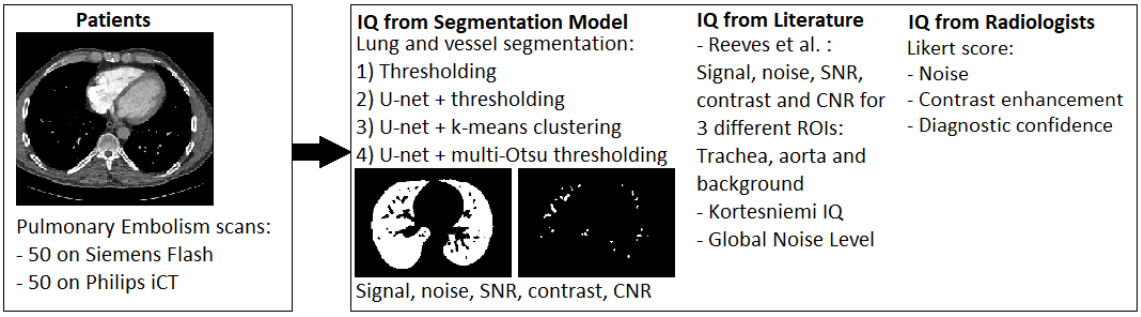


Figure 2.1: Pipeline of this study: 50 patients from each hospital were processed automatically. Using 4 different segmentation methods IQ metrics were calculated, simultaneously with IQ metrics from literature. Also, for subjective IQ, radiologists scored the scans.

2.1 Data

2.1.1 Patient dataset

We obtained CT images from 2 different centers: Hospital 1 (*Maastricht University Medical Center, Maastricht, The Netherlands*) and Hospital 2 (*Máxima Medical Center, Veldhoven, The Netherlands*). The patient scans were retrospectively collected and a waiver was obtained for this study by both the ethics committees from both hospitals.
The two retrospective datasets consist of 50 patients per center, scanned with a PE protocol, using a Siemens Flash in Hospital 1 and a Philips iCT in Hospital 2. The inclusion criteria were that patients were referred from the ER and were suspected of having a PE. In Hospital 1 iterative reconstruction (ADMIRE) with kernel I26f and 1mm slice thickness was used. In Hospital 2 iterative reconstruction (iDose) with kernel B (standard) and 2mm slice thickness was used. In Hospital 1 iopromide (Ultravist 300) was used as intravenous contrast injection and the contrast injection protocol was adapted to both patient body weight and kV-setting via a predefined formula. More information about this protocol can be found in the study of Hendriks et al. [19]. Further, a test bolus technique was used. In Hospital 2 an intravenous contrast injection of 60 ml iohexol (Omnipaque 300) was used with a flow rate of 4 ml/s, and the bolus tracking technique was used where the scan was started when a difference of 50 HU was reached. More information about CT-scanners, settings and contrast bolus timing can be found in Appendix A.1.
In order to obtain a dataset with more variation in IQ, in both datasets 10 scans were included that had to be repeated due to a bad timing of contrast agent injection, which causes the contrast between pulmonary vessels and possible blood clots to be too low for diagnosis. Also, for 20 patients in Hospital 1 and 10 patients in Hospital 2 a filtered back projection (FBP) was reconstructed retrospectively and included instead of the iterative reconstruction, which was expected to give a noisier image compared to the iterative reconstruction algorithms that were used in clinical practice [13].
Patient characteristics with respect to absence and presence of pathology are shown in Table 2.1. Patients with small insignificant PEs and patients with small noduli with a diameter < 1 cm were categorized as 'without significant abnormalities', since their areas were expected to be negligible in the 3D volumes that were analyzed in this study.

Table 2.1: Patient characteristics of the PE datasets

	Total # of patients	# of patients without significant abnormalities in lungs	# of patients with PE	# of patients with significant abnormalities in lungs
Hospital 1	50	26	4	20
Hospital 2	50	22	9	19

2.1.2 Phantom dataset

In both hospitals an anthropomorphic phantom (PBU-60 phantom, Kyoto Kagaku Co., Ltd., Kyoto, Japan) was scanned to check whether IQ also varied for different scanner settings. The phantom has been scanned on a Siemens Flash (Hospital 1) with the standard lung embolism protocol and on a Philips Ingenuity (Hospital 2) with the standard thorax protocol. For each scanner scans were made with varying tube current, tube voltages and reconstruction algorithms. On both scanners the tube currents ranged from 10 to 180 mAs and the tube voltages were either 80, 100 or 120 kVp. Additionally, for the Siemens Flash the reconstruction algorithms filtered back projection (FBP) and iterative reconstruction (ADMIRE) with strength 3 and 5 and with kernel I26f and a slice thickness of 1 mm were applied. For the Philips Ingenuity the reconstruction algorithms FBP and iterative reconstruction (iDose) with strength 3 and 5 and with kernel B (standard) and a slice thickness of 3 mm. More details about the phantom dataset can be found in Appendix D.

2.2 Image quality assessment

2.2.1 Image quality models using lung and pulmonary vessel segmentation

In order to determine the noise and contrast in the regions of interest, the lungs and pulmonary vessels were segmented in both the anthropomorphic phantom and patients. Segmentation was done in 3 dimensions from the lung apex to the base, indicated manually. In order to investigate how the lungs and pulmonary vessels can best be segmented for our purpose, segmenting was done using different methods, each resulting in a lung and vessel mask. In Figure 2.2 an overview of the segmentation methods is given step by step, also described below. Details on the artificial intelligence (AI) based methods are found in Appendix A.2.

Four methods are investigated, of which three include a first step of segmentation using a pre-trained U-net model created by Hofmanninger et al. [17]. This model has been trained on a large dataset and it includes air pockets, tumors and effusions.
In step 2, lungs and pulmonary vessels were separated, performed differently for each method:
Method 1 - Thresholding: Lungs were defined as a region within the patients' body with a Hounsfield Unit (HU) below a threshold of -500. Regions within the lung area with an HU above this threshold were defined as pulmonary vessel.
Method 2 - U-net + thresholding: Within the U-net lung mask the same thresholding was performed as in Method 1, so the region with HU below a threshold of -500 was defined as lung and the region with HU above the threshold was defined as pulmonary vessel.
Method 3 - U-net + K-means clustering: Within the U-net lung mask K-means clustering was performed using 4 clusters. K-means clustering is an unsupervised machine learning technique that divides all pixels into K clusters, where each pixel belongs to the cluster with the nearest mean. The cluster with the lowest HU was defined as lung and the cluster with highest HU was defined as pulmonary vessel.
Method 4 - U-net + Otsu thresholding: Within the U-net lung mask two thresholds were determined automatically for each patient using multilevel Otsu's thresholding, where the optimal thresholds were chosen by maximizing the between-class variance with an exhaustive search [20, 21]. The region within the resulting lung mask with an HU below the lowest threshold was defined as lung and the region with an HU above the highest threshold was defined as pulmonary vessel.

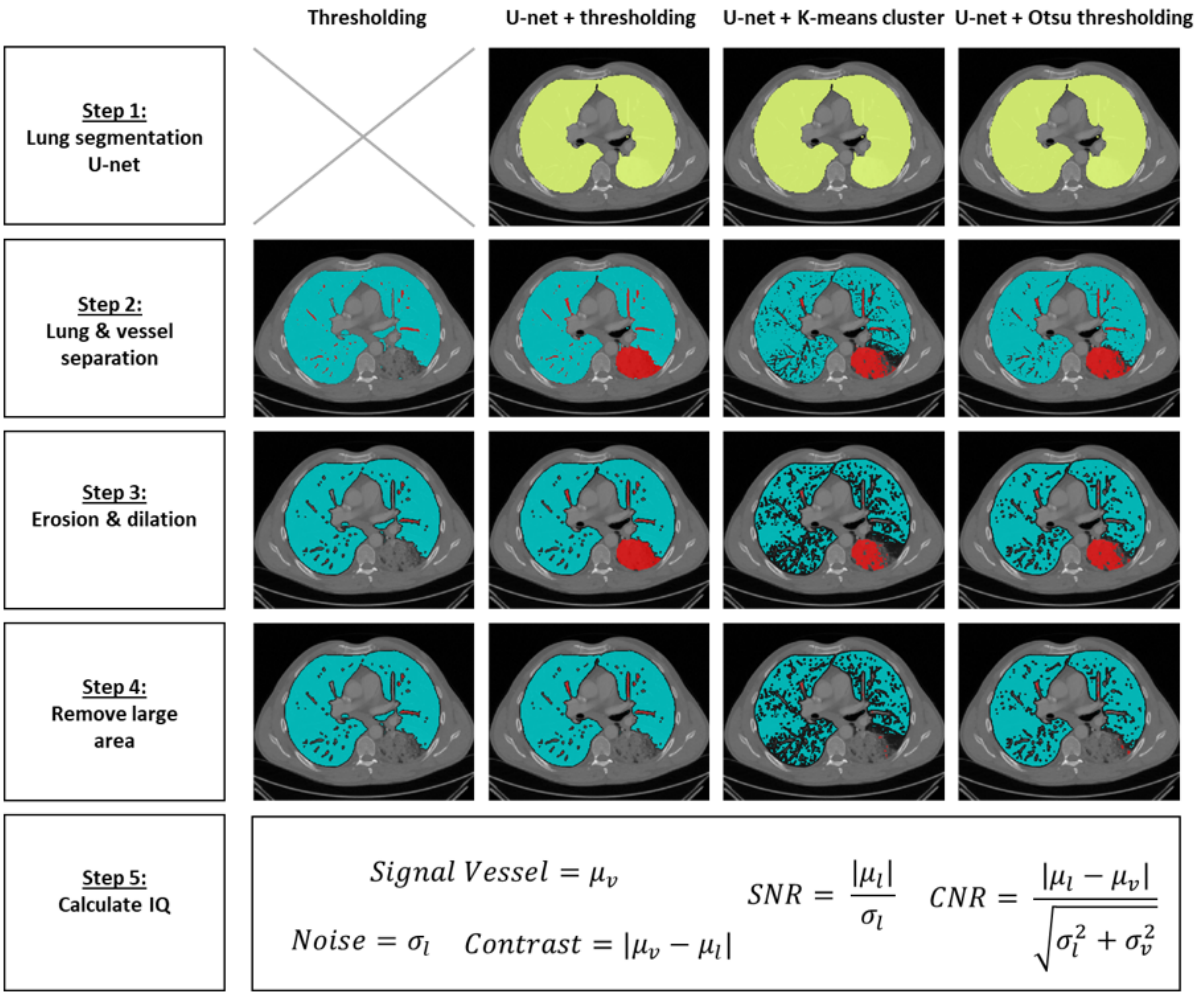


Figure 2.2: Pipeline of the segmentation method for one slice of a patient. In step 1 the lungs were segmented by a U-net model (in yellow). In step 2 the vessels and lungs were separated using different methods (lungs in blue and vessels in red). In step 3 erosion and dilation was applied and in step 4 large areas were removed from the vessel regions. Finally in step 6 the IQ was calculated from the segmentations.

In step 3 on the resulting lung masks for all methods, erosion with a disk of size 2 was applied to remove possible edges. On the resulting vessel masks first erosion with a disk with a size of 2 was applied to remove very small structures, followed by a dilation step of size 1. In this way, the larger vessels remained. The segmentation model was not always capable of distinguishing pulmonary vessels from pathology and some pathologies were significantly larger by area than the typical pulmonary vessel, such as tumors or emphysema. Therefore, in step 4 a constraint was added that sets the maximum area in the transverse plane that a pulmonary vessel was allowed to be to 2.5 cm^2. All regions with a larger area were removed from the pulmonary vessel mask since they were considered too likely to consist of abnormalities instead of pulmonary vessels.

Finally in step 5, the following IQ metrics were calculated. In the segmented lungs the noise σ_l was calculated, where the noise was defined as the standard deviation of the signal. Also, the signal-to-noise ratio was calculated for the lung ROI, which was defined as

$$SNR = \frac{|\mu_l|}{\sigma_l},$$

with μ_l the mean lung signal. Additionally, the mean signal in the pulmonary vessels μ_v and the contrast between the lungs and pulmonary vessels were calculated. Finally, the contrast-to-noise ratio

was calculated between the lungs and vessels according to

$$CNR = \frac{|\mu_l - \mu_v|}{\sqrt{\sigma_l^2 + \sigma_v^2}},$$

with σ_v the vessel noise.
The IQs were calculated for all segmented slices and the average IQ scores were taken as final IQs.

Choices made in the segmentation methods (threshold, cluster number, disk sizes and maximum vessel area) were based on analysis and visual inspection of results with varying parameters. More info on model parameters and choices can be found in Appendix C.

2.2.2 Image quality models from literature

In addition, IQ of the patients was determined using 3 different methods from literature.

First, a simplified 2D version of the model proposed by Reeves et al. [14] was implemented, where ROIs were selected manually in external air, trachea lumen air and descending aorta blood. Next, the signal was determined in the descending aorta blood, the noise and SNR were calculated for both the external air, trachea lumen air and descending aorta blood, and the contrast and CNR were determined between the trachea lumen air and descending aorta blood. This results in a total of 9 different IQ metrics.

Second, the Kortesniemi IQ [9] was determined, which relates to the noise. In short, a mask of 3 x 3 pixels was moved around each pixel in the 9 different possible positions and for each position the standard deviation within that mask was calculated. The lowest standard deviation per movement was chosen and saved in a standard deviation map. Pixels that contained air or sharp edges were excluded, by only taking into account pixels with a CT-value above -500 HU and removing the highest 5% of standard deviations from the map. From the resulting map the IQ was calculated according to

$$IQ = \frac{n_{sel}}{\sum_{i,j} \sqrt{S_{i,j}}},$$

with n_{sel} the number of pixels included and $S_{i,j}$ the saved standard deviation of the pixel in row i and column j. Normalization was applied using the reverse sigmoid curve equation. The Kortesniemi IQ was calculated for all slices that were also included in the segmentation method and the average IQ over these slices was used as the final Kortesniemi IQ. A higher score represents lower noise in the image.

Lastly, the GNL and GNL Air were determined, which relates to noise [8, 12]. Here, a mask of 7 x 7 pixels was moved around each pixel that was included. For the GNL this was done in the soft tissue range (0 - 100 HU) and for the GNL Air in the air range ($<$ -500 HU). The mean and mode were calculated from the standard deviations histograms. This results in 4 different noise metrics: GNL Mode, GNL Median, GNL Mode Air and GNL Median Air. For all metrics a higher score represents higher noise. Again, the GNL scores were calculated for all slices included in the segmentation method and the average GNL scores were used as final IQ metrics.

2.2.3 Radiologist scoring

To compare the automatically calculated IQ with subjective IQ, scoring by radiologists was used. The dataset of Hospital 1 was scored separately by one radiologist (2 years of experience) and one last year radiology resident. The dataset of Hospital 2 was scored separately by two radiologists (8 and 9 years of experience). A 5-point Likert scale was used for scoring (1 = very poor, 2 = poor, 3 = moderate, 4 = good, 5 = excellent) and for each scan radiologists were asked to give three different categorical scores: one for noise, one for contrast enhancement in the central until segmental pulmonary vessels and one for the diagnostic confidence to diagnose PE. The scans were presented in a random order on a diagnostic screen and radiologists were able to scroll through the scan volumes, to adjust the window settings and to zoom in or out. The scans were presented anonymously, without information about the scan parameters. The average scores of the two radiologists were calculated and used as the subjective IQ scores.

2.3 Statistics

For statistical analysis, first the automatically calculated IQ metrics from both the segmentation methods and the literature methods were compared with the Likert score diagnostic confidence. For this, correlation was checked separately for each calculated IQ metric using adjusted R^2. Also, using a Wilcoxon rank sum test it was determined whether the automatically calculated IQ metrics differ significantly between sufficient diagnostic confidence and insufficient diagnostic confidence according to subjective IQ, where a Likert score ≥ 3 was considered sufficient IQ and a Likert score < 3 was considered insufficient IQ. These procedures were repeated to determine the relation of noise related IQ metrics with Likert score noise and contrast related IQ metrics with Likert score contrast.

Next, to investigate for which segmentation method IQ metrics gave a highest overall correlation with Likert score diagnostic confidence, linear regression was used for all segmentation based IQ metrics. Collinearity was checked first and the metrics with highest variance inflation factor (VIF) were removed one by one, until all included metrics had a VIF below 10. After this, linear regression with backward elimination was applied to remove the least useful metrics. For each segmentation method the model's adjusted R^2 was reported.

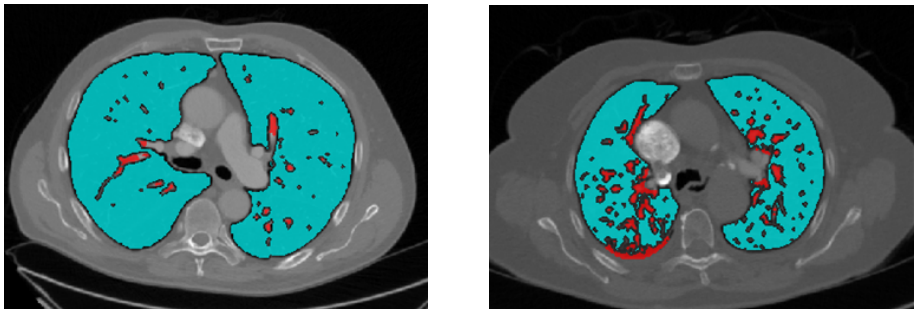
Finally, it was investigated whether the presence of pathology in the lungs influences the segmentation based IQ results, as segmentations may be less accurate when pathology is present. This was done by dividing the datasets in two groups: one group with all patients without significant pathology and one group with all patients with significant pathology (including PE). Using a Wilcoxon rank sum test it was determined whether the segmentation based IQ metrics have significant difference between both groups. This was also determined for the three different Likert scores, to check whether both groups were equally distributed over the different Likert scores.

3 Results

In this chapter the main results for the patient datasets are given. Results of the phantom dataset can be found in Appendix D and extended results of the patient datasets can be found in Appendix E.

3.1 Visual analysis of segmentations

The four segmentation methods developed on the anthropomorphic phantom showed that automated segmentation is possible. For the patient datasets, visual analysis of the segmentations showed that for patients without pathology present in the lungs segmentations were successful for all 4 methods. In segmentation method 1 the trachea is often classified as lungs. However, segmentations were sometimes influenced by pathology. The segmentation models were sometimes unable to distinguish pathology from pulmonary vessels, particularly if larger areas of effusion, pneumonia or emphysema were present that were not large enough to be excluded by the criteria that vessels may not be larger than 2.5 cm^2. This issue was present for all segmentation methods, but as method 3 and 4 obtained smaller and more homogeneous masks also less pathology was included. Segmentations using method 2 are visualized for one patient without significant pathology in the lungs in Figure 3.1a and for one patient with pleural effusion in Figure 3.1b, where excess fluid is classified as pulmonary vessel.



(a) (b)

Figure 3.1: Segmentation of lungs (blue) and pulmonary vessels (red), in (a) for a slice of a patient without pathology in the lungs and in (b) for a slice of a patient with pleural effusion in the lungs.

3.2 Likert scores

In Figure 3.2 the confusion matrices for both hospitals are presented for all 3 types of Likert scores. The scores of both radiologists are compared. It can be seen that in Hospital 1 the scores were more similar for the contrast enhancement Likert scores and diagnostic confidence Likert scores than for Hospital 2. For both Hospital 1 and 2 the contrast enhancement Likert scores are highly correlated with the diagnostic confidence Likert scores ($R^2 = 0.86/0.81$), whereas the noise Likert scores had low correlation with the diagnostic confidence Likert scores ($R^2 = 0.09/0.009$).

For the three types of Likert scores more patients were scored as sufficient IQ (Likert ≥ 3) than as insufficient IQ (Likert < 3). Respectively Hospital 1 and 2 consisted of 9 and 8 patients scores as insufficient diagnostic confidence and respectively 41 and 42 patients scored as sufficient diagnostic confidence. Both hospitals consisted of 8 patients scored as insufficient noise and 42 patients scored as sufficient noise. Lastly, Hospital 1 and 2 respectively consisted of 10 and 7 patients scored as insufficient contrast enhancement and respectively 40 and 43 as sufficient contrast enhancement.

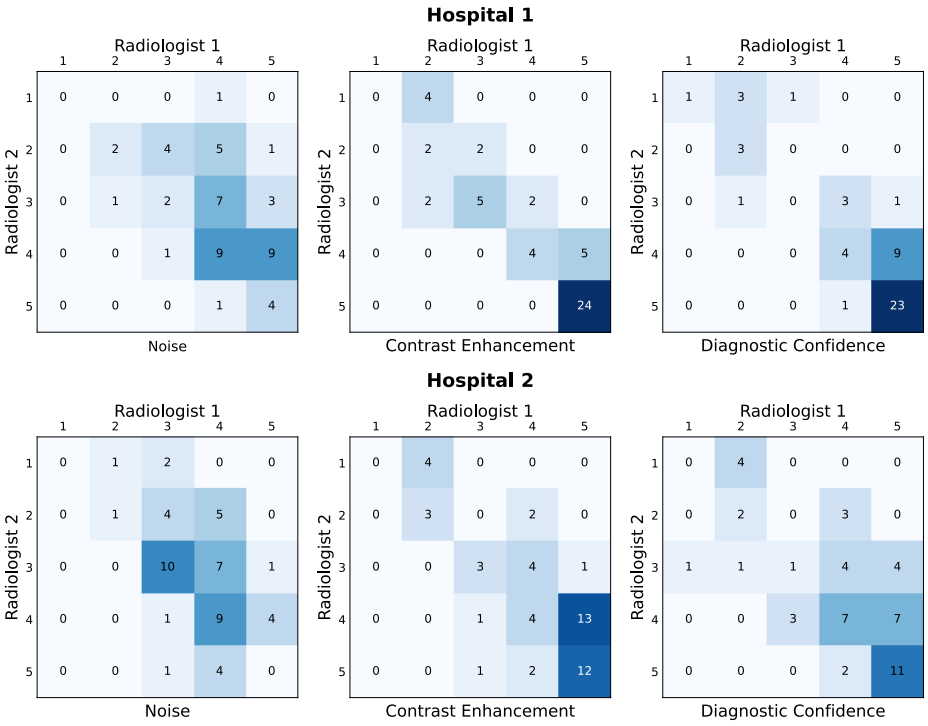


Figure 3.2: The confusion matrices of both hospitals are given for the three types of Likert scores, where the scores from both radiologists are compared per hospital.

3.3 IQ metrics against diagnostic confidence Likert scores

For each of the four segmentation methods five IQ metrics were determined: noise, SNR, signal vessel, contrast and CNR. This means that 20 different segmentation based correlations were checked with diagnostic confidence Likert scores (column 2 and 4 of Table 3.1). 17 of these correlations were significant in Hospital 1 and 8 were significant in Hospital 2. In Hospital 1, all 5 metrics were significant for segmentation method 1, and for the other segmentation methods 4 out of 5 correlations were significant. In Hospital 2, for all segmentation methods only signal vessel and contrast showed significant correlations. From literature correlations of 14 IQ metrics were checked with diagnostic confidence Likert scores, of which Hospital 1 and 2 had only 1 and 2 significant correlations. Highest correlations for Hospital 1 were contrast method 3 (adjusted $R^2 = 0.45$) and signal vessel method 1 (adjusted $R^2 = 0.42$) and for

Hospital 2 signal vessel method 3 (adjusted $R^2 = 0.25$) and signal vessel method 4 (adjusted $R^2 = 0.22$). These correlations can be seen in Figure 3.3, where it is visible that a clear relation is found between signal vessel and contrast with diagnostic confidence Likert scores. However, it is also clear that only a few low Likert scores are found and that some patients with a low Likert score have similar values of automated IQ metrics as patients with a higher Likert score.

A significant difference between insufficient and sufficient IQ for signal vessel and contrast metrics was found in Hospital 1 for all segmentation methods, but in Hospital 2 only for signal vessel of segmentation method 3 and 4 and contrast of segmentation method 4 (column 3 and 5 of Table 3.1). For both hospitals only 1 literature based IQ metric had a statistically significant difference. Other metrics were on itself not able to statistically discriminate between sufficient and insufficient IQ.

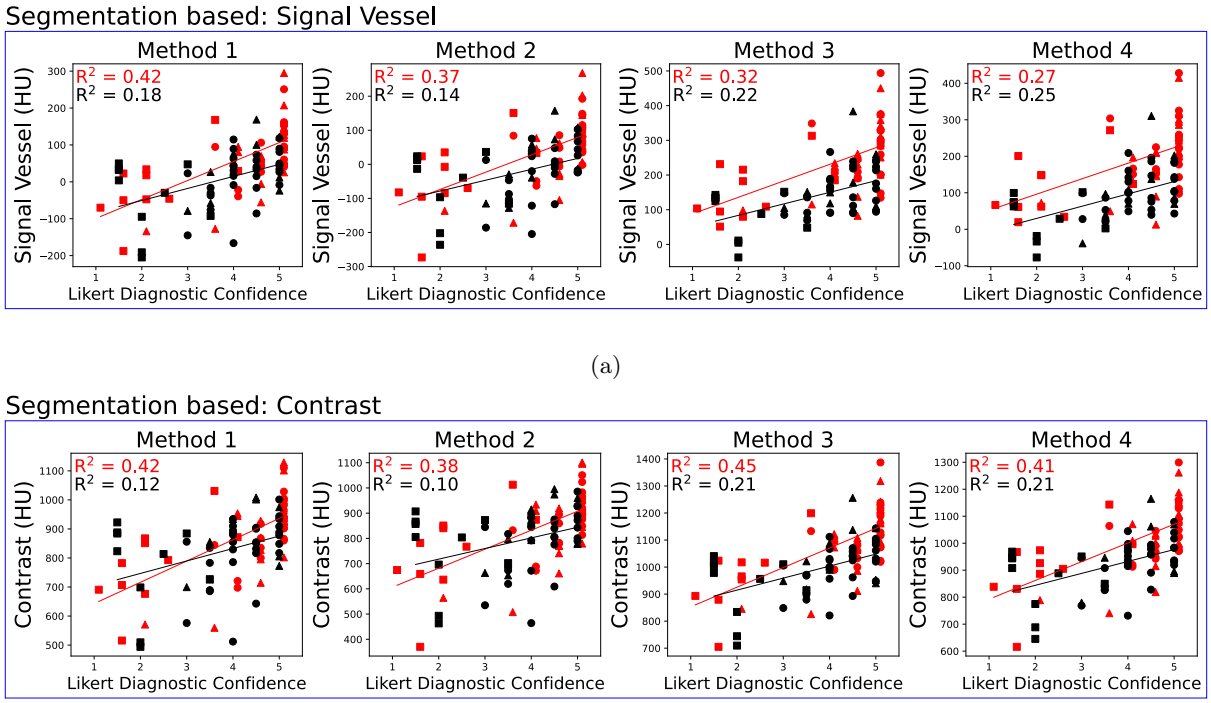


Figure 3.3: Signal vessel (a) and contrast (b) plotted against diagnostic confidence Likert scores for all 4 segmentation methods. For all significant correlations a linear fit with adjusted R^2 is included. Hospital 1 is in red and Hospital 2 in black. Circles represent normal scans, squares repeated scans and triangles FBP scans. For better visibility the x-axis of Hospital 1 has been shifted to the right by 0.1.

Table 3.1: Table with results for all IQ metrics related to diagnostic confidence Likert scores for both hospitals. The correlations (adjusted R^2) and the p-values between sufficient versus insufficient diagnostic confidence are given. Non-significant correlations or p-values are indicated with 'ns'.

	Hospital 1		Hospital 2	
IQ metric	R^2 linear correlation	p-value sufficient vs insufficient	R^2 linear correlation	p-value sufficient vs insufficient
Segmentation based IQ:				
Noise method 1	0.07	ns	ns	ns
Noise method 2	ns	ns	ns	ns
Noise method 3	0.08	ns	ns	ns
Noise method 4	0.09	ns	ns	ns
SNR method 1	0.14	ns	ns	ns
SNR method 2	0.31	0.01	ns	ns
SNR method 3	0.13	0.05	ns	ns
SNR method 4	0.20	0.02	ns	ns
Signal vessel method 1	0.42	<0.001	0.18	ns
Signal vessel method 2	0.37	<0.001	0.14	ns
Signal vessel method 3	0.32	<0.001	0.22	0.01
Signal vessel method 4	0.27	<0.001	0.25	0.007
Contrast method 1	0.42	<0.001	0.12	ns
Contrast method 2	0.38	<0.001	0.10	ns
Contrast method 3	0.45	<0.001	0.21	ns
Contrast method 4	0.41	<0.001	0.21	0.04
CNR method 1	0.10	ns	ns	ns
CNR method 2	0.06	ns	ns	ns
CNR method 3	ns	ns	ns	ns
CNR method 4	ns	ns	ns	ns
Literature based IQ:				
Kortesniemi IQ	ns	ns	0.11	ns
GNL Mode	ns	ns	ns	ns
GNL Median	0.17	0.004	ns	ns
GNL Mode Air	ns	ns	0.09	0.04
GNL Median Air	ns	ns	ns	ns
Noise Background	ns	ns	ns	ns
Noise Trachea	ns	ns	ns	ns
Noise Aorta	ns	ns	ns	ns
SNR Background	ns	ns	ns	ns
SNR Trachea	ns	ns	ns	ns
SNR Aorta	ns	ns	ns	ns
Signal Aorta	ns	ns	ns	ns
Contrast Aorta Trachea	ns	ns	ns	ns
CNR Aorta Trachea	ns	ns	ns	ns

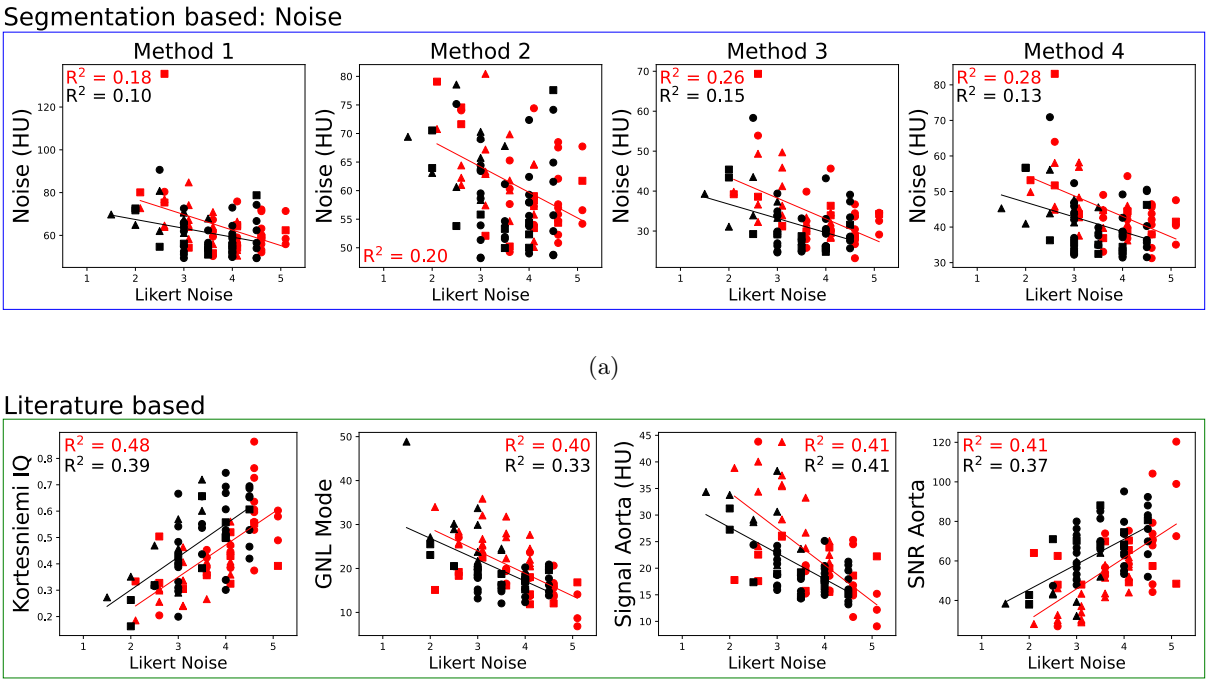
Results of linear regression with backward elimination can be found in Table 3.2. In all cases the contrast metric was excluded due to high collinearity with at least one of the other IQ metrics. For each dataset and segmentation method different IQ metrics were included in the final model, but the vessel signal is included in all cases. For both hospitals the regression model of segmentation method 2 (U-net + thresholding) has highest correlation (adjusted $R^2 = 0.48/0.33$).

Table 3.2: Results of linear regression with backwards elimination for all segmentation based IQ metrics for all 4 methods. Here 'Col' means excluded due to collinearity, 'Excl' means excluded during backwards elimination and 'Incl' means included in the final model.

	Hospital 1				Hospital 2			
Segmentation method	1	2	3	4	1	2	3	4
Noise	Excl	Excl	**Incl**	**Incl**	Excl	Excl	Excl	Excl
SNR	**Incl**	**Incl**	**Incl**	**Incl**	Excl	Excl	Excl	Excl
Signal vessel	**Incl**	**Incl**	**Incl**	**Incl**	**Incl**	**Incl**	**Incl**	**Incl**
Contrast	Col	Col	Col	Col	Col	Col	Col	Col
CNR	Excl	**Incl**	Excl	**Incl**	**Incl**	**Incl**	**Incl**	Excl
Model adjusted R^2	0.45	**0.48**	0.45	0.47	0.27	**0.33**	0.26	0.25

3.4 Noise based IQ results against noise Likert scores

Correlations were found between noise related IQ metrics and noise Likert scores (column 2 and 4 of Table E.1). For segmentation based IQ metrics, two automatic IQ metrics were checked per segmentation method (noise and SNR) and for most methods only the noise metric correlated with noise Likert scores in both hospitals. In Figure 3.4a the noise IQ metric is plotted against noise Likert scores per segmentation method and some correlation is found. For both Hospital 1 and 2 highest correlations were found for noise with segmentation method 3 (adjusted $R^2 = 0.26/0.15$) and 4 (adjusted $R^2 = 0.28/0.13$). From literature 11 noise based IQ metrics were checked, and for Hospital 1 and 2 respectively 9 and 7 of them had significant correlation. For literature methods higher correlations were found compared to the segmentation based IQ metrics, and for both hospitals they were highest for Kortesniemi IQ (adjusted $R^2 = 0.48/0.39$), GNL Mode (adjusted $R^2 = 0.40/0.33$), Noise Aorta (adjusted $R^2 = 0.41/0.41$) and SNR Aorta (adjusted $R^2 = 0.41/0.37$). In Figure 3.4b these highest correlating IQ metrics from literature are plotted against noise Likert scores, and it shows clearly that for both hospitals these correlations are higher compared to the segmentation based noise metrics. Also, it appears that Hospital 1 on average has slightly higher noise than Hospital 2 has.
For both hospitals noise IQ metrics from all segmentation methods had a statistically significant difference between insufficient and sufficient noise, whereas for all segmentation methods SNR was not able to statistically discriminate between sufficient and insufficient noise (column 3 and 5 of Table E.1). For Hospital 1 and 2 respectively 5 and 7 literature based noise related IQ metrics had a statistically significant difference.



(a)

(b)

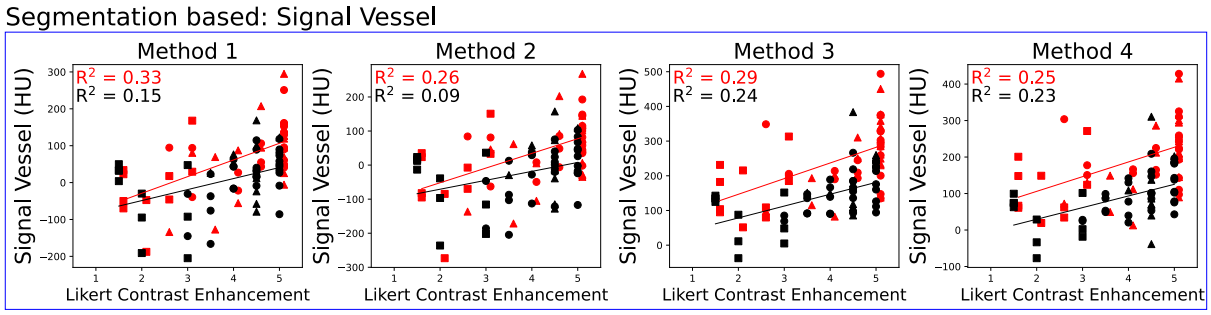
Figure 3.4: In (a) for the segmentation based IQ noise is plotted against noise Likert scores for all 4 segmentation methods. In (b) for the literature based IQ metrics both Kortesniemi IQ, GNL Mode, Noise Aorta and SNR Aorta are plotted against noise Likert scores. For all significant correlations a linear fit with adjusted R^2 is included. Hospital 1 is in red and Hospital 2 in black. Circles represent normal scans, squares represent repeated scans and triangles represent FBP scans. For better visibility the x-axis of Hospital 1 has been shifted to the right by 0.1.

3.5 Contrast based IQ results against contrast enhancement Likert scores

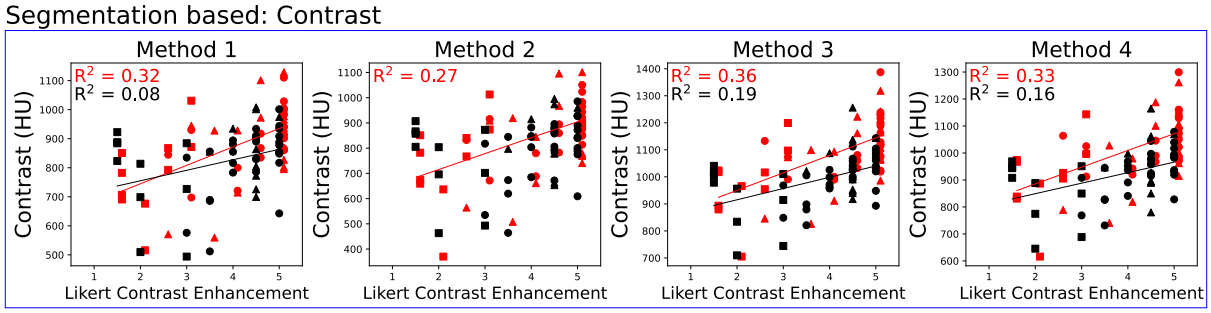
For contrast enhancement related IQ metrics per segmentation method three IQ metrics were checked (signal vessel, contrast and CNR), giving a total of 12 possible correlations. For both hospitals all signal vessel and contrast metrics had significant correlation with contrast enhancement Likert scores, except for contrast method 2 of Hospital 2 (column 2 and 4 of Table E.2). CNR only had a significant correlation for segmentation method 3 in Hospital 1. In Figure 3.5 the signal vessel and contrast are plotted against contrast enhancement Likert scores for all segmentation methods. Clear correlations can be seen, and it appears that on average the contrast enhancement of Hospital 1 is slightly higher than that of Hospital 2.

From literature 3 contrast enhancement related IQ metrics were checked. However, none of them gave significant correlations with contrast enhancement Likert scores.

For Hospital 1 and 2 respectively 8 and 3 segmentation based contrast related IQ metrics had significant difference between insufficient and sufficient contrast enhancement, and for both hospitals no literature based contrast related IQ metrics had significant differences (column 3 and 5 of Table E.2), indicating that they were not able to statistically discriminate between sufficient and insufficient contrast enhancement.



(a)



(b)

Figure 3.5: Signal vessel (a) and contrast (b) plotted against contrast enhancement Likert scores for all 4 segmentation methods. For all significant correlations a linear fit with adjusted R^2 is included. Hospital 1 is in red and Hospital 2 in black. Circles represent normal scans, squares repeated scans and triangles FBP scans. For better visibility the x-axis of Hospital 1 has been shifted to the right by 0.1.

3.6 Influence of pathology

In Table 3.3 the statistical analysis between patients with and without pathology is presented (see also Figure E.8 - E.11). It shows that for both hospitals 10 out of the 20 segmentation based IQ metrics have significant difference between patients with and without pathology. For Hospital 1 segmentation method 2 and 4 had lowest number of metrics with a significant difference (N = 2) and for Hospital 2 segmentation methods 1 and 2 had lowest number of metrics with a significant difference (N = 1). For Hospital 1 none of the Likert scores had significant difference between both groups, however for Hospital 2 significant differences were found for both diagnostic confidence Likert scores (p-value = 0.01) and contrast enhancement Likert scores (p-value = 0.03).

Table 3.3: Table of influence of pathology for segmentation based IQ metrics for each segmentation method per hospitals. The p-value is given for significant differences, and 'ns' if the p-value is not significant.

	Hospital 1				Hospital 2			
Segmentation method	1	2	3	4	1	2	3	4
Noise	0.005	0.003	ns	ns	0.006	0.004	0.01	0.004
SNR	0.03	ns	0.03	ns	ns	ns	0.04	0.03
Signal vessel	0.04	0.05	0.007	0.006	ns	ns	0.01	0.06
Contrast	ns	ns	ns	ns	ns	ns	ns	ns
CNR	ns	ns	0.002	0.003	ns	ns	0.0009	0.04

4 Discussion

In this study automatically determined IQ metrics based on noise and contrast were compared with subjective IQ scored by radiologists for two datasets from different centers. Important result is that automatically determined IQ metrics can be calculated that correlate with radiologist scoring, particularly vessel signal and contrast correlate with diagnostic confidence.

Because it was expected that image noise and contrast enhancement in the pulmonary vessels were not correlated, radiologists were asked to score IQ separately for both, and for the diagnostic confidence for detecting PE. Results did agree with this, showing that it was useful to ask for these three different Likert scores.

Correlations with diagnostic confidence Likert scores were highest for contrast related IQ metrics. This can be explained by the fact that for both hospitals the contrast enhancement Likert scores had much higher correlation with the diagnostic confidence Likert scores compared to the noise Likert scores with diagnostic confidence Likert scores. This seems to indicate that for PE diagnosis with sufficient diagnostic confidence, it is more important to have good contrast enhancement than it is to have low noise. This is also observed for repeated scans, as they were repeated due to insufficient contrast timing. Contrast related IQ metrics from literature (Reeves et al. [14]) did not have significant correlations with diagnostic confidence Likert scores. However, they did not focus on contrast for PE scans. The aorta does not seem the right place to measure signal and contrast for PE scans, as it is located further in the blood circulatory system from the place of contrast injection than the pulmonary vessels. Therefore, when the contrast peak reaches the pulmonary vessels it has not reached the aorta yet, and when it reaches the aorta the peak is no longer located in the pulmonary vessels. This means that segmentations of pulmonary vessels are required for automatically determining IQ for PE scans.

Noise related IQ metrics had lower correlation with diagnostic confidence Likert scores, but significant correlations were found with noise Likert scores. Correlations were highest for the noise related IQ metrics from literature, mostly for Kortesniemi IQ, GNL Mode and the noise and SNR in the aorta. The latter two do require segmentations in the aorta, whereas the Kortesniemi IQ and GNL Mode can be easily calculated without segmentations. Therefore, the Kortesniemi IQ and GNL Mode seem the best IQ metrics for automatically obtaining an objective measure for noise, better compared to the metrics based on segmentations. Radiologists were asked to score noise over the whole image. If they focused more on soft tissue regions this may explain why correlations for Kortesniemi IQ and GNL Mode had higher correlations, as both are determined over the soft tissue regions. On average, noise was found to be slightly higher in Hopsital 1 (see Figure E.12), which is likely due to differences in scanning protocol and different methods of automatic dose control that Siemens and Philips use. This agrees with the dose given to patients, which was on average lower in Hospital 1. Also, in Hospital 1 a slice thickness of 1 mm was used and in Hospital 2 the slice thickness was 2 mm, and noise increases for smaller slice thickness.

For the IQ metrics related with the contrast enhancement Likert scores very similar results were found as related to diagnostic confidence Likert scores, because contrast enhancement Likert scores had very high correlation with diagnostic confidence Likert scores. Correlations were highest for the signal vessel and contrast IQ metrics, which means that segmentations of the pulmonary vessels is needed for automatically calculating an objective measure for contrast enhancement in the pulmonary vessels. On average it was found that the contrast enhancement in Hospital 1 was higher. This is likely due to the differences that were found in the type, volume and flow of the contrast injection, as in Hospital 1 volume and flow rate is adjusted per patient based on weight and kV and in Hospital 2 fixed parameters are used for all patients. These are influential factors for the amount of contrast enhancement. Also, differences in techniques of determining scan delay for intravenous contrast administration, as Hospital 1 used a test bolus and Hospital 2 used bolus tracking.
Mostly, when significant correlations were found between automatically determined objective IQ and subjective IQ, the differences between sufficient and insufficient were also significant. However, this was not always the case.

For both hospitals the best performing segmentation method related to diagnostic confidence Likert scores was U-net + thresholding (adjusted $R^2 = 0.48/0.33$) according to linear regression with backwards elimination. This is remarkable, since this method did not have highest correlation in general

with the IQ metrics individually. Visually differences between the four segmentation metrics could be seen. Segmentation method 1 did sometimes include regions outside of the lungs, mostly consisting of the trachea. Also, method 3 and 4 had smaller and more homogeneous segmented areas compared to method 1 and 2, which made them also include less pathology. These differences may have caused the differences in IQ results and correlations.

For both hospitals 10 out of 20 IQ metrics had a significant difference between the groups of patients with and without pathology, particularly noise and signal vessel. For both hospitals the contrast metric did not have significant difference for any segmentation method. Overall, this seems to indicate that pathology does influence the results of the segmentation based IQ metrics. However, for Hospital 2 diagnostic confidence Likert scores and the contrast enhancement Likert scores did also have a significant difference between both groups, meaning that the patients with and without pathology were not equally distributed over the subjective IQ scores, which could explain some of the significant differences in objective IQ metrics. Significant differences in signal vessel might also be partially due to patients with significant PEs, as they have a blood clot with lower HU present in the vessels.

In literature, automated IQ metrics on images were shown to be promising [8–14]. The study of Franck et al. showed correlation between Kortesniemi IQ and mean visual grading analysis for thorax CT-scans in human cadavers [10] and the study of Brauer et al. demonstrated correlation with Kortesniemi IQ and GNL metrics with subjective IQ in abdomen CT scans on both an anthropomorphic phantom and patients [13]. Results of this study do agree with these findings. Additionally, this study showed correlations between contrast based IQ metrics calculated in pulmonary vessel and lung segmentations with subjective IQ in PE CT-scans, which to our knowledge has not yet been investigated in other studies. Contrast based IQ metrics in pulmonary vessels are found to be needed in order to automatically determine IQ in PE CT-scans.

This study does have some limitations. First of all, the golden standard used in this study consists of Likert scores, which is a subjective score and may vary per radiologist. To partially overcome this issue, two radiologists were asked to score each dataset and the average is taken. A comparing method is a more consistent subjective scoring method compared to Likert scoring, as is demonstrated by Hoeijmakers et al. [22]. However, with this method it is not possible to include multiple categorical scores as we needed for our study.
It was found that for most IQ metrics correlation was higher for Hospital 1 compared to Hospital 2. A possible explanation for this could be that the radiologists of Hospital 1 were more in agreement with each other compared to the radiologists of Hospital 2, as could be seen in Figure 3.2. This is likely because in Hospital 1 the radiologist and last year radiology resident were educated in the same hospital and are more frequently involved in research projects where images are scored using Likert scores. In Hospital 2 the radiologists were educated in different hospitals. Another explanation for the higher correlations in Hospital 1 compared to Hospital 2 could be differences in the datasets that were collected. Despite these differences the results were consistent between both hospitals: overall best and least performing IQ metrics were similar.
With including repeated and FBP scans we did manage to include lower IQ scans in our dataset. However, the dataset was still imbalanced between insufficient and sufficient IQ, which possibly induced a bias. In addition, apparently FBP scans were often scored with a higher IQ by radiologists than repeated scans were.
Also, in presence of pathology, segmentations were not always very accurate. Sometimes pathologies, such as lesions and effusions, were included in the resulting lung and vessel masks, meaning they were also included in the segmentation based IQ metrics. The segmentation models were especially unable to distinguish between pulmonary vessel and pathology. Also, in method 3 and 4 the segmentation clusters and thresholds are determined automatically and these can be influenced by the presence of pathology, which can directly influence the results of automated IQ metrics. In the study of Gang Nam et al. [23] a deep learning–based automatic pulmonary vessel segmentation algorithm was developed for noncontrast chest CT and found successful segmentations for both healthy and diseased lungs. A similar segmentation method could therefore solve these problems and make the segmentation based IQ metrics work well for all patients.
Additionally, a limitation is in the statistical analysis of the influence of pathology. In this technique the correlation between subjective and objective IQ is not taken into account. This could mean that the results are less accurate, especially when patients with and without pathology are unequally distributed

over the Likert scores. Therefore, using the current method the conclusion on whether pathology is an influential factor or not is to be considered as preliminary.[1]
Lastly, a limitation is that the linear regression with backward selection model has first removed all contrast metrics due to collinearity. This means that the finding that the U-net + thresholding method is the best segmentation method for PE scans related to diagnostic confidence, did not take into account the contrast IQ metric, which on its own had significant correlations with diagnostic confidence Likert scores and seems to be important for diagnostic confidence.

5 Conclusion

It is feasible to determine objective IQ metrics that correlate with subjective IQ, based on an automatic algorithm in PE CT-scans. The automated IQ metrics can be calculated for all patients on different CT-scanners and for different scanner settings. Correlations between noise and contrast IQ metrics were found with subjective IQ. Also, noise and contrast IQ metrics were able to differentiate between sufficient and insufficient IQ. For the diagnostic confidence of PE scans contrast IQ metrics are most important, for which segmentations of pulmonary vessels is required. The best way to make these segmentations is using a U-net and thresholding. Segmentations and consequently results seem to be less accurate for patients with pathology in the lungs. For automatically determining noise in PE CT-scans, segmentations are not required, as methods from literature that calculated noise metrics over mainly the soft tissue parts of the scan gave highest correlation with subjective noise image quality. Overall, this opens possibilities for continuous image quality monitoring for pulmonary embolism scans in clinical practice, which could also be used for scanning protocol optimization.

[1] To tackle this limitation another statistical analysis that does take the correlation between objective and subjective IQ into account is explored in Appendix F. However, this needs further exploration with a statistician.

References

[1] J. A. Kline, J. S. Garrett, E. J. Sarmiento, C. C. Strachan, and D. M. Courtney, "Over-testing for suspected pulmonary embolism in American emergency departments: The continuing epidemic," *Circulation: Cardiovascular Quality and Outcomes*, 2020.

[2] B. Hendriks, C. Jeukens, H. Gietema, J. Wildberger, and B. Martens, "Five-year experience in individualised CT pulmonary angiography: the incidence and causes for repeat scans," 2022, submitted.

[3] S. Rose, B. Viggiano, R. Bour, C. Bartels, J. P. Kanne, and T. P. Szczykutowicz, "Applying a New CT Quality Metric in Radiology: How CT Pulmonary Angiography Repeat Rates Compare Across Institutions," *Journal of the American College of Radiology*, vol. 18, pp. 962–968, 7 2021.

[4] A. M. Cornea, B. J. McCullough, L. M. Mitsumori, and M. L. Gunn, "Enhancement of the pulmonary arteries and thoracic aorta: comparison of a biphasic contrast injection and fixed delay protocol with a monophasic injection and a timing bolus protocol," *Emergency Radiology*, vol. 22, pp. 231–237, 6 2015.

[5] T. Niemann, I. Zbinden, H. W. Roser, J. Bremerich, M. Remy-Jardin, and G. Bongartz, "Computed tomography for pulmonary embolism: assessment of a 1-year cohort and estimated cancer risk associated with diagnostic irradiation,"

[6] N. Al-Zaher, F. Vitali, M. F. Neurath, and R. S. Goertz, "The Positive Rate of Pulmonary Embolism by CT Pulmonary Angiography Is High in an Emergency Department, Even in Low-Risk or Young Patients," *Medical Principles and Practice*, vol. 30, pp. 37–44, 2 2021.

[7] M. Kuroki, M. Nishino, T. Masaya, Y. Mori, V. Raptopoulos, P. Boiselle, S. Tamura, and H. Hatabu, "Incidence of Pulmonary Embolism in Younger Versus Older Patients Using CT," *Journal of thoracic imaging*, vol. 21, pp. 167–171, 11 2006.

[8] O. Christianson, J. Winslow, D. P. Frush, and E. Samei, "Automated technique to measure noise in clinical CT examinations," *American Journal of Roentgenology*, vol. 205, pp. W93–W99, 7 2015.

[9] M. Kortesniemi, Y. Schenkel, and E. Salli, "Automatic image quality quantification and mapping with an edge-preserving mask-filtering algorithm," *Acta Radiologica*, vol. 49, no. 1, pp. 45–55, 2008.

[10] C. Franck, A. De Crop, B. De Roo, P. Smeets, M. Vergauwen, T. Dewaele, M. Van Borsel, E. Achten, T. Van Hoof, and K. Bacher, "Evaluation of automatic image quality assessment in chest CT – A human cadaver study," *Physica Medica*, vol. 36, pp. 32–37, 4 2017.

[11] J. Sanders, L. Hurwitz, and E. Samei, "Patient-specific quantification of image quality: An automated method for measuring spatial resolution in clinical CT images," *Medical Physics*, vol. 43, pp. 5330–5338, 10 2016.

[12] A. Malkus and T. P. Szczykutowicz, "A method to extract image noise level from patient images in CT," *Medical Physics*, vol. 44, pp. 2173–2184, 6 2017.

[13] M. Brauer, C. Jeukens, C. Mihl, E. Laupman, J. Wildberger, B. Martens, and C. van Pul, "Relationship between automatically determined image quality metrics and subjective image quality in abdominal CT," 2022, submitted.

[14] A. P. Reeves, Y. Xie, and S. Liu, "Automated image quality assessment for chest CT scans," *Medical Physics*, vol. 45, pp. 561–578, 2 2018.

[15] A. Moore, J. Wachsmann, M. Chamarthy, L. Panjikaran, Y. Tanabe, and P. Rajiah, "Imaging of acute pulmonary embolism: An update," *Cardiovascular Drugs and Therapy*, vol. 8, 12 2017.

[16] A. Mansoor, U. Bagci, B. Foster, Z. Xu, G. Z. Papadakis, L. R. Folio, J. K. Udupa, and D. J. Mollura, "Segmentation and Image Analysis of Abnormal Lungs at CT: Current Approaches, Challenges, and Future Trends 1," 1148.

[17] J. Hofmanninger, F. Prayer, J. Pan, S. Röhrich, H. Prosch, and G. Langs, "Automatic lung segmentation in routine imaging is primarily a data diversity problem, not a methodology problem," *European Radiology Experimental*, vol. 4, 12 2020.

[18] E. M. Van Rikxoort and B. Van Ginneken, "Automated segmentation of pulmonary structures in thoracic computed tomography scans: a review," *Physics in Medicine & Biology*, vol. 58, p. R187, 8 2013.

[19] B. Hendriks, N. Eijsvoogel, M. Kok, B. Martens, J. Wildberger, and M. Das, "Optimizing pulmonary embolism computed tomography in the age of individualized medicine: A prospective clinical study," *Investigative Radiology*, vol. 53, p. 1, 02 2018.

[20] N. Otsu, "A Threshold Selection Method from Gray-Level Histograms," *IEEE Trans Syst Man Cybern*, vol. SMC-9, no. 1, pp. 62–66, 1979.

[21] P.-S. Liao, T.-S. Chen, and P.-C. Chung, "A fast algorithm for multilevel thresholding.," *J. Inf. Sci. Eng.*, vol. 17, pp. 713–727, 09 2001.

[22] E. Hoeijmakers et al., "Improving the subjective scoring of clinical CT images using a pairwise comparison method instead of a Likert scoring [Unpublished manuscript],"

[23] J. Gang Nam, J. Nathanael Witanto, S. Joon Park, S. Jin Yoo, J. Mo Goo, and S. Ho Yoon, "Automatic pulmonary vessel segmentation on noncontrast chest CT: deep learning algorithm developed using spatiotemporally matched virtual noncontrast images and low-keV contrast-enhanced vessel maps,"

[24] A. Oppelt, *Imaging Systems for Medical Diagnostics*. Publicis Corporate Publishing, Erlangen, 2005.

[25] T. M. Buzug, *Computed Tomography: From Photon Statistics to Modern Cone-Beam CT*. Springer, 2008.

[26] H. O. Tekin and `[U+FFFD]` Kara, "Analysis of filtering material and its effect on x-ray features by using computational method for medical imaging applications," *RAD Conference Proceedings*, vol. 1, pp. 133–135, 6 2016.

[27] "Basics of X-ray." `https://miac.unibas.ch/PMI/01-BasicsOfXray.html#`. Accessed October 27, 2022.

[28] "How a CT works." `https://ams.com/how-a-ct-works`. Accessed October 27, 2022.

[29] "Filtered back projection." `http://www.impactscan.org/slides/impactday/basicct/sld016.htm`. Accessed August 22, 2022.

[30] W. Stiller, "Basics of iterative reconstruction methods in computed tomography: A vendor-independent overview," *European Journal of Radiology*, vol. 109, pp. 147–154, 12 2018.

[31] M. Söderberg and M. Gunnarsson, "Automatic exposure control in computed tomography an evaluation of systems from different manufacturers," *Acta radiologica (Stockholm, Sweden : 1987)*, vol. 51, pp. 625–34, 07 2010.

[32] R. Clarke et al., "The 2007 recommendations of the international commission on radiological protection. icrp publication 103.," *Annals of the ICRP*, vol. 37, pp. 1–332, 5 2007.

[33] C. Lee, "How to estimate effective dose for CT patients," *European Radiology*, vol. 30, pp. 1825–1827, 2020.

[34] W. H. Kamr, A. M. El-Tantawy, M. M. Harraz, and A. I. Tawfik, "Pulmonary embolism: Low dose contrast MSCT pulmonary angiography with modified test bolus technique," *European Journal of Radiology Open*, vol. 7, 1 2020.

[35] M. Moradi and B. Khalili, "Qualitative indices and enhancement rate of CT pulmonary angiography in patients with suspected pulmonary embolism: Comparison between test bolus and bolus-tracking methods," *Advanced Biomedical Research*, vol. 5, no. 1, p. 113, 2016.

[36] T. Suckling, T. Smith, and W. Reed, "A retrospective comparison of smart prep and test bolus multidetector CT pulmonary angiography protocols," *Journal of Medical Radiation Sciences*, vol. 60, pp. 53–57, 6 2013.

[37] "CT Pulmonary Angiography." `https://introductiontoradiology.net/courses/rad/CTPA/05misc/technique.html`. Accessed August 23, 2022.

[38] W. Ertel, *Introduction to Artificial Intelligence*. Springer, 2017.

[39] C. C. Aggarwal, *Neural Networks and Deep Learning*. Springer, 2018.

[40] J. T. Page, Z. S. Liechty, M. D. Huynh, and J. A. Udall, "BamBam: Genome sequence analysis tools for biologists," *BMC Research Notes*, vol. 7, no. 1, 2014.

[41] J. Liao, Y. Wang, J. Yin, L. Liu, S. Zhang, and D. Zhu, "Segmentation of rice seedlings using the YCrCB color space and an improved Otsu method," *Agronomy*, vol. 8, 11 2018.

[42] Z. Meng, Y. Hu, and C. Ancey, "Using a Data Driven Approach to Predict Waves Generated by Gravity Driven Mass Flows," *Water 2020, Vol. 12, Page 600*, vol. 12, p. 600, 2 2020.

[43] H. Kim and Y. S. Jeong, "Sentiment Classification Using Convolutional Neural Networks," *Applied Sciences 2019, Vol. 9, Page 2347*, vol. 9, p. 2347, 6 2019.

[44] H. Wang and B. Raj, "On the Origin of Deep Learning On the Origin of Deep Learning," *undefined*, 2017.

[45] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9351, pp. 234–241, 2015.

[46] M. Z. Alom, M. Hasan, C. Yakopcic, T. Taha, and V. Asari, "Recurrent Residual Convolutional Neural Network based on U-Net (R2U-Net) for Medical Image Segmentation," 8 2018.

[47] "The phantom laboratory, catphan® manual." `https://www.phantomlab.com/s/catphan412_424manual.pdf`. Accessed November 21, 2022.

[48] A. F. Hayes, *Introduction to Mediation, Moderation, and Conditional Process Analysis: A Regression-Based Approach*. A Division of Guilford Publications, 2018.

A Theory

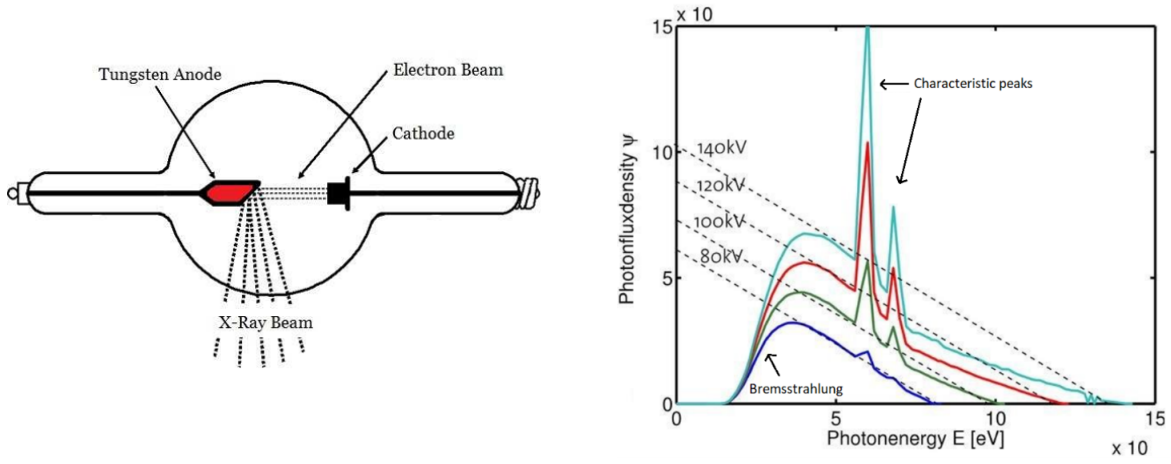
In this section, the theory behind the different methods used in this project are briefly explained. For the scope of this thesis a brief explanation of CT theory is considered sufficient. For more extensive explanation we refer to books [24, 25]. In A.1 the basics behind CT scanning are described, including a brief summary of the typical CT scanning methods for PE diagnosis. In A.2 the theory behind the artificial intelligence (AI) techniques used for automated segmentation is described.

A.1 Computed tomography scanning

Computed Tomography (CT) scanning is a medical imaging technique that uses X-rays to obtain high quality images of the inside of a human body.

A.1.1 X-rays

X-rays are generated in an X-ray tube (Figure A.1a), which consists of an anode and a cathode. When a cathode is heated, electrons are released and accelerated towards the anode due to a potential difference between the cathode and the anode. The electrons will interact with the anode, which is usually made of tungsten, resulting in characteristic X-rays and Bremsstrahlung. In Figure A.1b a spectrum of X-rays can be seen for a tube operated at different tube voltages.
Characteristic X-rays are emitted when an electron fills the vacancy after an inner shell electron is ejected due to hard collisions, which occur when the distance between the atom and the charged particle is small. Bremsstrahlung is emitted when the free electron gets close to the nucleus to which it is attracted due to its strong electric field. This causes the electron to lose energy, which is converted to radiation called Bremsstrahlung. The energy of the photon depends on the incoming orientation and the electron's speed.



(a) (b)

Figure A.1: In (a) a schematic drawing of an X-ray tube can be seen [26]. In (b) a spectrum of X-rays is given for a tube operated at 80, 100, 120 and 140 kV [27].

Afterwards, the X-rays will leave the tube via a window and are used for imaging. X-rays interact with matter, dominated by Compton scattering and the photoelectric effect for diagnostic X-ray imaging.
When an incident photon scatters from an atom and an outer shell electron is ejected it is called Compton scattering. The scattered photons increase the random noise and decrease the image contrast.
The photoelectric effect is the process where a photon is absorbed by an atom resulting in one of the orbital electrons to be ejected. The energy of the incident photon is then transferred to the atomic electron. These interactions can only occur if the energy of the absorbed photon is larger than the binding energy of the ejected electron and they can occur with electrons in the K, L, M or N shells.
Due to the above mentioned interactions, the X-ray beam attenuates as it travels through a patient according to the Lambert-Beer law,

$$I = I_0 \cdot e^{-\mu \cdot l},$$

with I the X-ray beam intensity, I_0 the initial X-ray beam intensity, μ the linear attenuation coefficient of the relevant tissue and l the distance the beam has travelled through that tissue.

The interaction of X-rays with tissues can lead to tissue damage. The ionizing radiation can interact with atoms of a DNA molecule, which can lead to DNA damage or mutations. These mutations can lead to cancer later in life.

A.1.2 Detector

Opposite to the X-ray tube a detector array is placed (Figure A.2), which consists of a scintillator and a photodiode. After interaction with the patient an X-ray will reach the detector and in the scintillating layer it will be absorbed, after which electrons will reach higher energy states. Photons are released as the electrons relax back to their ground state and these will be detected by the photodiode.
The X-ray tube and detector rotate around the patient. Modern scanners contain multiple parallel rows of detectors to be able to scan multiple slices at the same time, and are therefore called multislice computed tomography scanners.

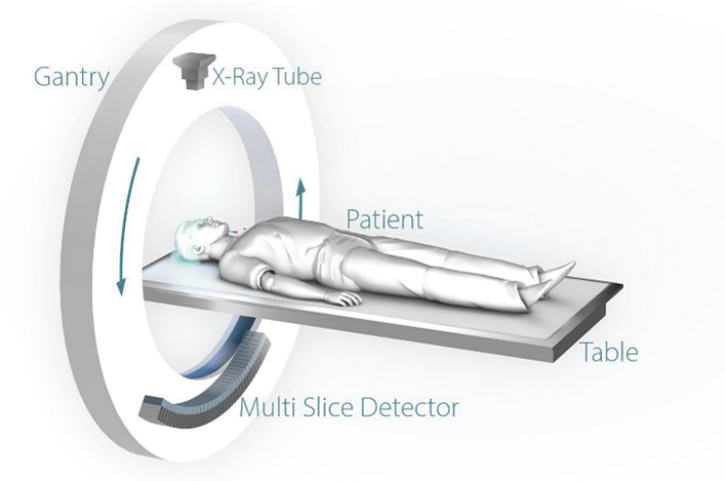


Figure A.2: Schematic overview of the main components of a CT scanner, including the gantry, X-ray tube, detector array and the table [28].

A.1.3 Image reconstruction

Different tissue types can be distinguished in X-ray images as they absorb X-rays differently, expressed in the linear attenuation coefficient. The attenuation property is expressed in Hounsfield Units (HU), defined as

$$1000 \cdot \frac{\mu - \mu_{water}}{\mu_{water}},$$

with μ_{water} the linear attenuation coefficient of water and μ the linear attenuation coefficient of the specific structure.
In the detector elements, the X-rays that travelled through a part of the body are collected for that particular projection. To turn all information from the rotating detector into an image of a cross section of the body, different reconstruction techniques can be used.

Filtered back projection (FBP)
From certain angles the detector will measure the attenuation profile, called projections. When a large number of projections are acquired, the attenuation profiles can be combined to form an image, called back projection. Since there is only a discrete sample of angles produced images will be blurry. FBP compensates for this by applying a filter or kernel before back projection. This kernel has direct influence on image noise and spatial resolution. Several kernels are used in clinical practice, depending on the clinical task. In Figure A.3 an example of FBP is shown of an object for a different number of projections together with a result of conventional back projection, which has significant lower spatial resolution. Note that in this example a parallel beam is used instead of the fan shaped beam that is normally used in CT scanners.

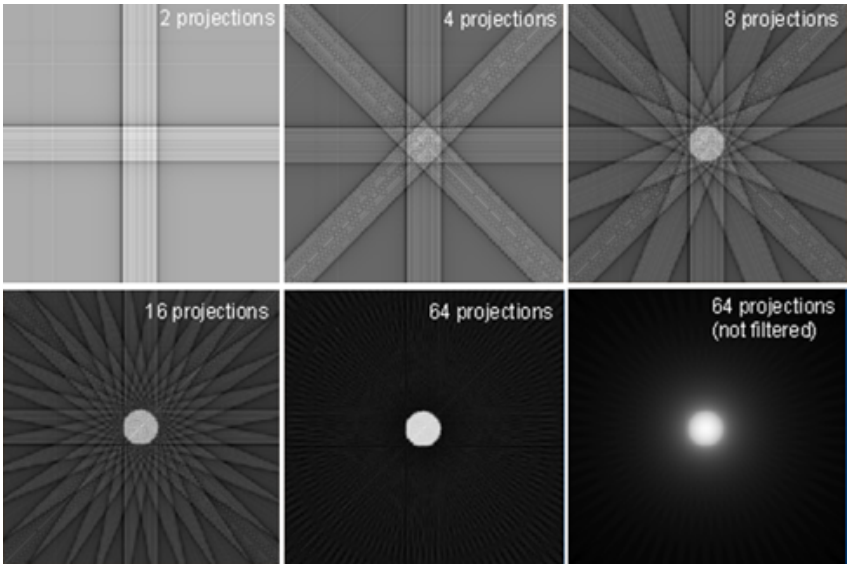


Figure A.3: FBP applied on an object for different numbers of projections, ranging from 2 to 64. On the right bottom also a conventional back projection result is given for 64 back projections. Adapted from [29].

Iterative reconstruction
Currently iterative reconstruction methods are mainly used in clinical practice, due to its ability to overcome noise associated with filtered back projection without increasing dose. The iterative reconstruction process contains a cycle of forward and back projection steps to iteratively improve reconstructed image data. In general it consists of the following steps [30]:

1. Based on a primary image, e.g. an FBP image, synthesized projections are simulated by forward projection.

2. Synthesized projection are compared with measured projections and from their difference a correction term is calculated.

3. Image estimate is updated by back projection of the correction term.

These steps are repeated until the differences in the images reach a preset stopping criterion. Iterative reconstruction is computationally more expensive, and each manufacturer has its own model. Examples of commercial models are ASIR (GE healthcare), VEO (GE healthcare), iDose (Philips), IMR (Philips), SAFIRE (Siemens) and ADMIRE (Siemens).

A.1.4 Scanner parameters

Next to the reconstruction method, a user can choose between multiple CT scanner settings before scanning a patient, like the pitch, the exposure and kVp.
While the gantry is rotating around the patient, the table will move through the gantry. The pitch is the displacement of the table in one gantry rotation divided by the total thickness of all simultaneously acquired slices. A lower pitch increases image quality (IQ) but it also increases patient dose.
The exposure (in mAs) is defined as the product of tube current and rotation time. With a higher exposure more electrons are released, which will increase the signal and results in higher IQ. However, increased exposure also increases patient dose.
Automatic exposure control (AEC) systems are added to CT scanners to adapt the exposure to the patient's shape, size and attenuation. The methods differ per manufacturer [31]. Generally the signal is kept constant by adjusting the exposure.

Radiation dose output is often expressed in the CT dose index, which is measured in mGy. $CTDI_W$, which is the weighted average of dose across a single slice, is defined as:

$$CTDI_W = \frac{2}{3}CTDI_{100,periphery} + \frac{1}{3}CTDI_{100,center},$$

where $CTDI_{100}$ is a linear measure of dose distribution over a 100 mm long ionization chamber and $CTDI_{100,periphery}$ and $CTDI_{100,center}$ are the $CTDI_{100}$ values at the surface and center respectively. The CTDI volume is commonly used to express the CT dose output and is defined as

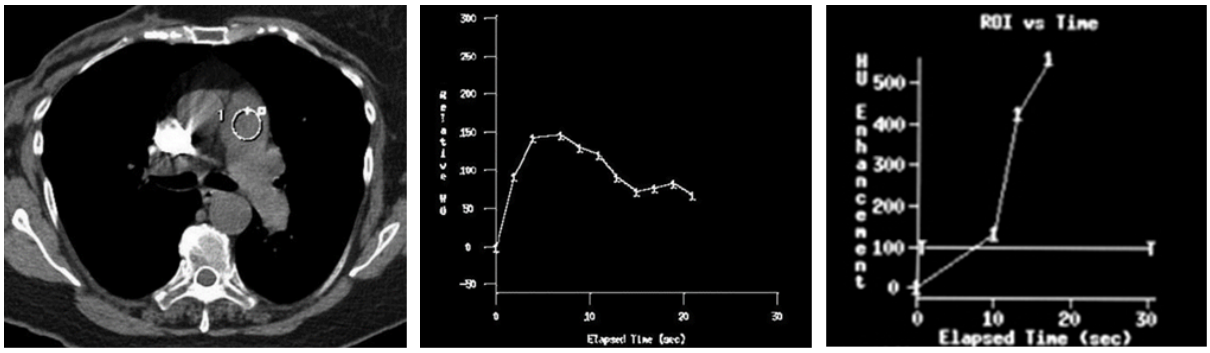
$$CTDI_{vol} = \frac{CTDI_W}{P},$$

where P is the pitch as described above.

This CTDI is just a scan protocol characteristic and can only be related to patient effective dose by more extensive calculations including area. This is beyond the scope of this thesis, and for more information about this we refer to the following papers [32, 33].

A.1.5 CT pulmonary angiogram (CTPA)

A pulmonary embolism (PE) is a blood clot that blocks the blood flow in a lung artery. PE is diagnosed using CTPA, which uses contrast enhancement in the pulmonary vessels to be able to distinguish the potential blood clot from pulmonary vessels. Timing of the intravenous contrast material administration is challenging. Test bolus and bolus tracking are the two most used techniques for determining the scan delay.
For the test bolus method some low dose axial images are taken at the level of the main pulmonary artery at a few moments in time. Manually an ROI is placed on the pulmonary artery (Figure A.4a), after which a small amount of contrast is injected. The delay time between contrast injection and the contrast peak reaching the pulmonary artery is now calculated (Figure A.4b). After this initial test bolus the actual PE scanning will start, where the time between contrast injection and scanning is based on the calculated delay time.
The bolus tracking technique also starts with some low dose axial images taken at the level of the main pulmonary artery at fixed moments, where an ROI is manually placed on the pulmonary trunk. The intravenous contrast medium is then injected, and after the Hounsfield unit in the pulmonary artery reaches a certain preset threshold, e.g. 100 HU (see Figure A.4c), the PE scan will start. [34, 35]
When a patient has certain symptoms, like shortness of breath or chest pain, they can be expected to have PE. These symptoms can also be caused by other abnormalities in the lungs, and therefore often other types of pathology are found in PE scans. In the PE datasets of this study mostly pleural effusions, pneumonia, emphysema or lung noduli were found next to PE. A pleural effusion is a build-up of excess fluid between the thin membranes that separates the lungs and chest cavity (pleura). A small amount of fluid present in the pleura is normal. Pneumonia is an infection that inflames the air sacs in the lungs, which are then filled with fluids and pus. Emphysema is a type of chronic obstructive pulmonary disease. Patients with emphysema have damage to the walls of the air sacs (alveoli) in the lungs, which makes it difficult to breath. Lung noduli are small round or oval-shaped growth with diameters up to 3 cm. Lung noduli are typically benign but they can also be malignant.



(a) (b) (c)

Figure A.4: In (a) an axial slice of a thorax CT scan is shown, where an ROI is placed on the pulmonary artery. In (b) for the test bolus method the HU is plotted against time, where the contrast peak can be seen. In (c) the HU is plotted against time for the bolus tracking method, where the scanning starts when the pulmonary artery reaches the threshold of 100 HU. Adapted from [36, 37].

A.2 Artificial Intelligence

Artificial intelligence is a relatively new field of study which can be used to automate processes. It concerns advanced mathematical methods to train a model from large annotated datasets. In this study, we use K-means clustering, Otsu-thresholding and a U-net model. The basics behind these 3 techniques will be summarized in this chapter. For more details on AI and machine learning we refer to books [38,39].

A.2.1 K-means clustering

K-means clustering is an unsupervised learning algorithm and it groups similar observations in a dataset based on their similarity. K-means clustering minimizes within-cluster variances, like squared Euclidean distances, between the clusters. It is applicable for n-dimensions, where n can be any real number and it requires all variables to be continuous.
K-means clustering requires a preset value for the number of clusters K. A number of K clusters centers will be randomly assigned and each data point is assigned to its nearest cluster center. Next the new centroid from the clustered group of points c_i is determined using

$$c_i = \frac{1}{|S_i|} \sum_{x_i \in S_i} x_i,$$

where $X = \{x_1, x_2, x_3, ..., x_n\}$ is the set of datapoints and S_i is the set of all points assigned to the ith cluster. Again the distances between each datapoint and the new cluster centers are calculated and each datapoint is assigned to its nearest cluster center. This will be repeated until no data point is reassigned or if a maximum number of iterations is reached. An example of this process of a 2 cluster set can be seen in Figure A.5.

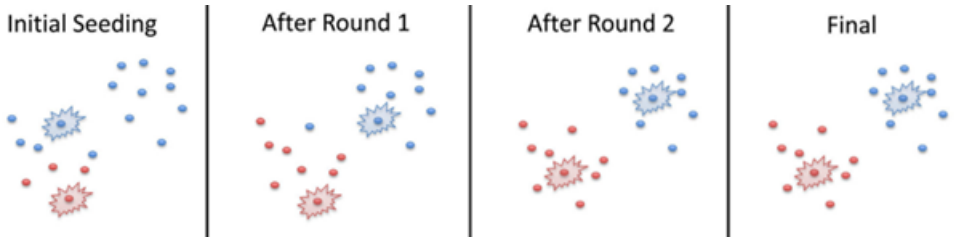


Figure A.5: Example of K-means clustering algorithm with 2 clusters, with the cluster center marked by a starburst. After two rounds the clusters have reached a ready-state [40].

A.2.2 Otsu-thresholding

Otsu-thresholding [20] is an image thresholding technique that iteratively searches for the threshold that minimizes the within-class variance of two classes. The within-class variance σ_w at any threshold t is defined as

$$\sigma_w^2(t) = \omega_0(t)\sigma_0^2(t) + \omega_1(t)\sigma_1^2(t),$$

with ω_0 and $_1$ the probability of the two classes separated by a threshold t and with σ_0^2 and σ_1^2 the variances of these two classes. The Otsu-thresholding algorithm iteratively calculates the within-class variance for all possible thresholds t and the threshold that corresponds to the maximum within-class variance is the resulting Otsu-threshold. In Figure A.6 an example of a histogram with the Otsu-threshold visualized is given. Multi-Otsu thresholding is an adapted version in which multiple thresholds and multiple classes are generated from a single image. This is however computationally more demanding and therefore calculating multiple thresholds will take a longer time.

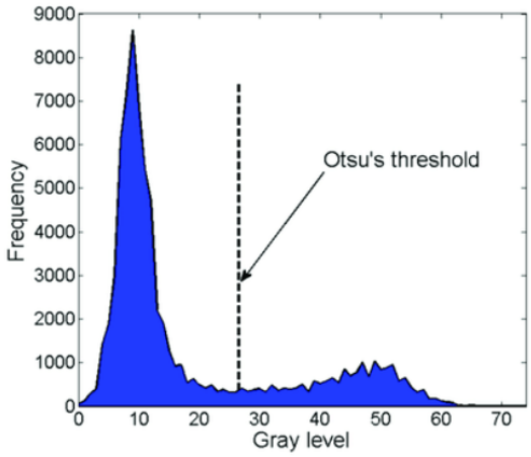
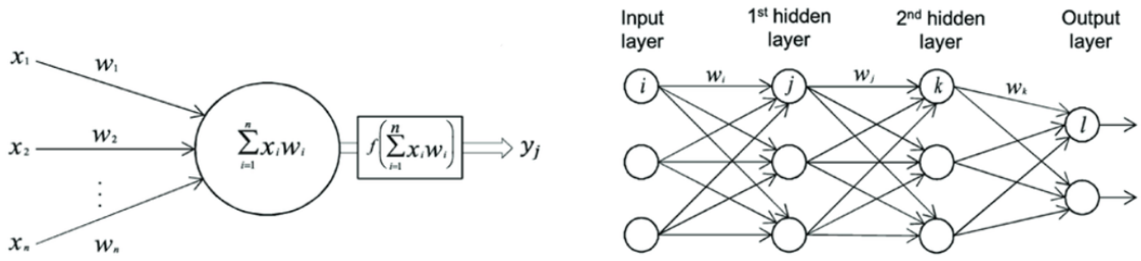


Figure A.6: Example of a histogram where Otsu's threshold is visualized. It can be seen that the histogram is divided into two classes [41].

A.2.3 U-net

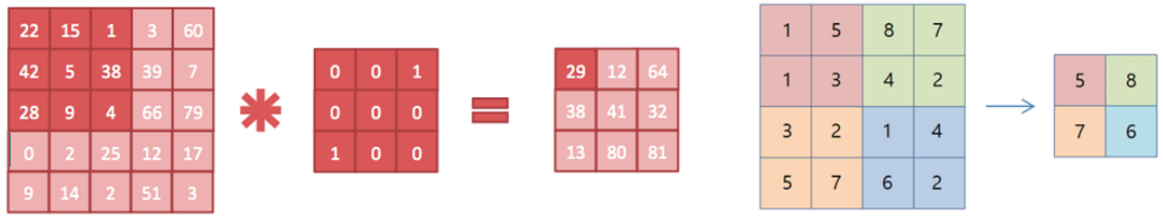
A convolutional neural network is a deep neural network that is often used for images. The dataset will contain a set of images, labeled for a certain outcome that the network should learn to recognize. It consists of fully connected neuron networks, which means that each neuron in one layer is connected to all neurons in the next layer. In a neuron (Figure A.7a) an input vector $\vec{x}$ enters the neuron and is multiplied with a weight vector $\vec{w}$. After this a bias is added and an activation function is applied. The network is used to calculate the relation between input data and output, without the need to define features. It typically consists of one input layer, multiple hidden layers and an output layer, as shown in Figure A.7b.



(a) (b)

Figure A.7: Schematic view of a neuron (a) and a fully connected neural network (b) [42].

Multiple operations take place in a CNN. A convolutional operation is often used to find features of the image. It uses a filter that represents a certain feature which is applied over the whole image. The shape of the output depends on the filters size and the displacement of the filter after each operation. A simplified example of a two dimensional convolution can be seen in Figure A.8a. Another operation that is often used is max pooling, see Figure A.8b. This filter returns the maximum pixel values and thereby reduces the noise in an image, which decreases the image dimensions. This can be prevented by adding pixels with a value of zero around the image, called zero padding. The weights and biases used in the filters during training will be learned by the CNN.



(a) (b)

Figure A.8: Illustration of a two dimensional convolution operation (a) and a max pooling operation (b) [43, 44].

A U-net is a CNN that is developed for medical image segmentation [45]. The architecture is modified which makes it possible to work with fewer training images. The architecture of a U-net is u-shaped, as is also illustrated in Figure A.9, consisting of a contracting and an expansive path. In the contracting path both convolutional and max pooling operations are applied repeatedly, similar to a CNN. This increases the feature information while reducing the spatial information. In the expansive path convolutional and up-convolutional operations are used. Due to the up-convolutional operations the spatial information is increased again. The expansive and contracting paths are also concatenated to help the learning process and the preserve spatial information about the original input.

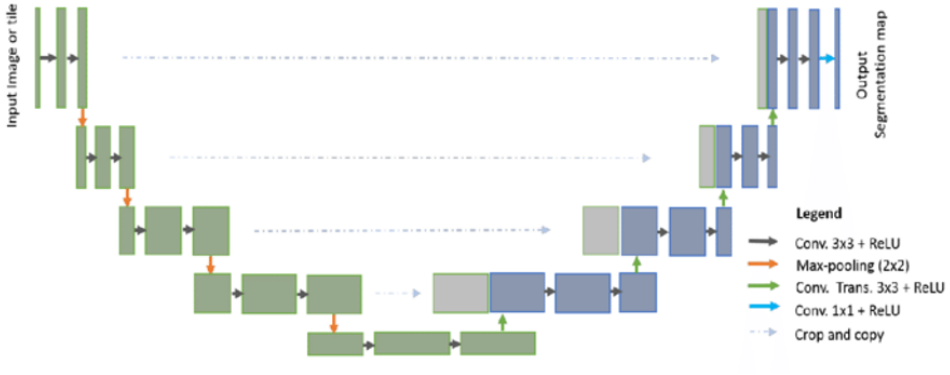


Figure A.9: Example of a U-net architecture. This U-net was used for sentiment classification [46].

After the models architecture is defined a U-net model can be trained. In the training process it should find the optimal weights and biases that give the model a better output. The optimization problem is described by a loss function, which is minimized during the training process. Many different loss functions $J(\theta)$ can be used and in general they are a function of the weights and biases, θ, described as

$$J(\theta) = \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}(y_i, \hat{y}_i),$$

with y_i the ground truth for training example i, $\hat{y}_i$ the model output of the training example i, N the total number of training samples and $\mathcal{L}$ the used loss function.

B Lung noduli study

The first part of this project consisted of a lung noduli study in the Máxima Medical Center with an already existing and available dataset used as a testing set for development of our method. In this appendix this part of the study is described briefly. The methods are similar to the pulmonary embolism study, which is described in more detail in Chapter 2.

B.1 Introduction

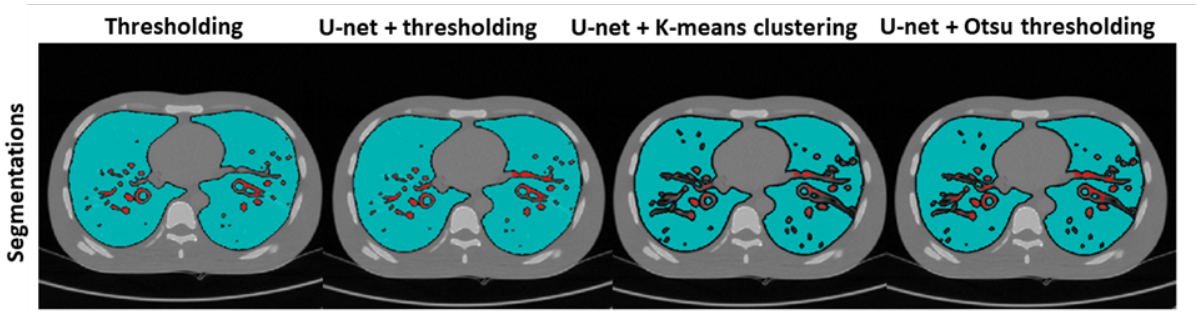
In the Máxima Medical Center Veldhoven various types of scanners are used. Exposure to radiation can cause damage to patients, and therefore it is important to optimize the scan protocols of all scanners according to the ALARA principle: to use the lowest dose that still gives acceptable image quality (IQ) to answer the clinical question. Currently scanning protocols are usually optimized using a physical phantom or radiologist scoring. However, a phantom does not resemble a patient's anatomy and manual scoring by radiologists is time consuming and results are subjective. Also, both methods cannot be used for continuous monitoring of image quality.
The aim of this study is to develop an algorithm that automatically determines noise and contrast-related image quality metrics in thorax CT. This method is used to check image quality and dose relation on two scanners: a GE PET-CT and a Philips iCT. To focus on the tissue of interest, an automatic lung and pulmonary vessel segmentation method is developed.

B.2 Methods

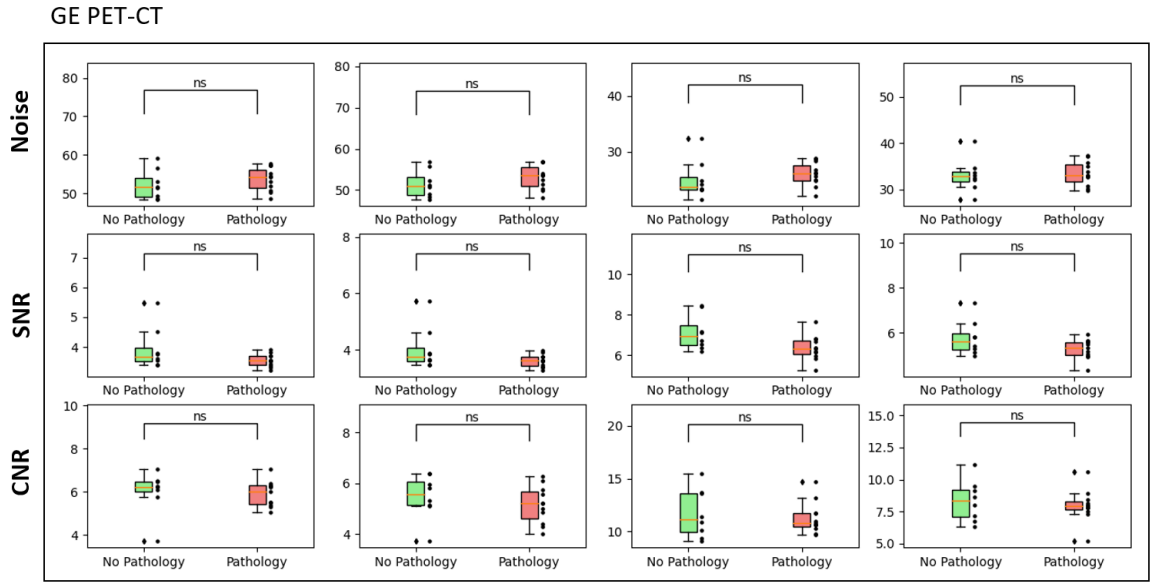
A retrospective dataset was used consisting of 40 patients who had a thorax CT scan for noduli detection. 20 patients were scanned on the GE PET-CT and 20 patients were scanned on the Philips iCT.
Similar as to the pulmonary embolism study, for all patients lungs and pulmonary vessels are segmented with 4 different methods: thresholding, U-net and thresholding, U-net and K-means clustering and U-net and multi-Otsu thresholding. Within the resulting lung masks the noise and signal-to-noise ratio is calculated and the contrast-to-noise ratio (CNR) is calculated between the lung and vessel masks.
Also, 5 different noise related IQ metrics from literature are calculated: Kortesniemi IQ, GNL Mode, GNL Median, GNL Mode Air and GNL Median Air [8, 9, 12].
Additionally, for each patient presence of lung pathology (lesions, emphysema, effusion, etc.) was recorded, as this may influence the automatic segmentations and the calculated IQ scores. To see whether the presence of pathology in the lungs gave significant differences in the resulting IQ scores a Wilcoxon rank sum test was used.

B.3 Results and discussion

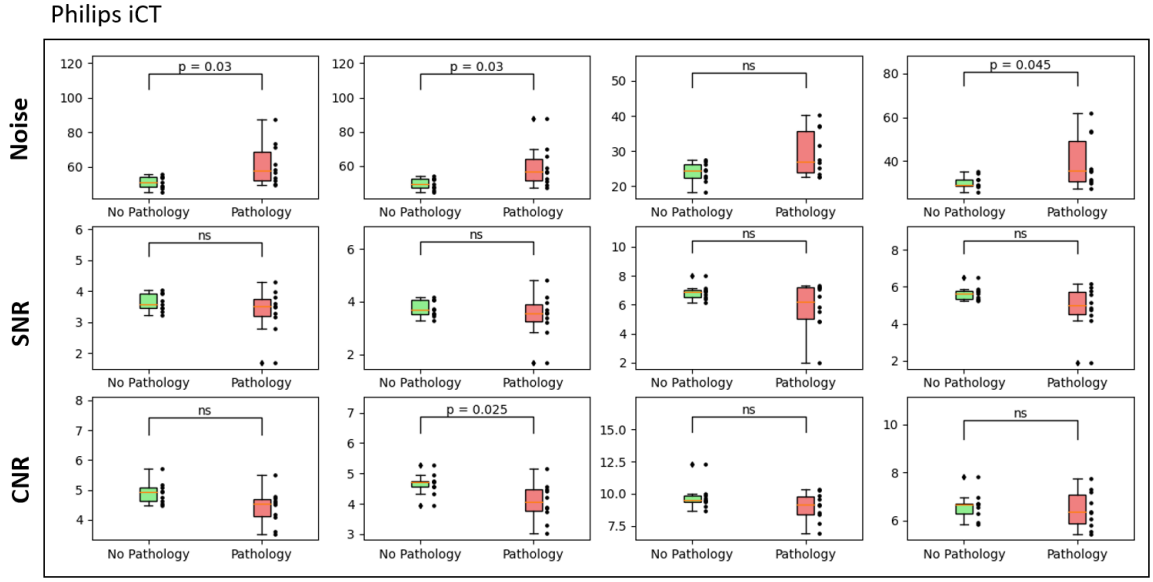
Visual analysis of the segmentations showed that for all patients without pathology in the lungs segmentations were successful, see Figure B.1a. However, for some patients with pathology in the lungs segmentations were less accurate, which sometimes lead to partial inclusion of pathology in the lung or vessel masks.
For patients scanned on the GE PET-CT no significant differences were found between patients with and without pathology present in the lungs, see Figure B.1b.
For patients scanned on the Philips iCT 3 methods (thresholding, U-net + thresholding and U-net + Otsu thresholding) showed significant differences in noise for patients with and without pathology, see Figure B.1c. However, no significant differences were found for SNR and CNR, except the CNR with the method U-net + thresholding.
These results indicate that pathology sometimes is an influential factor, but not in all cases. The fact that results varied between both scanners may be caused by differences in severity of pathology present in the patients included in the datasets.



(a)



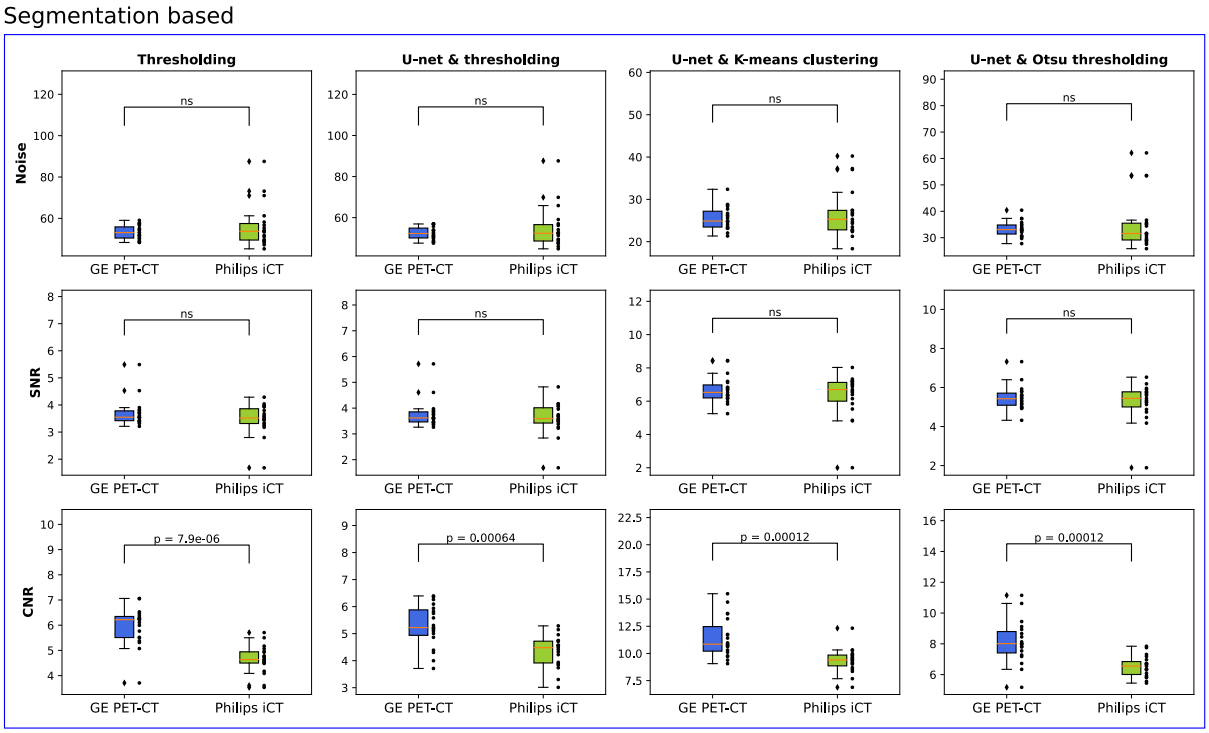
(b)



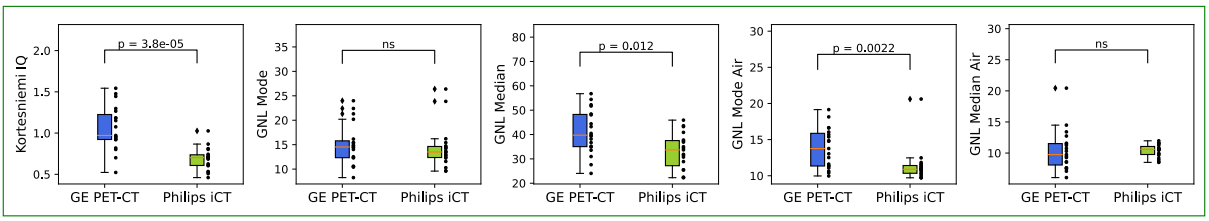
(c)

Figure B.1: In (a) examples of segmentations are given for all 4 methods for an anthropomorphic phantom (PBU-60 phantom, Kyoto Kagaku Co.). Results of noise, SNR and CNR for all methods are given for patients without pathology (green) and patients with pathology (red) scanned on a GE PET-CT in (b) and scanned on a Philips iCT in (c), including the according p-values. For non significant p-values 'ns' is indicated.

Differences in noise and SNR between both scanners were not significant. However, CNR did differ significantly between both scanners for all segmentation methods, see figure B.2a. Here on average the patients scanned on the GE PET-CT had higher CNR.
In Figure B.2b it can be seen that for the noise based IQ metrics from literature Kortesniemi IQ, GNL Median and GNL Median Air significant differences were found between both scanners, where for the Kortesniemi IQ noise was on average lower for the GE PET-CT and for the GNL Median and GNL Median Air noise was on average lower for the Philips iCT.
The differences found are likely reflecting differences in protocol settings. In Figure B.3 the CTDIvol is plotted against mAs for the patients of both datasets, and it can be seen that on average the GE PET-CT used a higher dose which is mainly caused by a few outliers that received higher dose.



(a)



(b)

Figure B.2: In (a) results are given for noise, SNR and CNR for all 4 segmentation methods. Here the two different scanners are compared: the GE PET-CT (blue) and the Philips iCT (green). In (b) the two scanners are compared for the noise related literature based methods. The according p-values are included and for non significant p-values 'ns' is indicated.

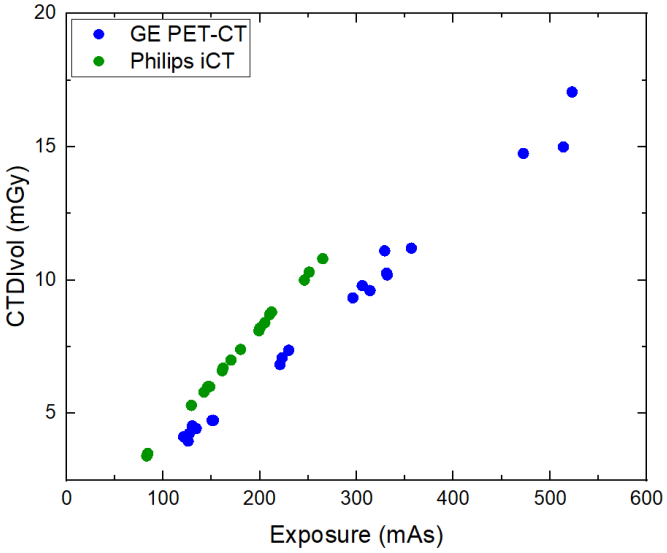


Figure B.3: The CTDIvol plotted against exposure for patients scanned on the GE PET-CT (blue) and Philips iCT (green). All patients were scanned at 100 kV.

B.4 Conclusion

Automatically calculating IQ metrics in thorax CT scans is feasible. Segmentations for patients with pathology in the lungs were less accurate, and this could also be seen for some IQ metrics where IQ results were significantly different between patients with and without pathology. This indicates that the algorithm is not always robust in determining IQ in the presence of pathology.
The algorithm is able to analyze scans from different scanners and is able to demonstrate inter-scanner differences.

C Segmentation: Choices of Model Parameters

For the lungs and pulmonary vessel segmentation models, a number of choices and assumptions were made. In this appendix the reasoning behind these choices will be given.

C.1 Thresholding: Threshold value

In the thresholding methods a fixed threshold is used to segment the lungs and pulmonary vessels. Pixels within the body with a value below this threshold are defined as lungs and pixels with a value above this threshold within the segmented lungs are defined as pulmonary vessel. To choose which threshold value is used, 9 different thresholds have been tested for 7 patients, varying from -900 to -100 HU in steps of 100 HU. Results can be seen in Figure C.1 for both signal, noise, SNR and CNR. A lower threshold means that only a few pixels are defined as lungs and with higher thresholds the lung area becomes larger, until at some point (-100 HU) regions outside the lungs are classified as lungs, see Figure C.2. Based on visual inspection of the segmented lung and pulmomary vessel masks a threshold between -600 and -400 seemed a reasonable choice. Based on this and on Figure C.1, where IQ metrics were most stable at a threshold of -500 HU, a threshold of -500 HU is chosen.

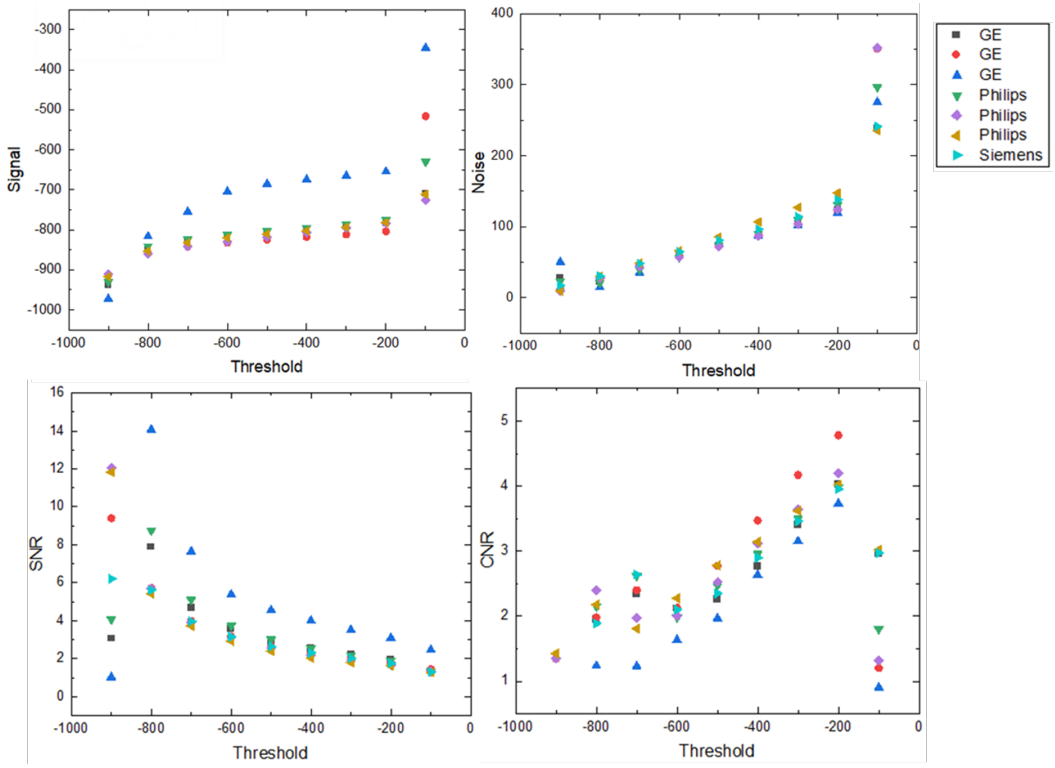


Figure C.1: Signal, noise, SNR and CNR results for the thresholding method for 7 patients against threshold value in HU. 3 patients are scanned on a GE PET-CT, 3 on a Philips iCT and 1 on a Siemens Flash.

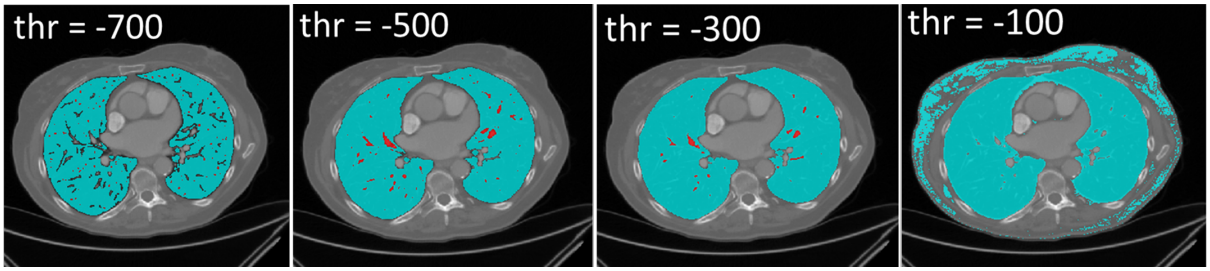


Figure C.2: Segmentations with the thresholding method for one patient using different thresholds: -700, -500, -300 and -100. In blue the lung mask and in red the vessel mask is given.

C.2 K-means clustering: number of clusters

For the method that used K-means clustering the lungs are divided into a preset number of clusters, and the cluster with lowest HU is classified as lungs and the cluster with highest HU is classified as pulmonary vessel. To test what number of clusters can best be used, the model is tested for clusters varying between 2 and 9 for 3 different patients. Visual results for the Kyoto phantom can be seen in Figure C.4. With a lower cluster number more pixels are defined as either lung or vessel and with a higher cluster number a larger area within the lungs is not defined as either of these, which makes the segmented areas more homogeneous. This can also be seen in Figure C.3, where the noise decreases as the cluster number increases. It was also observed that with less clusters more pathology was included in the segmentations compared to more clusters. However, when there are too many clusters, the areas of the lungs and vessels become too small which might no longer give representative results for the patient. Based on visual inspection and analysis of the results, a cluster number of 4 is chosen, as this gave the optimal balance between homogeneous segmented regions with less pathology included without the lung and vessel masks becoming too small.

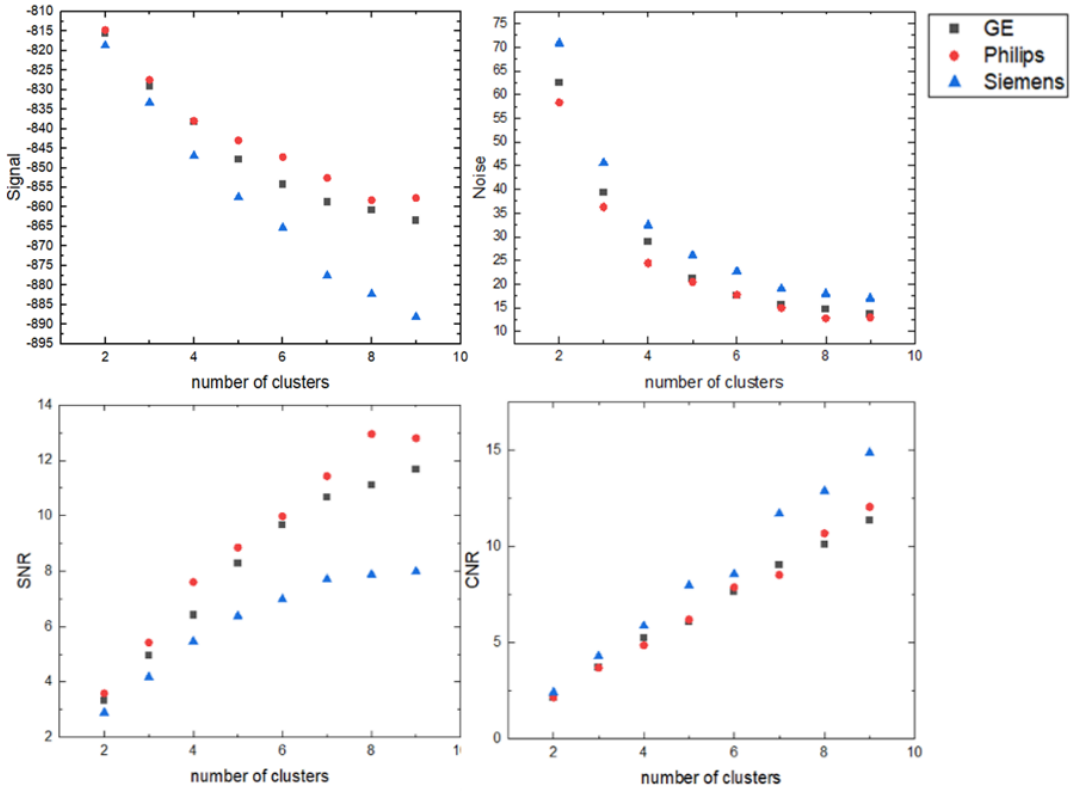


Figure C.3: Signal, noise, SNR and CNR results for the U-net + K-means clustering method for 3 patients against number of clusters. 1 patient is scanned on a GE PET-CT, 1 on a Philips iCT and 1 on a Siemens Flash.

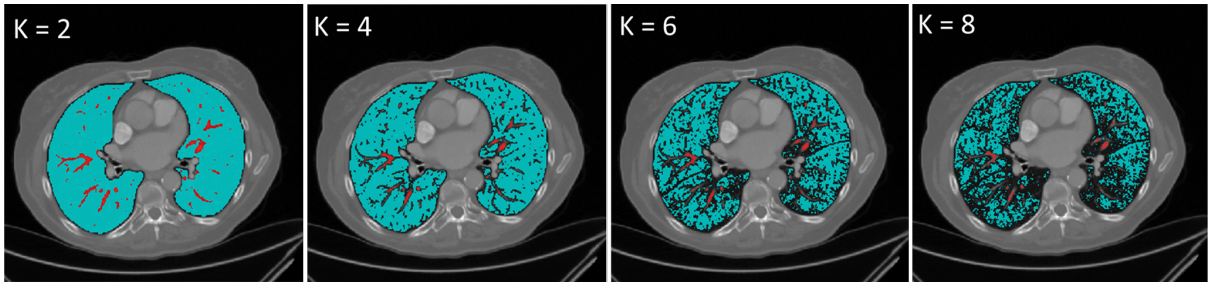


Figure C.4: Segmentations with the U-net + K-means clustering method for one patient using different numbers of clusters: K = 2, 4, 6 and 8. In blue the lung mask and in red the vessel mask is given.

C.3 Erosion and dilation

After the segmentations erosion and dilation were applied. Erosion can be used to remove the edges of segmentations and remove small structures. Erosion in combination with dilation can be used to remove small structures whilst still keeping the edges. To test whether and how much erosion and dilation can best be used, multiple different combinations are tested for 3 different patients scanned on a GE PET-CT, Philips iCT and Siemens Flash and segmented using the thresholding method. For this a disk shape is used with varying sizes, see Figure C.5. Here it can be seen that in the lungs removing a small edge influences the results, but the amount of influence decreases as more edges are removed. This is not the case in the vessels.
In Figure C.6a different sizes of erosion are applied to the lungmask of a patient. Here it can be seen that for larger erosion more edges are removed. In Figure C.6b different sizes of erosion and dilation are applied to the vesselmask of a patient. It can be seen that when erosion of size 1 is applied the vessels become smaller, and they become even smaller with erosion of size 2. When first erosion of size 2 is applied and second dilation of size 1, some edges and a few small vessel structures are removed.
Based on analysis of the graphs and visual inspection of the images, it is chosen to use erosion of disk size 2 for the lungs, since this removes the edges with most influence on IQ results. Applying more erosion did not influence IQ results much. For the pulmonary vessels, an erosion of size 2 and dilation of size 1 is chosen, as this gives an optimal balance between removing some edges and small vessels without making the vessel areas too small.

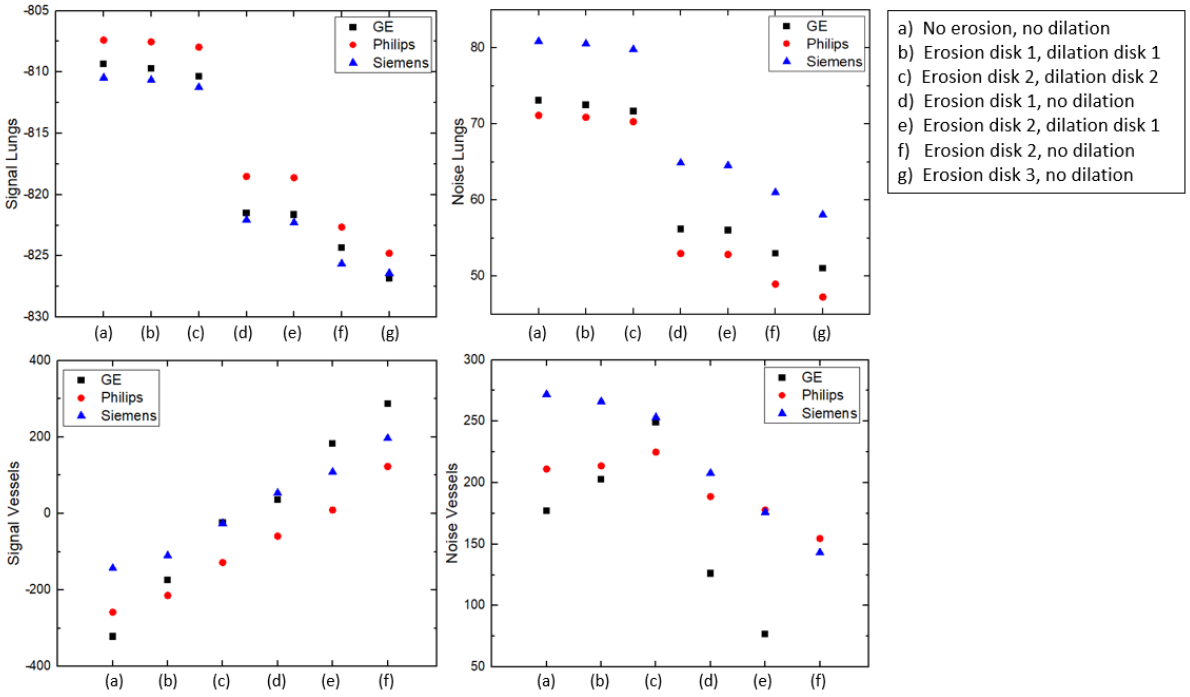
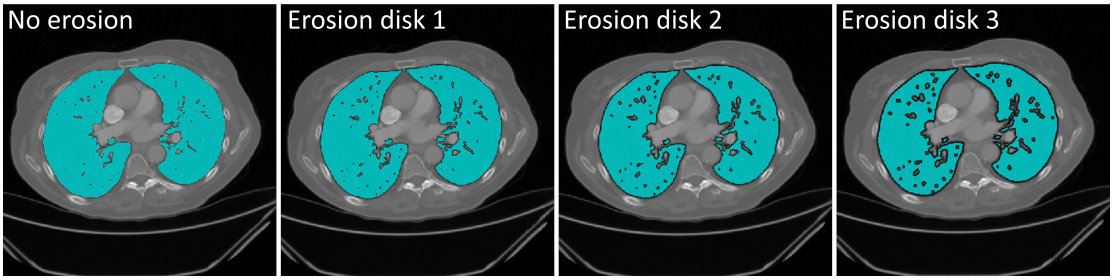
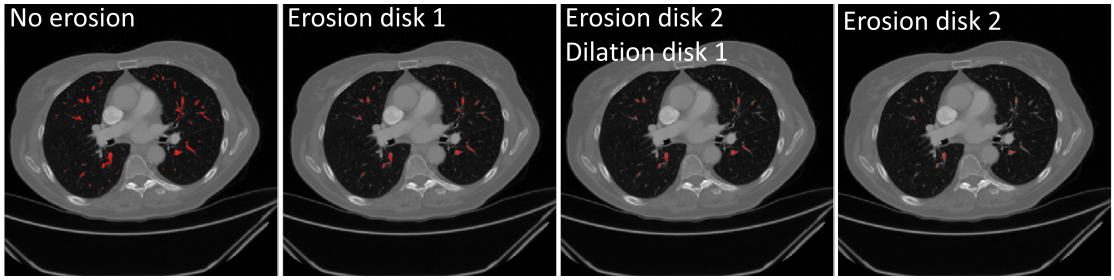


Figure C.5: Signal and noise in lungs and vessel results for the thresholding method for 3 patients against erosion and dilation type. The meaning of these types can be found in the legend. 1 patient is scanned on a GE PET-CT, 1 on a Philips iCT and 1 on a Siemens Flash.



(a)



(b)

Figure C.6: In (a) lung segmenations (in blue) are given for one patient with different disk sizes of erosion (0, 1, 2, 3) applied to it. In (b) vessel segmentations (in red) are given with different disk sizes of erosion (0, 1, 2) and dilation (0,1).

C.4 Maximum vessel area

A problem of the vessel segmentations is that the segmentation models are not always able to distinguish between actual blood vessels and pathology. Often pathology, such as tumors and effusion, are larger in area compared to the typical pulmonary vessel. Therefore a restriction can be added, saying that if a 2D segmented vessel is larger than X, it is too likely that it consists of pathology and is therefore excluded from the vessel mask. Choosing X too small leads to removing a lot of regions that consist of pulmonary vessels, and choosing X too large leads to still including large regions of pathology.
For 3 patients scanned on a Philips iCT and varying types of pathology (no pathology, tumor and effusion), different restriction areas are tested for all segmentation methods, see Figure C.7. It can be seen that for all segmentation methods the vessel signal has reached a stable value for the patient without pathology at a restriction area of around 250 mm^2. The vessel signal of patients with pathology has not yet reached a stable point at 250 mm^2, since there are still parts in the vesselmask with a larger area. These parts consist of pathology, and we want to remove these parts. Therefore it is chosen to use a vessel area restriction of 250 mm^2: at this threshold only a limited amount of vessels will be removed whereas large pathology is removed.

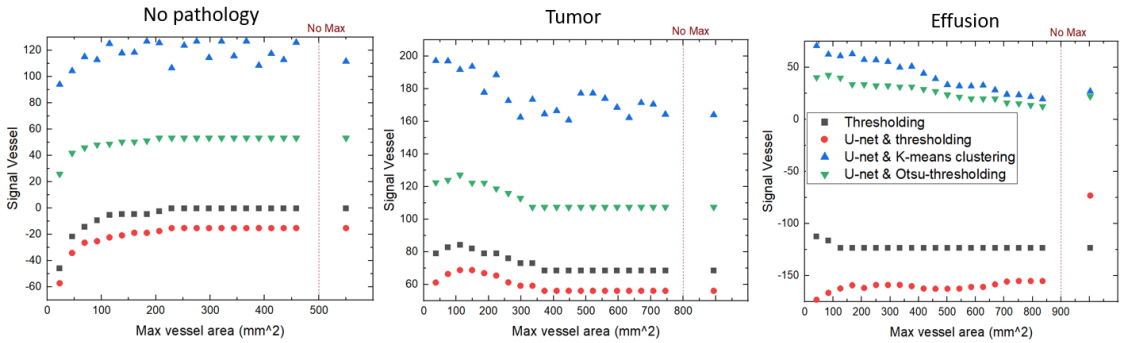


Figure C.7: Signal vessel against the maximum allowed vessel area for all 4 segmentation methods for 3 patients: 1 patient without pathology, 1 patient with a tumor and 1 patient with an effusion (= fluid in lungs). On the right from the red line the results are given for no maximum vessel area. All 3 patients are scanned on a Philips iCT

In Figure C.8 images of the original segmentations without vessel area restriction and the resulting segmentations with the restriction that the vessel area cannot be larger than 250 mm^2 are given for the 3 different patients. Here it can be seen that indeed no vessels are removed but that large pathology is removed.

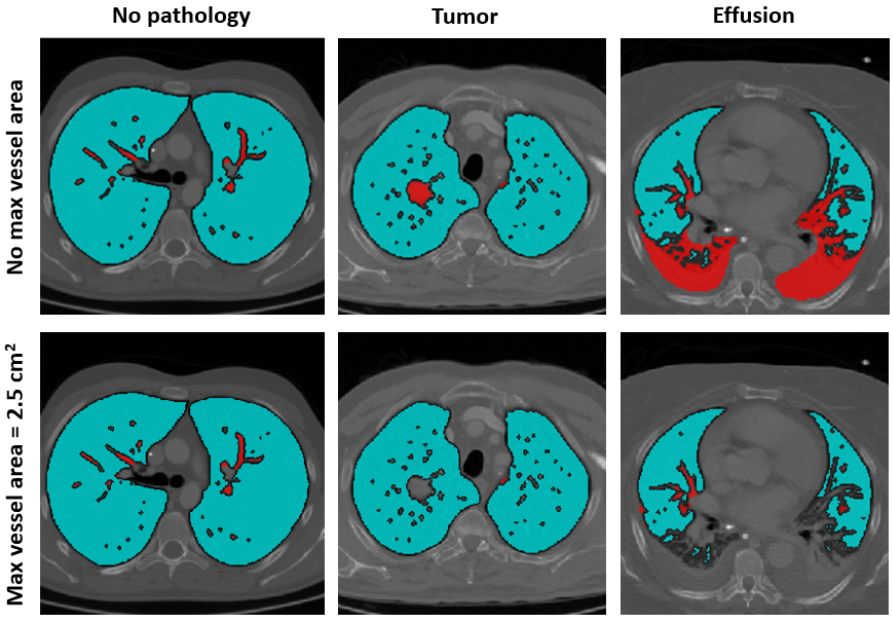


Figure C.8: Example of 3 patients, of which one has no pathology, one has a tumor and one has effusion in the lungs. For these patients segmentations of lungs (blue) and pulmonary vessels (red) are visualized for both no max vessel area (top) and a max vessel area of 2.5 cm^2 (bottom).

D Phantoms study

The outline of our study is shown in Figure D.1. We obtained CT images from 2 different centers (*Maastricht University Medical Center, Maastricht* and *Máxima Medical Center, Veldhoven*), each with a different scanner. The Catphan phantom is used to check physical performance of each scanner, the anthropomorphic phantom to test performance of the metrics for different scanner settings and finally CT scans from patients were included to test the new method. In this chapter the methods and results for the Catphan phantom and the anthropomorphic phantom will be given.

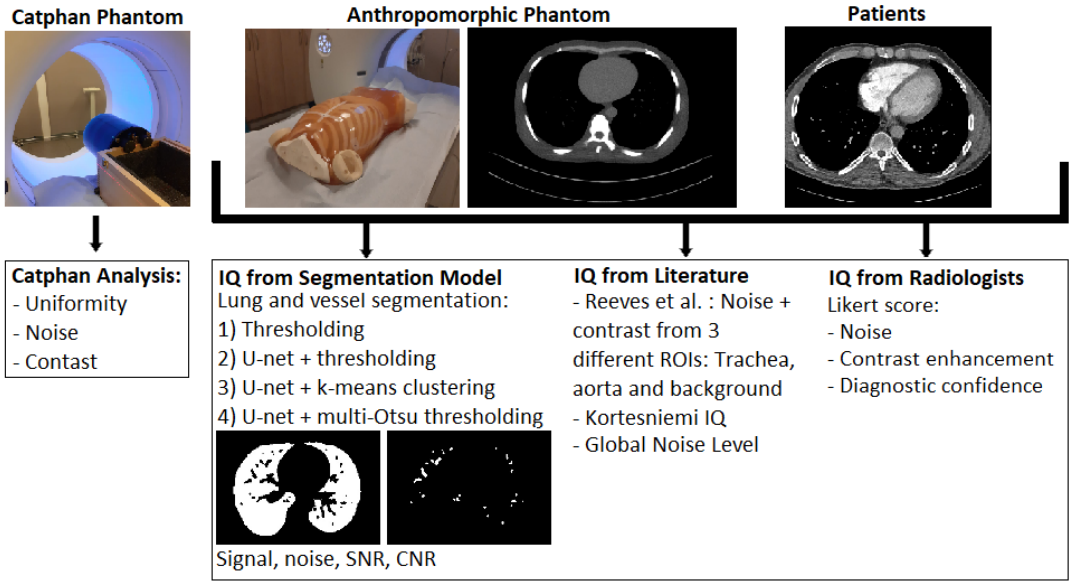


Figure D.1: Pipeline of this study

D.1 Methods

Catphan phantom

In this study scanners from 2 different medical centers were used. In Hospital 1 (*Maastricht University Medical Center, Maastricht*) a Siemens Flash was used and in Hospital 2 (*Máxima Medical Center, Veldhoven*) a Philips iCT and a Philips Ingenuity were used. To compare the different scanners, a Catphan phantom was measured on both scanners (Catphan 412 on the Siemens Flash and Cathan 500 on the Philips iCT and Philips Ingenuity) with both a basic protocol that was kept the same on each scanner and the standard lung embolism protocol that is used on the specific scanner. The scanner settings of the basic protocol can be found in Table D.1. The slice thickness was 3 mm and the pitch was 1. These scans were all reconstructed with FBP and iterative reconstruction (ADMIRE) with strenght 3 with kernel Br40 in Hospital 1 and with FBP and iterative reconstruction (iDose) with strength 3 with kernel B (standard) in Hospital 2. For the lung embolism protocol all standard settings were used that are normally used on patients for PE detection, except for slice thickness and kV and mAs. A slice thickness of 2 mm was used and the dose settings can be found in Table D.2. These scans were all reconstructed with FBP and the standard reconstruction method used in each hospital for PE scans. Noise and SNR are determined in the region of air (HU = -1000) and acrylic (HU = 100), as their HU values are closest to the theoretical HU values of lungs and blood. Additionally, the contrast and CNR between air and acrylic is determined. Lastly, uniformity is checked in both hospitals by placing 5 ROIs at the following locations: center, top, right, bottom and left. In each ROI the mean signal is determined, and the deviations between the mean signal in the peripheral ROIs and the central ROI is determined. A difference of max 5 HU is accepted.
For more details about the Catphan phantoms and measurements we refer to the manual [47].

Table D.1: Scanner settings of basic protocol of Catphan scans.

Number	kV	mAs
1	80	250
2	100	250
3	120	250
4	140	250

Table D.2: Scanner settings of the PE protocol of Catphan scans.

Number	kV	mAs
1	100	150
2	100	238
3	120	150
4	120	250

Anthropomorphic phantom

In both hospitals an anthropomorphic phantom (PBU-60 phantom, Kyoto Kagaku Co., Ltd., Kyoto, Japan) was scanned to check whether IQ also varied for different scanner settings. The phantom has been scanned on a Siemens Flash (Hospital 1) with the standard lung embolism protocol and on a Philips Ingenuity (Hospital 2) with the standard thorax protocol. For each scanner scans were made with varying tube current, tube voltages and reconstruction algorithms. On both scanners the tube currents ranged from 10 to 180 mAs and the tube voltages were either 80, 100 or 120 kVp. Additionally, for the Siemens Flash the reconstruction algorithms filtered back projection (FBP) and iterative reconstruction (ADMIRE) with strength 3 and 5 and with kernel I26f and a slice thickness of 1 mm were applied. For the Philips Ingenuity the reconstruction algorithms FBP and iterative reconstruction (iDose) with strength 3 and 5 and with kernel B (standard) and a slice thickness of 3 mm were applied.
Since the anthropomorphic phantom used on the Siemens Flash has previously been used for practicing biopsies, the resulting images contained some holes. After the phantom was scanned the holes in the images have been filled manually so that segmentations could still be made. The regions where the holes were present were not taken into account in the final image quality scores.
Next for all scans the lungs and pulmonary vessels will be segmented using the methods explained in Chapter 2.2.1. From these segmentations the signal, noise and SNR are determined in the lungs and CNR is determined between lungs and vessels. These IQ metrics are plotted against CTDIvol, to check the relation between IQ and dose.

D.2 Results and discussion

Catphan phantom

In Figure D.2 the results can be found for the Catphan measurements with the basic protocol. It is found that the noise in air is different between both scanners. The noise in air is higher for Hospital 1 and does not change much when kVp increases. This is also reflected in the SNR air and the CNR air acrylic. Other than that the results of both scanners look quite similar.
In Figure D.3 the results for the LE measurements on the Catphan phantom are compared between both hospitals. Here more differences in the results can be seen, which are likely due to their different protocols and reconstruction techniques. However, the y-axis of these plots have only a small range, which means that the found differences are in truth not that large.
For all Catphan scans uniformity was checked and none of the deviations were higher than 5 HU, which means that uniformity was accepted for all scans.

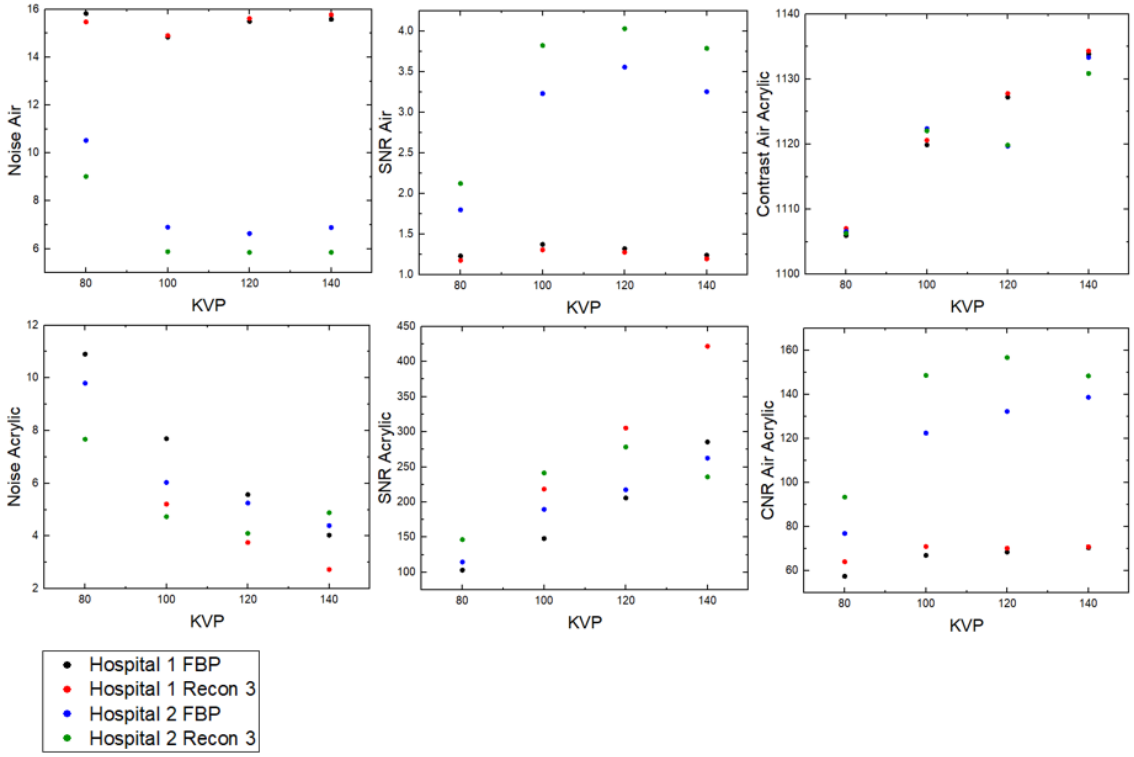


Figure D.2: Results of the basic Catphan measurements, where noise and SNR can be seen against KVP in both air and acrylic, and contrast and CNR between air and acrylic against KVP for both hospitals. Below the legend can be found

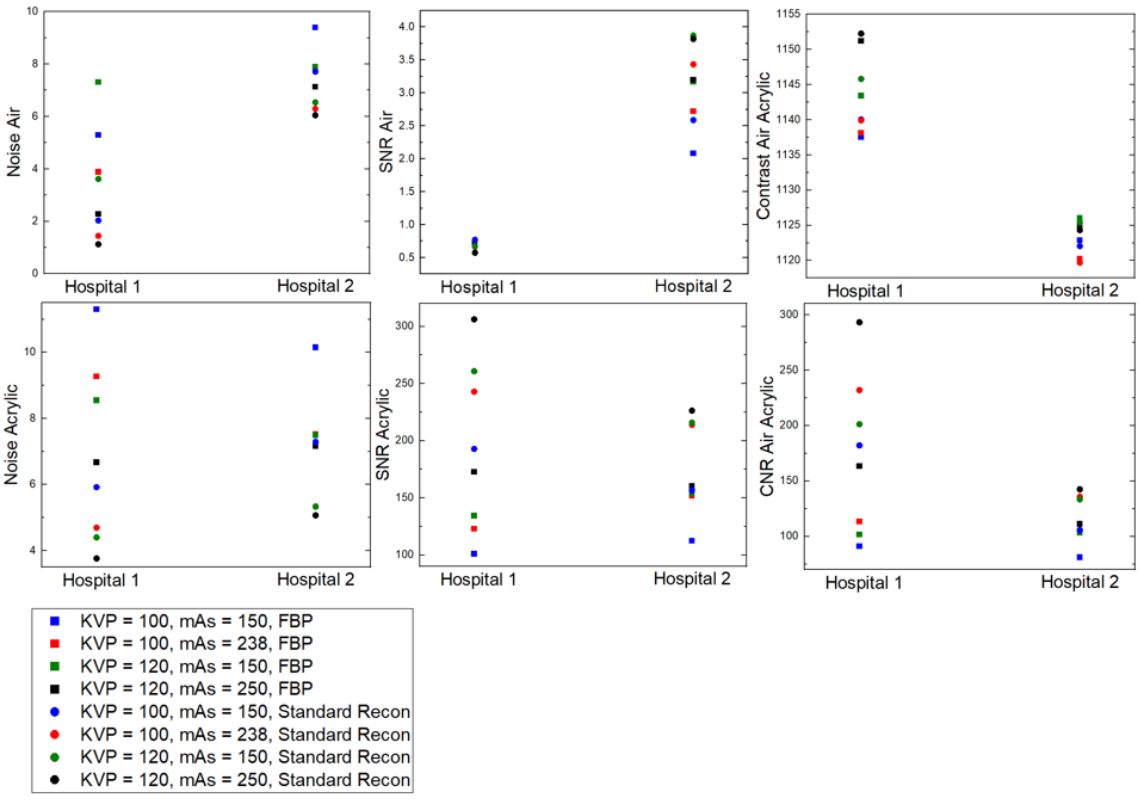


Figure D.3: Results of the PE Catphan measurements, where results for noise and SNR in both air and acrylic and contrast and CNR between air and acrylic are compared between both hospitals. Below the legend can be found.

Anthropomorphic phantom

In Figure D.4 the results for the anthropomorphic phantom scanned in Hospital 1 are given and in Figure D.5 the results for the anthropomorphic phantom scanned in Hospital 2 are given. As expected for both hospitals the noise decreases and SNR and CNR increases with higher CTDIvol. The correlations do however vary between the 4 different segmentation methods. Visually, when the CTDIvol increases the IQ of methods 3 and 4 seems to increase more. For both hospitals only some slight variations are found in IQ at the different reconstruction methods. Image quality does differ between the different hospitals. Signal, noise and SNR were higher in Hospital 2 compared to Hospital 1, whereas CNR was lower in Hospital 2. Differences could be expected as protocol and scanners were not the same: in Hospital 1 the PE protocol with slice thickness of 1 mm was used and in Hospital 2 the thorax protocol with a slice thickness of 3 mm was used and thinner slices have higher noise.

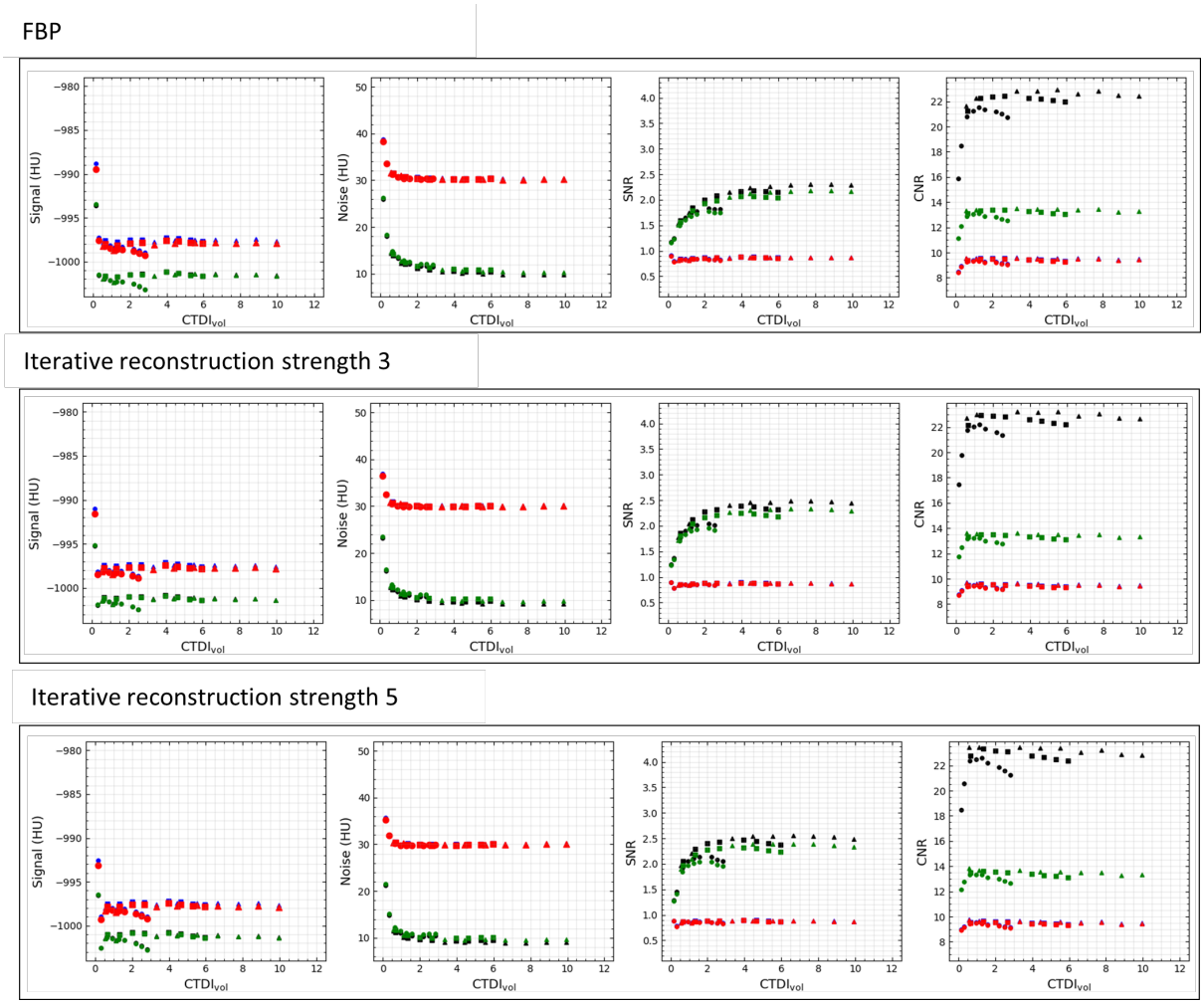


Figure D.4: Results of the Kyoto scanned in Hospital 1. The signal, noise, SNR an CNR are plotted against CTDIvol for method 1 (blue), method 2 (red), method 3 (black) and method 4 (green). Circles represent scans at 80 kV, squares represent scans at 100 kV and triangles represent scans at 120 kV.

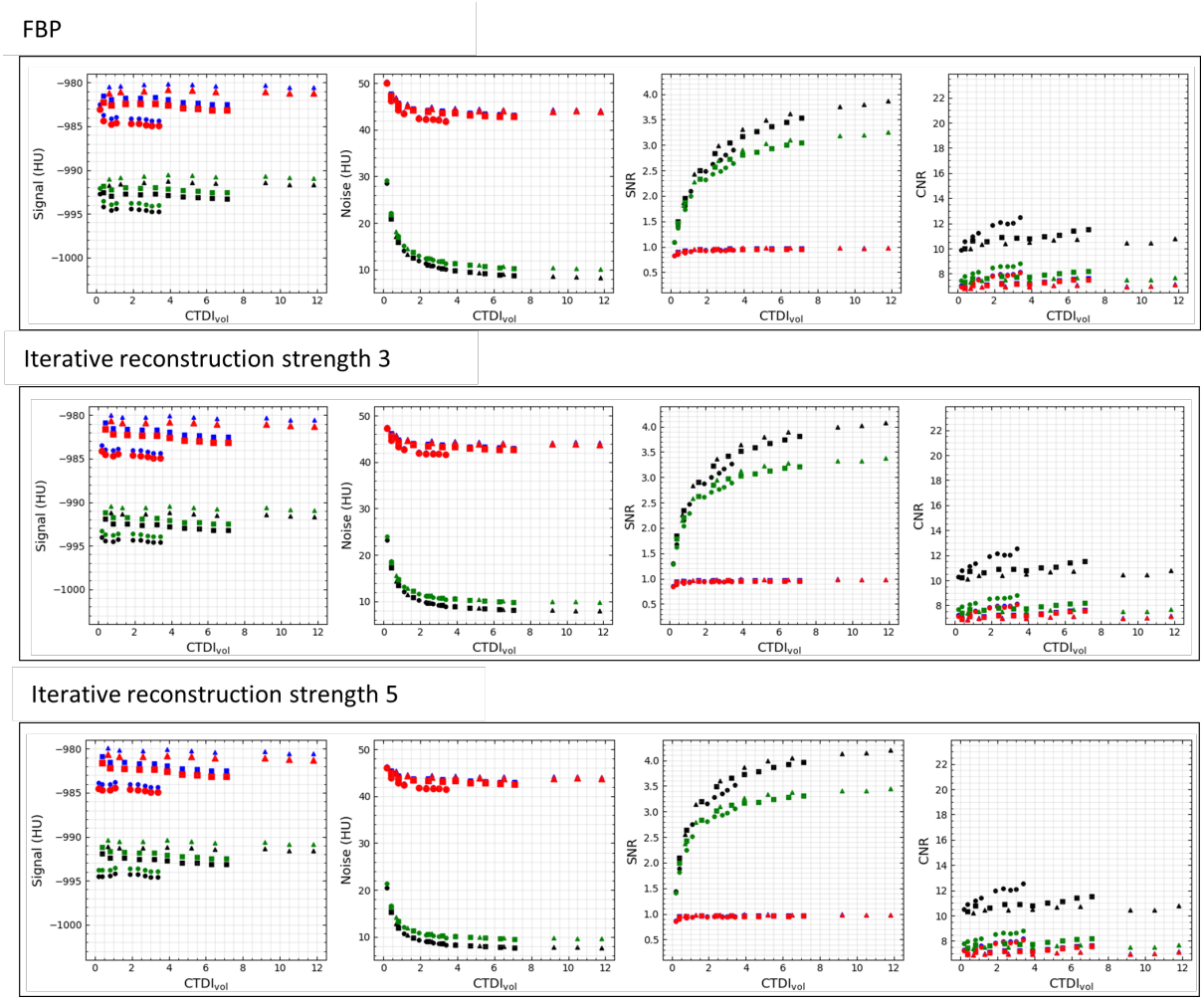


Figure D.5: Results of the Kyoto scanned in Hospital 2. The signal, noise, SNR an CNR are plotted against CTDIvol for method 1 (blue), method 2 (red), method 3 (black) and method 4 (green). Circles represent scans at 80 kV, squares represent scans at 100 kV and triangles represent scans at 120 kV.

D.3 Conclusion

To conclude, the IQ metrics versus dose do behave as expected. For the Catphan measurements both scanners and protocols did have similar results, except for some differences that were found in the signal and noise in the air regions. As expected, measurements on an anthropomorphic phantom showed correlations for segmentation based image quality and dose for all segmentation methods, where IQ increased with increasing dose. IQ metrics determined from segmentations using U-net + K-means clustering and U-net + Otsu thresholding increased most with increasing dose.

E PE Study: extended results

In this appendix all results are presented for the pulmonary embolism study. First for the different Likert scores, second IQ metrics with respect to the Likert score diagnostic confidence, third to the Likert score noise and last to the Likert score contrast enhancement. Additionally, the results of patients with and without pathology are compared and lastly the dose of the scans from both hospitals is compared.

E.1 Likert scores

In Figure E.1 the Likert scores are plotted against each other. It can be seen that Likert score contrast enhancement for both hospitals has high significant correlation with Likert score diagnostic confidence (adjusted $R^2 = 0.81/0.86$). Likert score noise however has low correlation with both Likert score diagnostic confidence and Likert score contrast enhancement, and most are not significant.

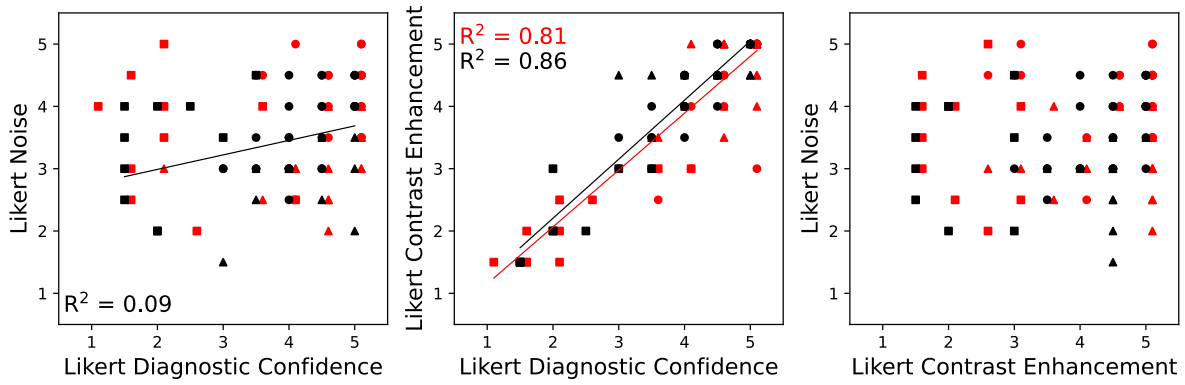


Figure E.1: The different types of Likert scores plotted against each other, where a linear fit and adjusted R^2 is included for significant correlations. Hospital 1 is in red and Hospital 2 in black. Circles represent normal scans, squares represent repeated scans and triangles represent FBP scans. For better visibility the x-axis of Hospital 1 has been shifted by 0.1.

E.2 Diagnostic confidence

In Figure E.2 all segmentation based IQ metrics and in Figure E.3 all literature based IQ metrics are plotted against Likert score diagnostic confidence for both hospitals. For significant correlations a linear fit with adjusted R^2 is included.

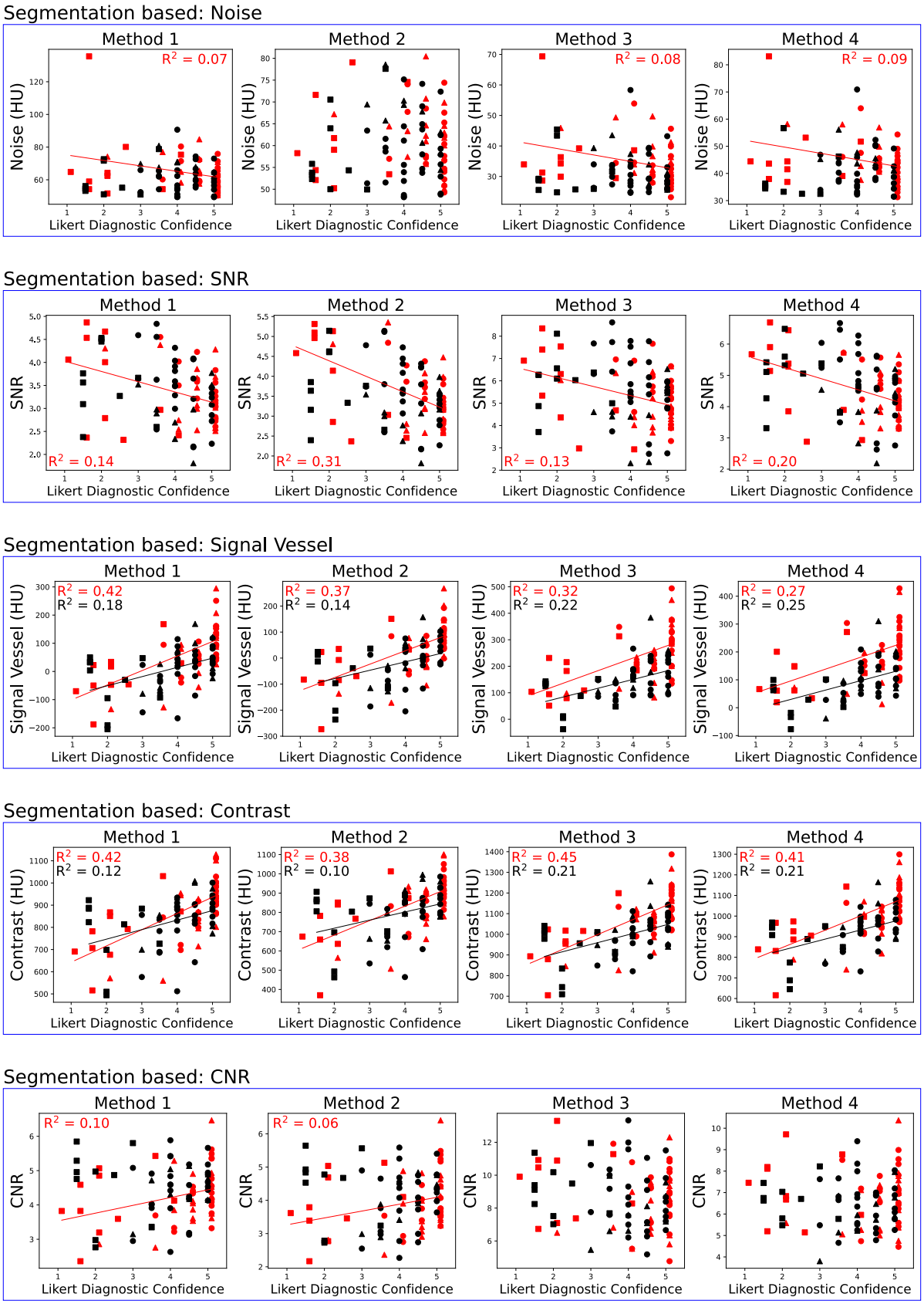


Figure E.2: All segmentation based IQ metrics for all segmentation methods against Likert score diagnostic confidence. A linear fit with adjusted R^2 is included when the correlation is significant. Hospital 1 is in red and Hospital 2 in black. Circles represent normal scans, squares represent repeated scans and triangles represent FBP scans. For better visibility the x-axis of Hospital 1 has been shifted by 0.1.

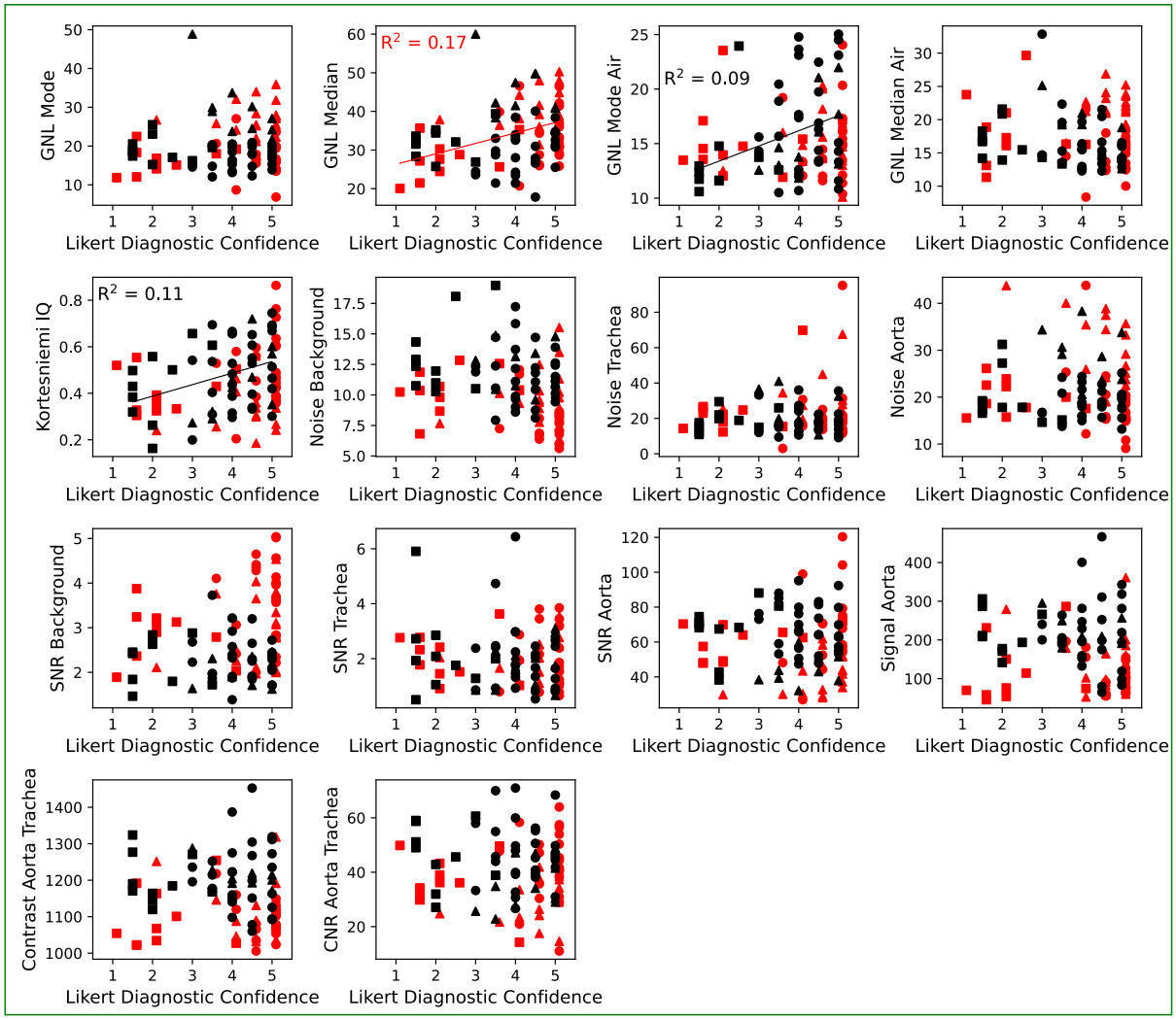


Figure E.3: All literature based IQ metrics against Likert score diagnostic confidence. A linear fit with adjusted R^2 is included when the correlation is significant. Hospital 1 is in red and Hospital 2 in black. Circles represent normal scans, squares represent repeated scans and triangles represent FBP scans. For better visibility the x-axis of Hospital 1 has been shifted by 0.1.

E.3 Noise

In Table E.1 the correlations (adjusted R^2) between all noise related IQ metrics and Likert score noise for both hospitals can be seen. Also the p-values between sufficient versus insufficient noise are given. In Figure E.4 all noise related segmentation based IQ metrics and in Figure E.5 all noise related literature based IQ metrics are plotted against Likert score noise.

Table E.1: Table with results for all noise related IQ metrics related to Likert score noise for both hospitals. The correlation (adjusted R^2) and the p-value between sufficient versus insufficient noise are given. Non-significant correlations or p-values are indicated with 'ns'.

	Hospital 1		Hospital 2	
IQ metric	R2 linear correlation	p-value sufficient vs insufficient	R2 linear correlation	p-value sufficient vs insufficient
Segmentation based IQ				
Noise method 1	0.18	<0.001	0.10	0.01
Noise method 2	0.20	0.001	0.06	0.01
Noise method 3	0.26	<0.001	0.15	<0.001
Noise method 4	0.28	<0.001	0.13	0.004
SNR method 1	ns	ns	ns	ns
SNR method 2	ns	ns	0.06	ns
SNR method 3	ns	ns	ns	ns
SNR method 4	ns	ns	ns	ns
Literature based IQ				
Kortesniemi IQ	0.48	0.001	0.39	<0.001
GNL Mode	0.40	ns	0.33	<0.001
GNL Median	0.18	ns	0.13	0.009
GNL Mode Air	ns	ns	ns	ns
GNL Median Air	0.28	ns	0.31	0.009
Noise Background	0.25	0.01	ns	ns
Noise Trachea	0.06	0.008	0.11	0.04
Noise Aorta	0.41	0.03	0.41	<0.001
SNR Background	0.17	ns	ns	ns
SNR Trachea	ns	ns	ns	ns
SNR Aorta	0.41	0.02	0.37	<0.001

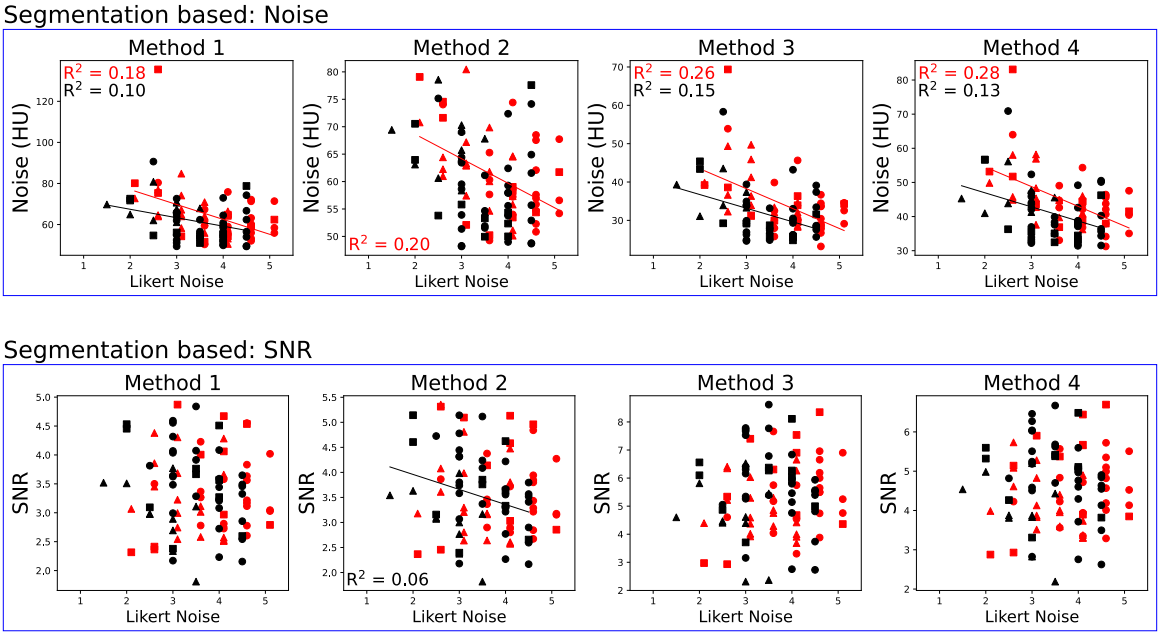


Figure E.4: All noise related segmentation based IQ metrics for all segmentation methods against Likert score noise. A linear fit with adjusted R^2 is included when the correlation is significant. Hospital 1 is in red and Hospital 2 in black. Circles represent normal scans, squares represent repeated scans and triangles represent FBP scans. For better visibility the x-axis of Hospital 1 has been shifted by 0.1.

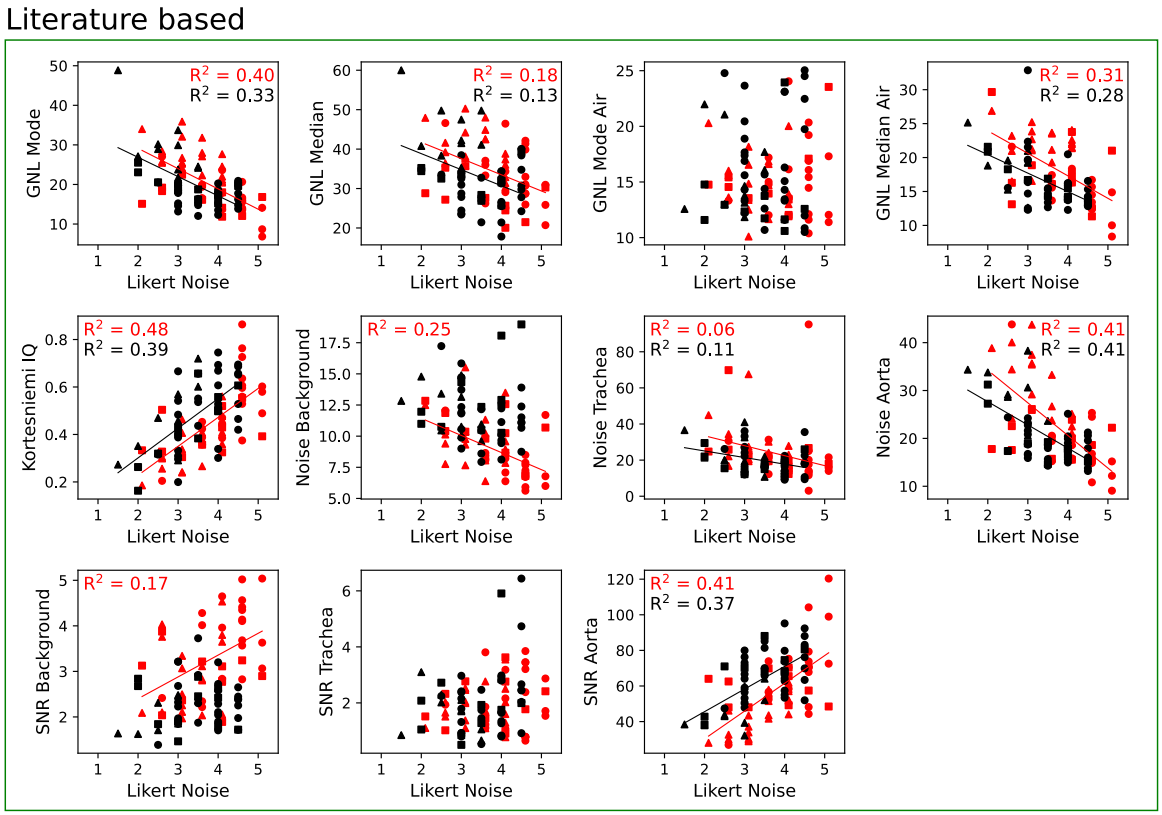


Figure E.5: All noise related literature based IQ metrics against Likert score noise. A linear fit with adjusted R^2 is included when the correlation is significant. Hospital 1 is in red and Hospital 2 in black. Circles represent normal scans, squares represent repeated scans and triangles represent FBP scans. For better visibility the x-axis of Hospital 1 has been shifted by 0.1.

E.4 Contrast enhancement

In Table E.2 the correlations (adjusted R^2) between all contrast related IQ metrics and Likert score contrast enhancement for both hospitals can be seen. Also the p-values between sufficient versus insufficient contrast enhancement are given. In Figure E.6 all contrast related segmentation based IQ metrics and in Figure E.7 all contrast related literature based IQ metrics are plotted against Likert score contrast enhancement.

Table E.2: Table with results for all contrast related IQ metrics related to Likert score contrast enhancement for both hospitals. The correlation (adjusted R^2) and the p-value between sufficient versus insufficient contrast enhancement are given. Non-significant correlations or p-values are indicated with 'ns'.

	Hospital 1		Hospital 2	
IQ metric	R2 linear correlation	p-value sufficient vs insufficient	R2 linear correlation	p-value sufficient vs insufficient
Segmentation based IQ				
Signal vessel method 1	0.33	<0.001	0.15	ns
Signal vessel method 2	0.26	0.001	0.09	ns
Signal vessel method 3	0.29	0.003	0.24	0.04
Signal vessel method 4	0.25	0.005	0.23	0.02
Contrast method 1	0.32	<0.001	0.08	ns
Contrast method 2	0.27	<0.001	ns	ns
Contrast method 3	0.36	<0.001	0.19	ns
Contrast method 4	0.33	<0.001	0.16	ns
CNR method 1	ns	ns	ns	ns
CNR method 2	ns	ns	ns	0.05
CNR method 3	0.07	ns	ns	ns
CNR method 4	ns	ns	ns	ns
Literature based IQ				
Signal Aorta	ns	ns	ns	ns
Contrast Aorta Trachea	ns	ns	ns	ns
CNR Aorta Trachea	ns	ns	ns	ns

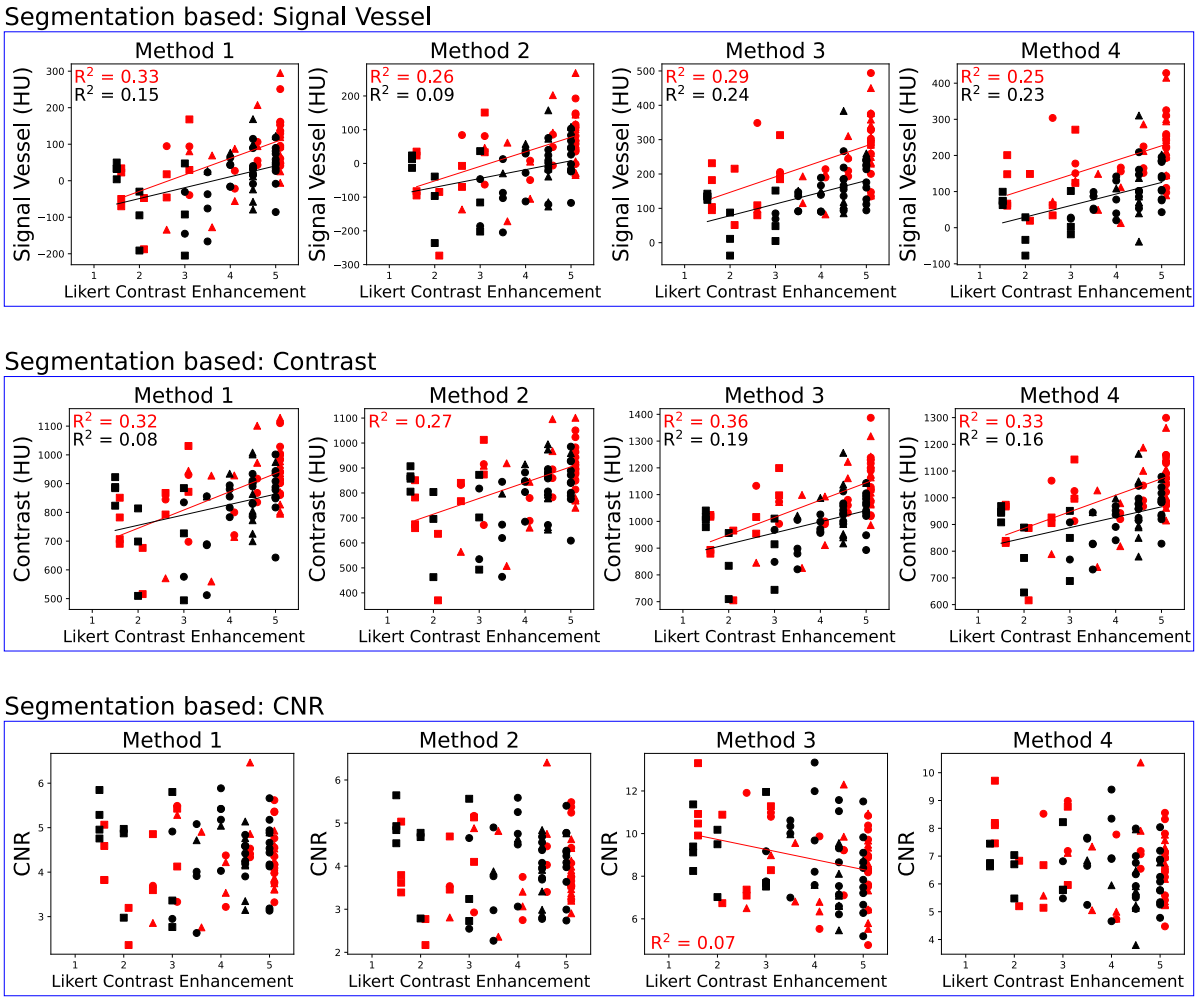


Figure E.6: All contrast related segmentation based IQ metrics for all segmentation methods against Likert score contrast. A linear fit with adjusted R^2 is included when the correlation is significant. Hospital 1 is in red and Hospital 2 in black. Circles represent normal scans, squares represent repeated scans and triangles represent FBP scans. For better visibility the x-axis of Hospital 1 has been shifted by 0.1.

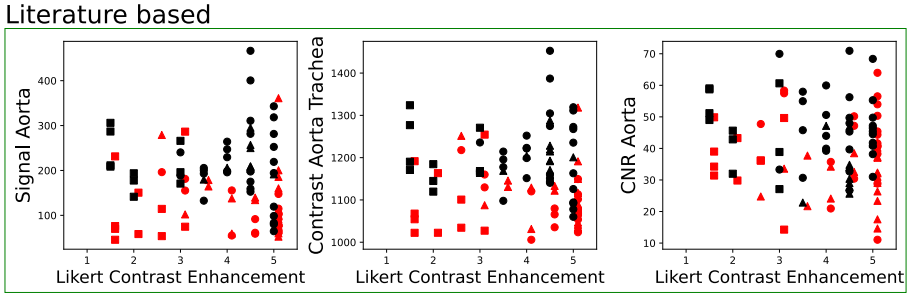


Figure E.7: All contrast related literature based IQ metrics against Likert score contrast. A linear fit with adjusted R^2 is included when the correlation is significant. Hospital 1 is in red and Hospital 2 in black. Circles represent normal scans, squares represent repeated scans and triangles represent FBP scans. For better visibility the x-axis of Hospital 1 has been shifted by 0.1.

E.5 Influence of pathology

In Figure E.8 and E.9 results are given for all patients for Hospital 1, where the segmentation based IQ results for all methods and the Likert scores are compared between patients with and without pathology. In Figure E.10 and E.11 similar graphs can be found for the patients scanned in Hospital 2.

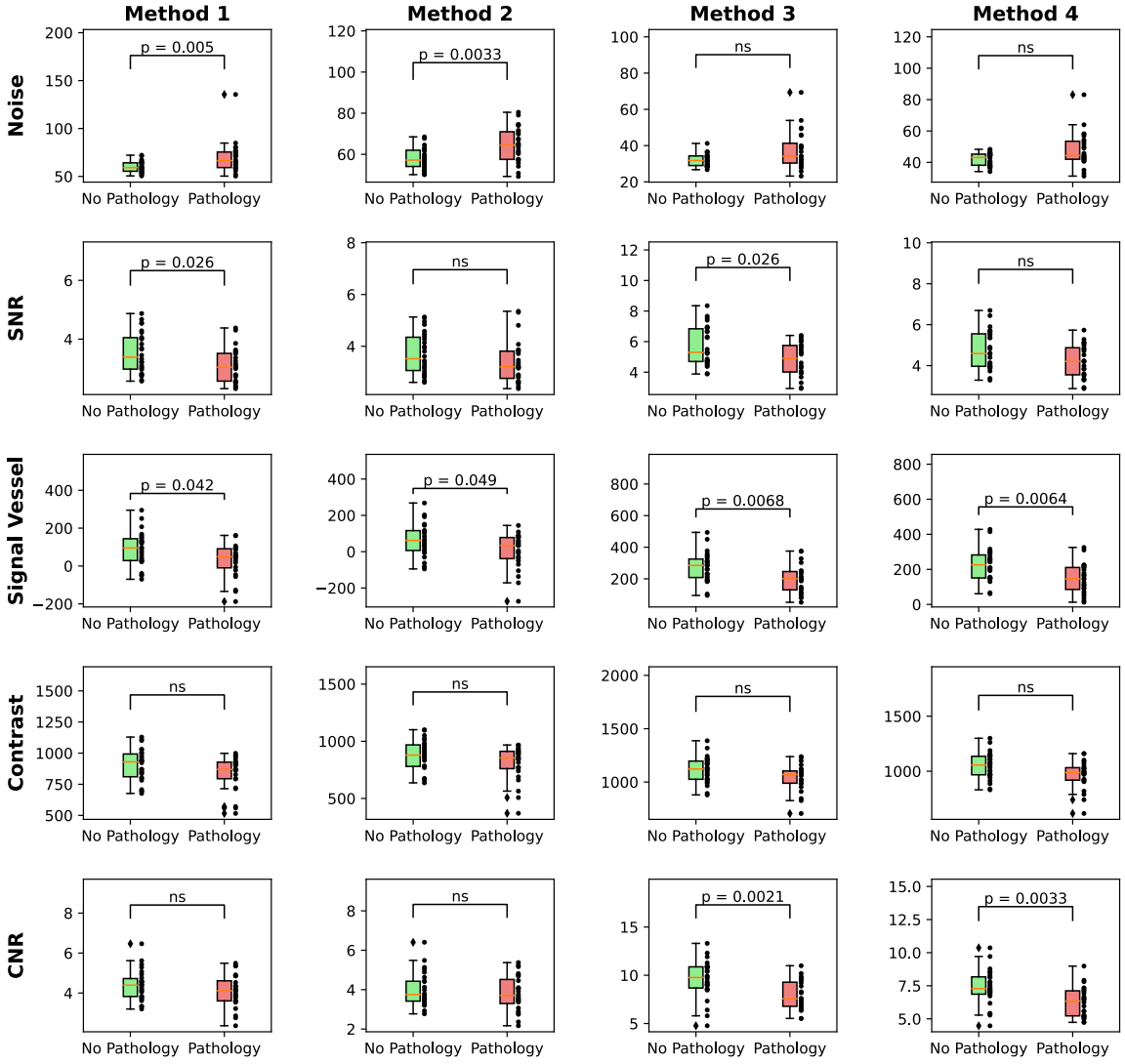


Figure E.8: Results of noise, SNR, signal vessel, contrast and CNR for all methods are given for patients without pathology (green) and patients with pathology (red) scanned in Hospital 1, including the according p-values. For non significant p-values 'ns' is indicated.

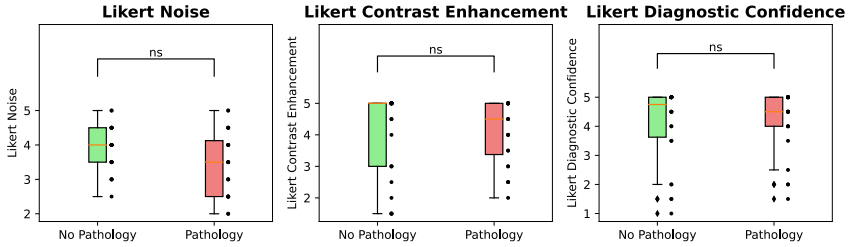


Figure E.9: The 3 different Likert scores (Noise, contrast enhancement and diagnostic confidence) are given for patients without pathology (green) and patients with pathology (red) scanned in Hospital 1, including the according p-values. 'ns' is indicated if the p-values were not significant.

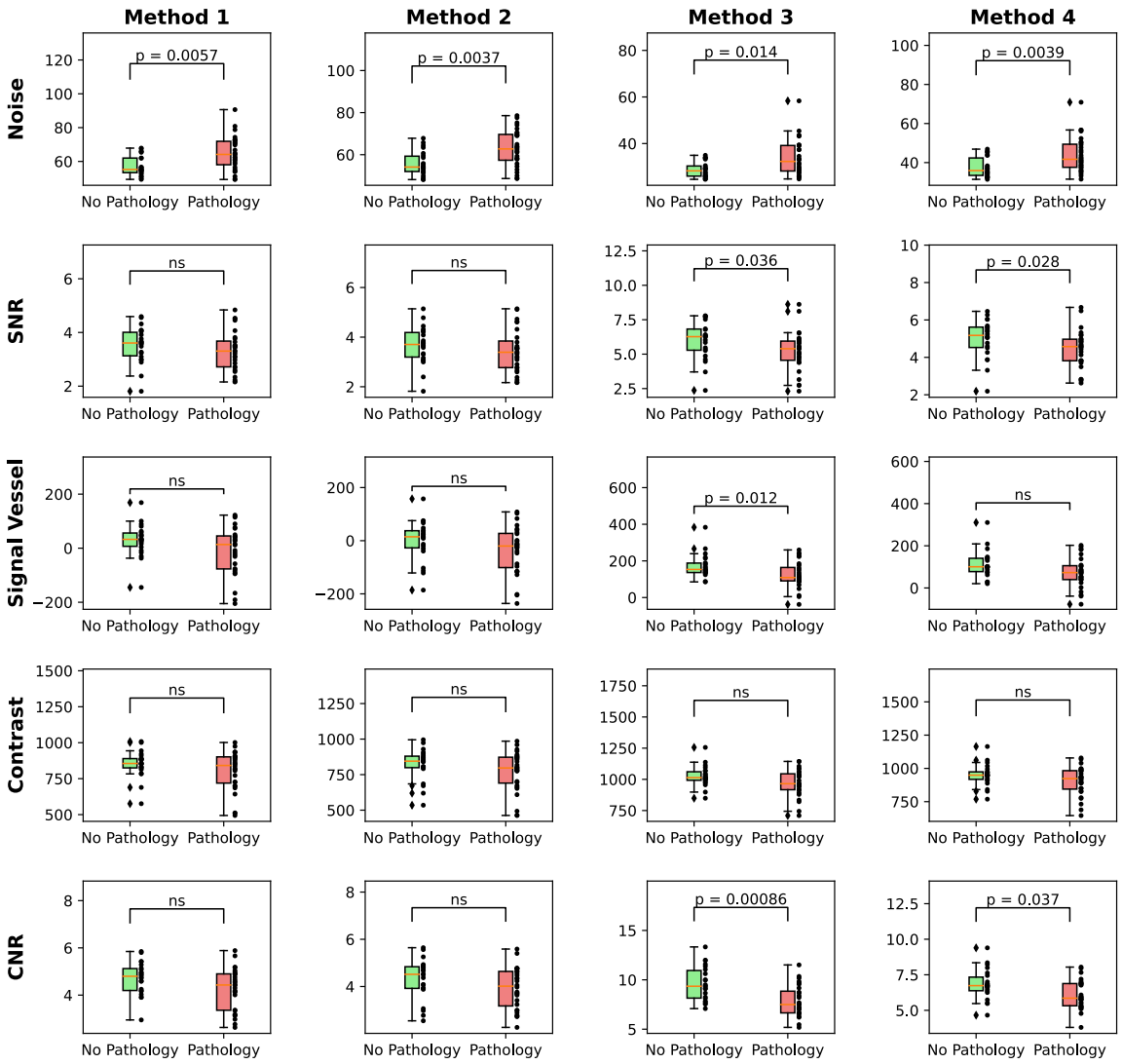


Figure E.10: Results of noise, SNR, signal vessel, contrast and CNR for all methods are given for patients without pathology (green) and patients with pathology (red) scanned in Hospital 2, including the according p-values. For non significant p-values 'ns' is indicated.

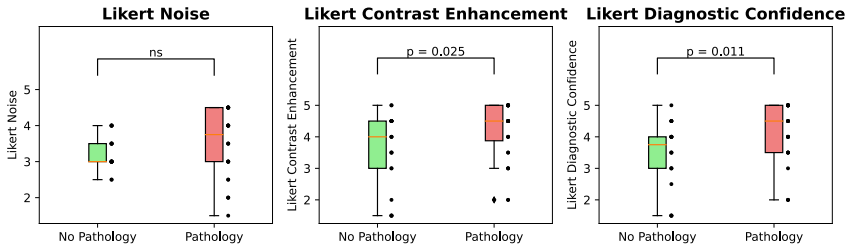


Figure E.11: The 3 different Likert scores (Noise, contrast enhancement and diagnostic confidence) are given for patients without pathology (green) and patients with pathology (red) scanned in Hospital 2, including the according p-values. For non significant p-values 'ns' is indicated.

E.6 Dose info of both datasets

In Figure E.12 the CTDIvol is plotted against mAs for both datasets. Here it can be seen that the average mAs and CTDIvol of Hospital 2 are higher than of Hospital 1, mostly because a few patients are scanned with a higher mAs in Hospital 2.

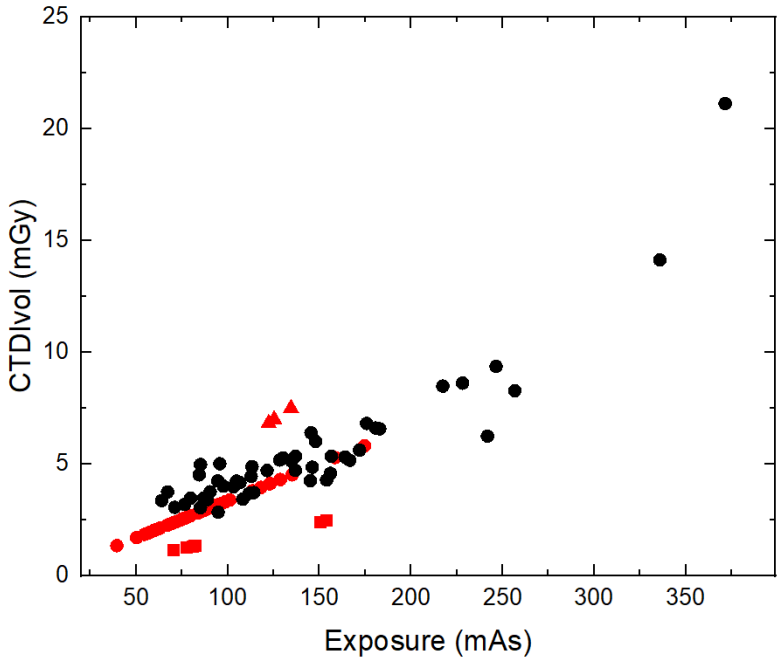


Figure E.12: The CTDIvol plotted against exposure for Hospital 1 (red) and Hospital 2 (black). Squares represent scans at 80 kV, circles at 100 kV and triangles at 120 kV.

F Moderator linear regression

In this Appendix the statistical method of regression with a moderator is explored. Moderation occurs when the relationship between two variables is dependent of a third variable, called the moderator variable. Moderated linear regression is defined by Hayes [48]:
"The effect of X on some variable Y is moderated by W if its size, sign, or strength depends on or can be predicted by W. In that case, W is said to be a moderator of X's effect on Y, or that W and X interact in their influence on Y."
The simple linear moderation model is given by

$$Y = i_Y + b_1 X + b_2 W + b_3 XW + e_Y,$$

with intercept i_Y, slopes b_1, b_2 and b_3 and the model's error term e_Y. In Figure F.1 a conceptual diagram of this simple linear moderation model is given.

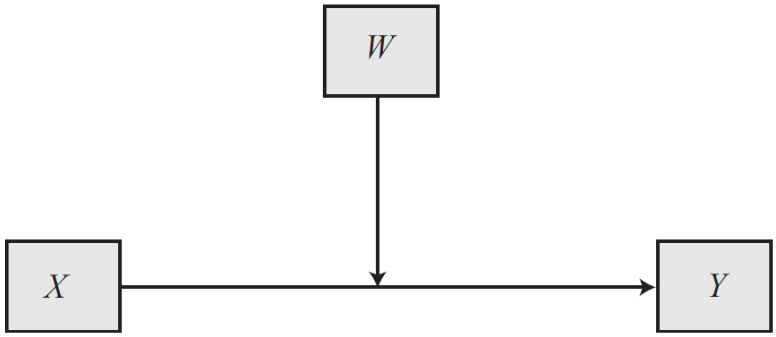


Figure F.1: Conceptual diagram of a simple moderation model [48].

The model's interaction ($X \times W$) can be determined. If the interaction term is statistically significant, this means that the effect of X on some variable Y is moderated by W.

Moderated linear regression was applied for the segmentation based IQ metrics, to see whether the effect of calculated IQ metrics on the radiologist Likert scores is moderated by the presence of pathology in the lungs. Significance of the interaction terms was calculated seperately for all segmentation based IQ metrics against diagnostic confidence Likert scores, and no interaction terms were found to be significant. Also, this was repeated for the noise related segmentation IQ metrics against noise Likert scores and contrast enhancement related segmentation IQ metrics against contrast enhancement Likert scores, and again no significant interaction terms were found.
This means that there is not sufficient evidence to conclude that there is a significant moderation effect of the presence of pathology in the lungs, suggesting that the effect is probably very small. However, this method still needs further exploration with a statistician before conclusions can be drawn based on these results.

G Abstracts

We have submitted two abstracts about this study. The first one was for the NVKF conference in Woudschoten and focused on the thorax noduli study. It must be noted that after we submitted this abstract an extra step was included in the method: remove segmented vessels with an area larger than 2.5 cm^2 as they are too likely to consist of pathology. Therefore the results in this abstract are slightly different from the results given in Appendix B. The second one was for the ECR 2023 conference in Vienna and was focused on the pulmonary embolism study. Both abstracts can be seen below.

G.1 Abstract NVKF conference Woudschoten 2022

Automatic Image Quality Parameter Determination for Thorax CT

Lisan van Haren, Cécile Jeukens, Evie Hoeijmakers, Emmeline Laupman, Carola van Pul

Background: To prevent that patients receive more dose than necessary, CT scanning protocols should be optimised according to the ALARA principle. Currently, image quality (IQ) is usually determined using a physical phantom or radiologist scoring. However, a phantom does not resemble patients' anatomy and manual scoring by radiologists is time consuming and results are subjective.

Objectives: The aim of this study is to develop an algorithm that automatically determines noise and contrast-related image quality metrics in thorax CT. To focus on the tissue of interest, an automatic lung and pulmonary vessel segmentation method is developed.

Methods: A dataset of 40 patients who had a thorax CT scan for noduli detection is collected retrospectively, of which 20 were scanned on a Philips iCT and 20 on a GE PET-CT. For all patients lungs and pulmonary vessels are segmented with 4 different methods: thresholding, U-net plus thresholding, U-net plus K-means clustering and U-net plus multi-Otsu thresholding. The signal-to-noise ratio (SNR) is calculated in the resulting lung masks and the contrast-to-noise ratio (CNR) is calculated between the lung and vessel masks.

Results: Visual analysis of the segmentations showed that for patients without pathology in the lungs segmentation was successful (Figure 1A). However, segmentations were influenced by pathology, leading sometimes to partial inclusion of pathology in masks and thus to differences in SNR/CNR between patients with and without pathology. Differences in CNR were largest for the thresholding method and smallest for the U-net-multi-Otsu-thresholding method, but were not statistically significant (Figure 1B). This indicates that the presence of pathology is not an influential factor. On the other hand (Figure 1C), CNR did differ significantly between the two scanners, likely reflecting differences in protocol settings and iterative reconstruction algorithms.

Conclusion: Automatically calculating IQ metrics in thorax CT scans is feasible. Segmentations for patients with pathology in the lungs were less accurate, however IQ results were not significantly different, indicating that the algorithm is robust in determining IQ even in the presence of pathology. The algorithm can analyse scans from different scanners and is able to demonstrate inter-scanner differences.

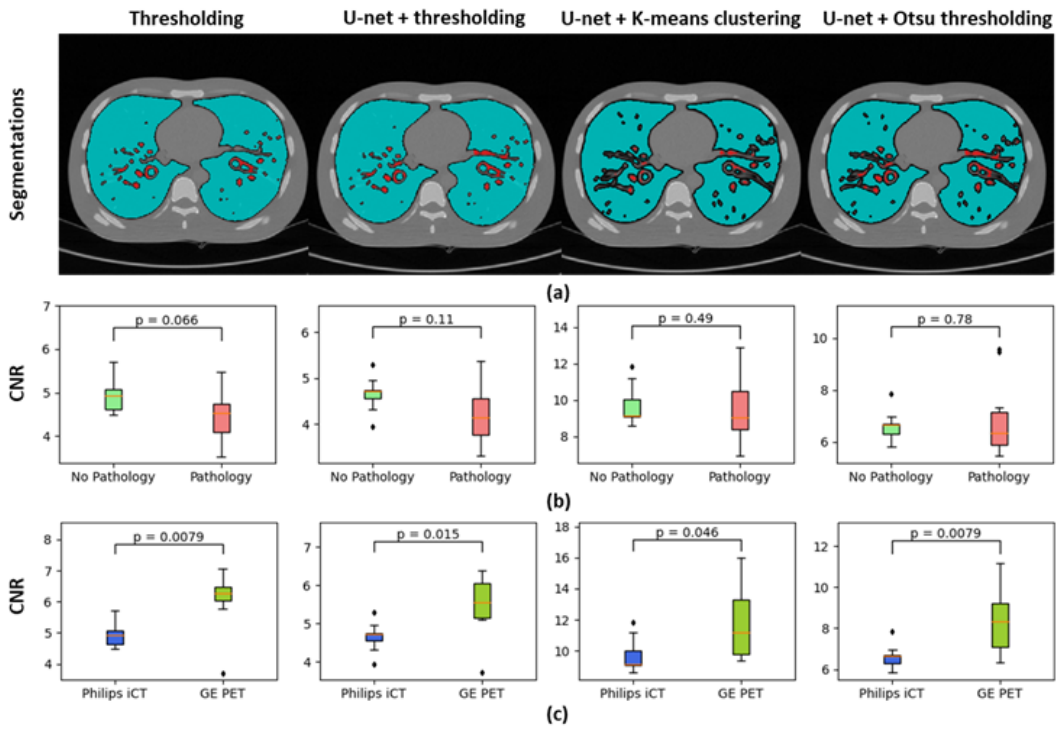


Figure 1: A) examples of segmentations are given for all 4 methods for an anthropomorphic phantom (PBU-60 phantom, Kyoto Kagaku Co.). B) results of CNR for all methods are given for patients without pathology (green) and patients with pathology (red) scanned on a Philips iCT. C) results of CNR are given for patients without pathology scanned on a Philips iCT (blue) and on a GE PET-CT (green).

G.2 Abstract ECR 2023 conference Vienna

Automatic Image Quality Parameter Determination for Pulmonary Embolism CT-Scans

Lisan van Haren, Carola van Pul, Evie Hoeijmakers, Emmeline Laupman, Mark van der Vlies, Bibi Martens, Babs Hendriks, Cécile Jeukens

Purpose: To develop an algorithm that automatically determines noise and contrast-related image quality (IQ) metrics for quality control purposes in pulmonary embolism (PE) CT-scans, and to assess agreement between objective IQ metrics and subjective IQ scoring by radiologists.

Methods: In two hospitals, two datasets of 50 consecutive clinical PE CT-scans were retrospectively collected, including repeated scans, having a lower IQ. We developed an algorithm based on lungs and pulmonary vessels segmentation using a U-net and K-means clustering. In these 3D segmentations five IQ metrics were calculated: noise and signal-to-noise ratio in lungs, mean signal in vessels, and contrast and contrast-to-noise ratio between lungs and vessels. Additionally, five noise IQ metrics used in literature were calculated. In each hospital, two radiologists scored noise, contrast attenuation of the pulmonary arteries and diagnostic confidence separately using a 5-point Likert scale (1=poor to 5=excellent). Additionally, presence of lung pathology (PE, lesions, emphysema, effusion) was recorded, as this may influence the automatic segmentation and calculated IQ-scores. Regression analysis was performed to assess correlation between IQ metrics and Likert scores, reporting adjusted R^2 and significance.

Results: The algorithm was able to automatically calculate all IQ metrics for all scans. Multiple significant correlations were found between IQ metrics and all 3 different Likert scores. For diagnostic confidence strongest correlations were observed for contrast (R^2= 0.45/0.21) and vessel signal metrics (R^2= 0.32/0.22). Results were consistent between both hospitals: overall best and least performing IQ metrics were similar. Lung pathology did not significantly influence the IQ scores.

Conclusion: It is feasible to determine objective IQ metrics that correlate with subjective IQ, based on an automatic algorithm in PE CT-scans. This opens possibilities for continuous image quality monitoring in clinical practice.