Applicability of a clinical cardiac CT protocol in post mortem studies

Abstract

Objective: Confirmation whether an optimized clinical cardiac CT scan protocol is also optimal for post mortem cardiac CT scans without iodine contrast or the reconstruction parameters should be changed.
Materials and methods: 27 CT volumes (three cases for three reconstruction kernel with three different iterative reconstruction settings) were graded by six readers in order to find the optimal reconstruction parameters. The scans were performed on a Siemens Definition Flash CT scanner using 120kV tube potentials.
Results: The study has shown that from the investigated options the softest cardiac kernel with the strongest iterative reconstruction were preferred by the readers (I26 Safire 3).
Conclusion: The results indicate that the scan protocol which was adopted from clinical practice is applicable in forensic radiology too even though iodine contrast agent was not administered.
Keywords: image quality, observer study, CT, cardiac, post mortem

1. Introduction

The number of performed forensic computed tomography (CT) scans is considerable smaller than in general clinical practice. While special circumstances might require specialized methodology[1], clinical scans form the basis for recommended protocols for routine post mortem scans. The post mortem alterations[2] and the general lack of iodine contrast administration might introduce a need for scan protocols optimized for forensics medicine. Cardiac pathology is the most common cause of sudden unexpected death, and dedicated post mortem cardicac CT (PMcCT) is an experimental attempt to better visualize pathology when angigraphy has not been performed.

The aim of this study was to evaluate whether readers find the same cardiac CT scan protocol best for post mortem investigation as is recommended for clinical scans even though inferior vena cava (IVC) contrast was not administered.

Filtered back projection (FBP) is considered the gold standard reconstruction for decades in CT imaging. The main drawback of FBP is the fact that it cannot take noise properties of the signal into account. Iterative reconstruction (IR) was introduced to more accurately model the physical and statistical

phenomena. IR achieves similar-to-FBP image quality at reduced dose, or less noise at same dose level. However, iterative reconstruction also changes the noise power spectrum, and the textures in the image. Depending on the modeling strategy, whether it is image-space only, sinogram-space, or both, and what physical phenomena are included in the model, finding the optimal parameters are still an open question. FBP has more limited options to mitigate noise. Different kernel-families were developed by the vendors to find a good balance between sharpness and noise. Often the iterative reconstructions are derived from the FBP reconstruction, and they try to yield the same sharpness as their FBP counterparts but with reduced noise level. Detailed review about the state of the art iterative algorithms was published by Geyer et al.[3]

Image quality (IQ) assessment is a major challenge in radiology due to the multivariate optimization problem such as dose, sharpness, noise, among others[4,5,6]. Several methods are applied in order to determine the best clinical settings, as it presented in a review from Manssons et al.[7].

While special circumstances, e.g. presence of foreign objects, might cause image artifacts[8], arguable post mortem alterations mostly affect the subjective image quality. If readers evaluate several different reconstructions, would they prefer the same reconstruction settings as it was recommended in the clinical practice? This question is the core of this paper, and a visual grading study was performed to answer it.

Arguably, one of the most frequently used methods for subjective IQ assessment in radiology is the visual grading (VG) with or without reference image[7]. It has advantages, it scales well with the number of images, and has disadvantages, e.g. the difficulty of choosing criteria. The European Commission (EC) released guidelines for image quality criteria for computed tomography[9] which helps to standardize them. Similarly, the statistical analysis of VGA studies went through a long evolution from t-tests to visual grading characteristics (VGC)[10] and visual grading regression (VGR)[11].

This paper focuses on two choices for protocol optimization: the reconstruction kernel selection and the use of iterative reconstruction. These options were investigated using published IQ criteria, and analyzed with VGR.

Many other parameters could be investigated, including dose, slice thickness, tube voltage, but in this study these parameters were kept constant, and only the effect of the various kernels and iterative reconstruction parameters was studied.

2. Materials and methods

2.1. Data collection and readers

Three cardiac scans, later referenced as *cases*, performed on a Siemens Somatom Definition Flash dual-source multi-slice CT scanner (Siemens AG, Forchheim,

Scan parameter	Value
tube potential	120kV
tube current	400mAs (effective)
total detector collimation	38.4mm
dose modulation	on
slice thickness	1 mm
pitch	0.6
window center	200 HU
window width	600 HU

Table 1. Scan and reconstruction parameters

Germany) [1] were selected for the study using three different reconstruction kernels with filtered back projection (FBP) and two iterative reconstruction (Safire 2 and 3). B26, B36 and B46 kernels were used for FBP, and the closely related I26, I36, I46 kernels were used for Safire reconstructions. The three scans with three kernels and the three reconstruction options yielded 27 CT volumes in total.

FBP and Safire 2 and 3 were chosen because in general these are the gold standard and the most preferred options[12], respectively. Scan and reconstruction parameters are presented in Table 1. Automatic dose modulation were used, and the table contains the effective tube current.

In this paper, B26/I26 notation refers to the group of B26 with FBP and I26 with Safire 2 and Safire 3. B36/I36 and B46/I46 follow the same logic.

These CT volumes were presented to six readers on calibrated 10bit medical display (EIZO Radiforce G22[2]) from approximately 60cm distance and using the display windows set by the scanner. No time constraints were given. The volumes were presented blinded, anonymized and in randomized order for each reader, both with respect to the reconstruction parameters and to the cases.

The three of the six readers were experienced radiographers, and three of the six were second year radiographer students. At the beginning of the reading session the objectives of the study and the use of the evaluation software were explained to the readers and they could ask as questions if they wished.

However, one of the readers (reader 4) made a mistake during recording responses, and was allowed to restart the study. While including or excluding this reader does not change the outcome, later we decided to exclude the reader's responses from the study. However, the responses are available in the online dataset.

[1] https://www.healthcare.siemens.com/computed-tomography/dual-source-ct/somatom-definition-flash, visited on 21st April, 2017

[2] http://www.ampronix.com/eizo-radiforce-g22.html, visited on 21st April, 2017

Identifier	Criterion
C1	Visually sharp reproduction of the heart
C2	Reproduction of the left venticular
C3	Visually sharp reproduction of the pleuromediastinal border
C4	Sharp/clear demarcation of the aortic wall
C5	General impression of contrast
C6	General impression of noise
C7	General impression of artifacts

Table 2. Criteria for visual grading.

2.2. Visual grading

Seven criteria were presented to the readers in their native language, and they were asked to score the CT volumes on a 5-grade scale where 1 was the worst and 5 was the best. The translation of the criteria and the specific meaning of the grades were presented in Table 2 and 3, respectively. The criteria C1-C4 are based on EC guidelines[9], and C5-C7 are meant to represent general perceived image quality.

ViewDEX presentation software[13] were used to display the volumes and criteria, and record the responses and the response times. The response times were only used for quality control (see in Discussion part), and were not taken into account as image quality descriptors.

2.3. Data analysis

Visual grading uses ordinal scale which means that higher score belongs to better results but the difference between 2 and 3 is not necessarily the same as between 3 and 4. The aim is to determine which kernel and which iterative reconstruction peform the best while the external factors (e.g. reader differences) are taken into account.

For this purpose, visual grading regression[11] was developed, which is model based on ordinal logistic regression. This model predicts the logarithm of odds ratio (OR) changes if a risk factor is present. These odds ratios are a bit harder to interpret than differences between mean values, but they are mathematically more founded. For details of VGR we refer back to the original publication[11].

However, the mean VGA score (VGAS)[7] is still widely used, and despite their shortcomings[14], shows the magnitude of the difference between options, and makes the results more comparable to other VGA studies. These two models, VGR and VGAS were the core for analyzing the data.

VGAS is calculated as follows:

$$\mathrm{VGAS} = \frac{\sum_{I,O} S_c}{N_I N_R} \tag{1}$$

where S_c is the given score, and it should be averaged for the number of readers(N_R) and the number of images (N_I).

Criteria	Score	Confidence level
C1, C2, C3, C4 (see Table 2)	5	very well reproduced – the structure had completely distinct shape
	4	well reproduced – the structure was clearly reproduced
	3	adequate reproduced – the structure was moderately reproduced
	2	poorly reproduced - the structure was vaguely reproduced
	1	not reproduced - the structure could not be discerned
		Contrast with regards to diagnostic
C5	5	very good
	4	good
	3	medium
	2	low
	1	unacceptable
		Noise level and artefact with regards to diagnostics are
C6, C7	5	not disturbing at all
	4	barely disturbing
	3	moderately disturbing
	2	quite disturbing
	1	highly disturbing

Table 3. Scoring levels for visual grading

For using ordinal scale and taking into account the different factors (kernels, iterative reconstruction, readers differences, etc.), and deciding whether an effect is significant, one should use the regression model.

The data analysis were conducted with the R open source statistical language[15]. The ordinal logistic regression model is calculated with the MASS package[16] from R. Using R's notation, the model is the following:

$$\text{polr}(\text{score} \sim \text{IR} + \text{Kernel} + \text{ReaderID} + \text{CaseID}) \tag{2}$$

where *polr* denotes the ordinal logistic regression function, score is the predicate, and IR, Kernel, ReaderID and CaseID are the non-numeric groups for iterative reconstructions, reconstruction kernels, readers and the scanned volume, respectively. This model takes into account the iterative reconstructions, the kernels and also the differences between the readers and between the cases. These later two factors might influence the scores, therefore, they should be taken into account.

For each criterion, the odds ratios with corresponding p-values were determined. The odds ratios in the ordinal logistic regression model are multiplicative factors, if more than two options are present.

Tables 4 and 5 report the result of the statistical model of VGR for different iterative reconstruction options and for the different kernels, respectively. Note that the model uses B26 with FBP as baseline, and results show the increased or decreased odds if a parameter was changed. To ease the interpretation, the most preferred options are presented in the last column of the tables.

Criterion	FBP < Safire 2			Safire 2 < Safire 3			preferred
	odds ratio	p	95% CI	odds ratio	p	95% CI	
C1	1.35	0.458	0.61-3.03	3.04	0.009	1.34-7.06	Safire 3
C2	1.23	0.614	0.55-2.74	2.52	0.025	1.13-5.70	Safire 3
C3	1.01	0.977	0.42-2.42	4.21	0.002	1.71-10.80	Safire 3
C4	1.20	0.666	0.52-2.79	1.88	0.141	0.81-4.38	Safire 3
C5	0.77	0.521	0.34-1.71	1.35	0.461	0.61-3.03	Safire 3
C6	1.20	0.658	0.53-2.76	3.32	0.004	1.47-7.66	Safire 3
C7	0.69	0.380	0.30-1.57	1.15	0.734	0.51-2.62	Safire 3
Overall	1.01	0.974	0.48-2.15	2.67	0.011	1.26-5.73	Safire 3

Table 4. Ordinal logistic regression results for iterative reconstruction options. Most preferred option is summarized in the last column. Significant ($p<0.05$) values are indicated with green color.

Criterion	B26/I26 < B36/I36			B36/I36 < B46/I46			preferred
	odds ratio	p	95% CI	odds ratio	p	95% CI	
C1	0.87	0.722	0.39-1.92	0.48	0.080	0.21-1.09	B26/I26
C2	0.58	0.180	0.26-1.28	0.68	0.341	0.30-1.51	B26/I26
C3	0.55	0.190	0.22-1.33	0.27	0.005	0.11-0.67	B26/I26
C4	0.86	0.725	0.37-2.00	0.27	0.003	0.11-0.63	B26/I26
C5	0.64	0.281	0.29-1.43	0.44	0.053	0.19-1.01	B26/I26
C6	0.47	0.073	0.21-1.07	0.12	<0.001	0.05-0.29	B26/I26
C7	0.83	0.655	0.37-1.86	0.81	0.609	0.36-1.83	B26/I26
Overall	0.59	0.150	0.28-1.21	0.28	0.001	0.13-0.59	B26/I26

Table 5. Ordinal logistic regression results for reconstruction kernels. Most preferred option is summarized in the last column. Significant ($p<0.05$) values are indicated with green color.

	FBP	Safire 2	Safire 3
B26/I26	3.12	3.22	3.46
B36/I36	3.20	2.97	3.29
B46/I46	2.74	2.93	3.15

Table 6. Overall mean scores for for the different kernels and reconstruction options. Highest value is indicated with green color.

All criteria were related to an overall image quality. The line *overall* refers to the mean of the scores given averaging the scores over the criteria. As mentioned before, calculating VGAS is a very basic approach, but it is still frequently used and to give a complete picture Table 6 presents the mean values for this overall image quality score. Due to the shortcomings of this simple model, it would be misleading to present p-values and confidence intervals for this model.

3. Results

Tables 4 and 5 show that I26 and Safire 3 were the most preferred reconstruction options. Safire 3 performed significantly better than Safire 2 or the baseline FBP reconstruction. On the other hand, B26/I26 performed best in all but one case. However, difference between B26/I26 and B36/I36 was not statistically significant in any of the cases. In general B46/I46 performed worse than the other two options, and in half of the cases it was statistically significantly worse. These results are in agreement with the simpler VGAS in the Table 6.

The logistic model assumes that there is no interaction between the covariates. Due to the number of observations in this study, interaction between the covariates could not be investigated to get statistically significant results for the large number of free parameters.

In addition to the regression model, Spearman's rank order correlation was used to investigate the responses to the different criteria and to analyze the agreement between readers. Correlation plot in Fig. 1 shows that scores for all of the criteria are strongly significantly correlated ($p<0.001$) with each other. However, the linear correlation coefficients between the criteria are varying, and especially low for criterion C7.

Similarly, the Spearman's rank order correlation between the readers' responses are presented in Fig. 2.

4. Discussion

4.1. Image quality

I26 with Safire 3 was the most preferred technique in this study for all criteria and for the overall image quality. The difference between Safire 3 and other options were statistically significant for some criteria (C1,C2,C3, C6, and

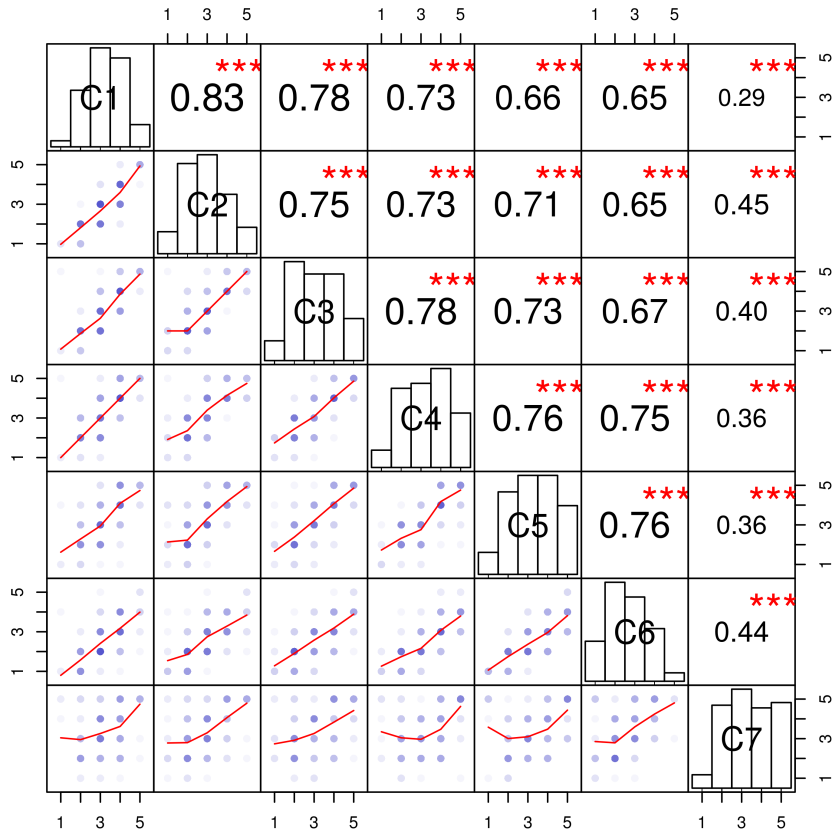

Figure 1. Spearman correlation of the scores between the criteria. Distribution of the scores are plotted with blue density maps with local curve fit. Significance levels are: p<0.05(*), p<0.01(**), p<0.001(***).

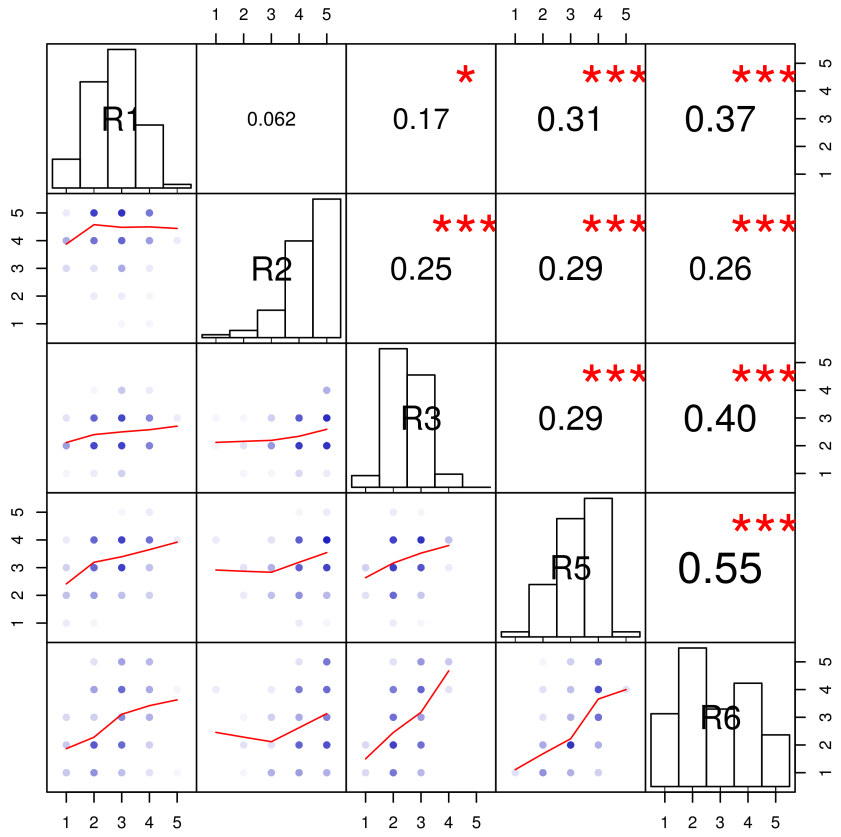

Figure 2. Spearman correlation of the scores between the readers. Distribution of the scores are plotted with blue density maps with local curve fit. Significance levels: p<0.05(*), p<0.01(**), p<0.001(***).

overall), the difference between the B26/I26 and B36/I36 were not statistically significant. The magnitude of the differences are easier to observe in Table 6 for VGAS. The difference in favor of Safire 3 is consistent but small. Evaluation of the clinical relevance of this difference is beyond the scope of this paper, and might be minimal.

The importance of sharpness and noise level depends on the diagnostic task. For the criteria in this study B26/I26 was slightly, but not significantly, better than the sharper B36/I36 kernel. For other criteria B36/I36 might perform better than B26/I26.

This study did not aim to determine the best possible kernel for a general post mortem cardiac study, but rather confirm the applicability of an existing clinical protocol in a forensic setup. If the baseline method would perform worse than any of the alternatives, then it would indicate this protocol couldn't be derived from clinical protocol, and definitely would need a dedicated protocol optimization study.

The results imply that in this case there is no indication of sub-optimal performance. Presence of foreign objects, sever tissue degradation or other substantial differences from clinical volumes might require different protocols.

4.2. Response time analysis

There was no time constraint set for the reading session, but response times were recorded for each displayed volume each readers. and second-round response times for reader #4, who wanted to restart the study after a few volumes to correct her answers.

The response time graph (Fig. 3) is included for two reasons. First, it is a legitimate concern that readers became tired during the study and they lost attention. A linear fit after the fifth response shows that average the response times slightly decreased as the study proceeded. This change is approximately 0.5sec/volume which is statistically not significant, and arguably negligible. Second reason for the graph is to demonstrate that the readers used similar amount of time even though they were not constrained. The response time for the very first task was excluded for all readers because it contained the preparation time too.

4.3. Outliers

One reader gave counter-intuitive responses for the last criterion. Does inclusion or exclusion of these responses change the final results? The above analysis included all the responses. Excluding these presumed outliers would slightly decrease the p-values, but would not change significance. We did not consider them influential and rather kept the original data set.

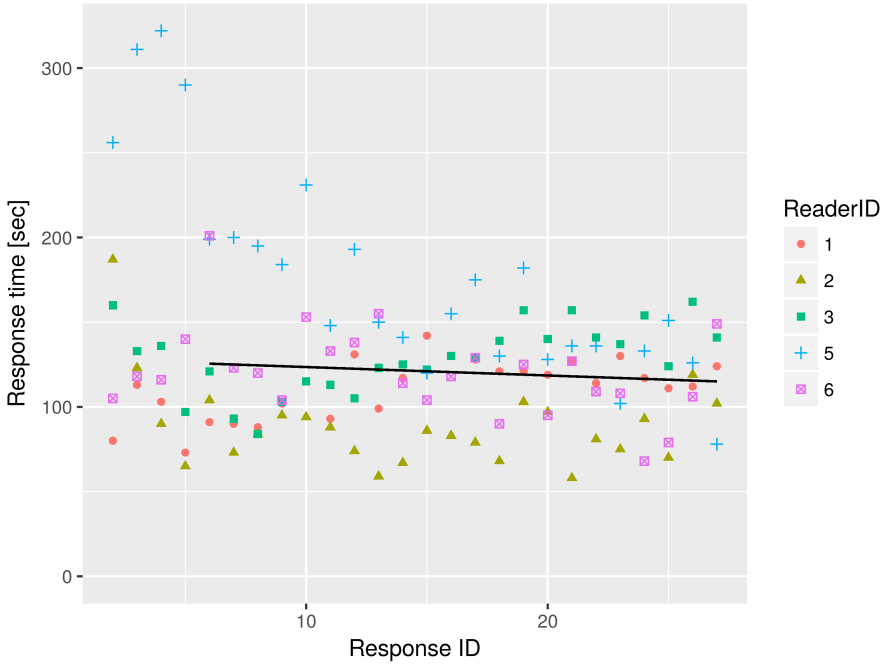

Figure 3. Readers' response time. After a few try initial grading, the response times stabilized. The decrease in response time is not significant (p=0.321).

4.4. Inter-observer differences

Both the response time analysis and the outlier points indicate some inter-observer differences. This difference might originate from the challenging evaluation of forensic volumes, but it also might be an underlying difference between the readers. We felt important to disclose this limitation of the study, and presented it in the most concise form in Fig. 2. Note that reader 4 was removed from study, as it was mentioned in the Materials and methods section.

4.5. Future work

Visual grading analysis has known short-comings. Most notably, VGA requires large number of responses to differentiate between similar quality images. The readers also might change their preference scale during the reading session which is known as adaptation[17]. To overcome these limitations, pairwise comparisons[18] (PC) might follow VG sessions. While PC makes easier to choose the best from two options, the number of comparisons increase quadratically with the number of options. Similarly to VG, strict criteria are required for PC to avoid being a „beauty contest'' and other systematic errors.

5. Conclusion

The study found the same kernel and iterative reconstruction (I26 Safire 3) optimal in forensic CT as in clinical use, despite the differences between the clinical and forensic setups.

Acknowledging the limitations of the study in the discussion part, there is no indication that applying a clinical protocol in post mortem scans would change the readers' preferred reconstruction, as long as the criteria are the same and the post mortem alterations are not sever. Statistically significant difference in preference does not necessarily mean difference in diagnostic performance.

References

[1] C. O'Donnell, N. Woodford, Post-mortem radiology—a new sub-speciality?, Clinical Radiology. 63 (2008) 1189–1194. doi:10.1016/j.crad.2008.05.008.

[2] C. Jackowski, W. Schweitzer, M. Thali, K. Yen, E. Aghayev, M. Sonnenschein, P. Vock, R. Dirnhofer, Virtopsy: Postmortem imaging of the human heart in situ using MSCT and MRI, Forensic Science International. 149 (2005) 11–23. doi:10.1016/j.forsciint.2004.05.019.

[3] L.L. Geyer, U.J. Schoepf, F.G. Meinel, J.W. Nance, G. Bastarrika, J.A. Leipsic, N.S. Paul, M. Rengo, A. Laghi, C.N.D. Cecco, State of the Art: Iterative CT Reconstruction Techniques, Radiology. 276 (2015) 339–357. doi:10.1148/radiol.2015132766.

[4] X. Zheng, T.M. Kim, R. Davidson, S. Lee, C. Shin, S. Yang, CT x-ray tube voltage optimisation and image reconstruction evaluation using visual grading analysis, in: B.R. Whiting, C. Hoeschen (Eds.), Proceedings of SPIE - The International Society for Optical Engineering, 2014: p. 903328. doi:10.1117/12.2043201.

[5] M.K. Kalra, M.M. Maher, T.L. Toth, L.M. Hamberg, M.A. Blake, J.-A. Shepard, S. Saini, Strategies for CT Radiation Dose Optimization, Radiology. 230 (2004) 619–628. doi:10.1148/radiol.2303021726.

[6] F. Zarb, L. Rainford, M.F. McEntee, Developing optimized CT scan protocols: Phantom measurements of image quality, Radiography. 17 (2011) 109–114. doi:10.1016/j.radi.2010.10.004.

[7] L.G. Månsson, Methods for the Evaluation of Image Quality: A Review, Radiation Protection Dosimetry. 90 (2000) 89–99. doi:10.1093/oxfordjournals.rpd.a033149.

[8] J.F. Barrett, N. Keat, Artifacts in CT: recognition and avoidance., Radiographics : A Review Publication of the Radiological Society of North America, Inc. 24 (2004) 1679–1691. doi:10.1148/rg.246045065.

[9] H. Menzel, H. Schibilla, D. Teunen, European guidelines on quality criteria for computed tomography, Luxembourg: European Commission. (2000).

[10] M. Båth, L.G. Månsson, Visual grading characteristics (VGC) analysis: A non-parametric rank-invariant statistical method for image quality evaluation, British Journal of Radiology. 80 (2007) 169–176. doi:10.1259/bjr/35012658.

[11] Ö. Smedby, M. Fredrikson, Visual grading regression: Analysing data from visual grading experiments with regression models, British Journal of Radiology. 83 (2010) 767–775. doi:10.1259/bjr/35254923.

[12] A.D. Hardie, R.M. Nelson, R. Egbert, W.J. Rieter, S.V. Tipnis, What is the preferred strength setting of the sinogram-affirmed iterative reconstruction algorithm in abdominal CT imaging?, Radiological Physics and Technology. 8 (2015) 60–63. doi:10.1007/s12194-014-0288-8.

[13] M. Håkansson, S. Svensson, S. Zachrisson, A. Svalkvist, M. Båth, L.G. Månsson, ViewDEX: An efficient and easy-to-use software for observer performance studies, Radiation Protection Dosimetry. 139 (2010) 42–51. doi:10.1093/rpd/ncq057.

[14] M. Båth, Evaluating imaging systems: Practical applications, Radiation Protection Dosimetry. 139 (2010) 26–36. doi:10.1093/rpd/ncq007.

[15] R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2014. http://www.r-project.org/.

[16] W.N. Venables, B.D. Ripley, Modern Applied Statistics with S, Fourth, Springer, New York, 2002. http://www.stats.ox.ac.uk/pub/MASS4.

[17] H. Helson, Adaptation-level theory, Harper & Row, New York, 1964. http://psycnet.apa.org/psycinfo/1965-00332-000.

[18] L.L. Thurstone, A law of comparative judgment., Psychological Review. 34 (1927) 273–286. doi:10.1037/h0070288.