Britta König*, Nika Guberina, Hilmar Kühl, and Waldemar Zylka

Validation of iterative CT reconstruction by inter and intra observer performance assessment of artificial lung foci

https://doi.org/10.1515/cdbme-2020-3137

Abstract: We investigate the suitability of statistical and model-based iterative reconstruction (IR) algorithm strengths and their influence on image quality and diagnostic performance in low-dose computer tomography (CT) protocols for lung–cancer screening procedures. We evaluate the inter– and intra–observer performance for the assessment of iterative CT reconstruction. Artificial lung foci shaped as spheres and spicules made from material with calibrated Hounsfield units were pressed within layered granules in lung lobes of an anthropomorphic phantom. Adaptively, a soft–tissue– and fat–extension ring were attached. The phantom with foci was scanned using standard high contrast, low-dose and ultra low-dose protocols. For reconstruction the IR algorithm ADMIRE at four different strength levels were used. Two ranking tests and Friedman statistics were performed. Fleiss κ and modified Cohen's κ_{ney} were used to quantify inter– and intra–observer performance. In conjunction with the standard lung kernel BL75 radiologists evaluated medium to high IR strength, with preference to S_4, as suitable for lung foci detection. When varying reconstruction kernels the ranking became more random than with varying phantom diameter. The inter–observer reliability shows poor to slight agreement expressed by $\kappa < 0$ and $\kappa = 0 - 0.20$. For the intra-observer reliability non–agreement with $\kappa_{ney} = 0 - 0.20$ and moderate agreement with $\kappa_{ney} = 0.60 - 0.79$ for the first ranking test, and almost perfect agreement with $\kappa_{ney} > 0.90$ for the second ranking test was observed. In conclusion, our validation suggests radiological preference of medium to high iteration strengths, especially S_4, for lung foci detection. An investigation of the correlation between diagnostic experience and the subjective perception of IR reconstructed CT images still needs to be investigated.

***Corresponding author: Britta König,** Westphalian University, Campus Gelsenkirchen, Germany, and University of Duisburg-Essen, britta.koenig@stud.uni-due.de
Nika Guberina, University of Duisburg-Essen, University Hospital Essen, Germany, nika.guberina@uk-essen.de
Hilmar Kühl, St. Bernhard-Hospital Kamp-Lintfort GmbH, University of Duisburg-Essen, Germany, hilmar.kuehl@uni-due.de
Waldemar Zylka, Westphalian University, Campus Gelsenkirchen, Germany, waldemar.zylka@w-hs.de

Keywords: CT, iterative reconstruction, lung nodule detection, inter– and intra–observer reliability, low–dose, image quality, phantom

1 Introduction and Background

The Belgian-Dutch randomized-controlled NELSON Trial in 2017 demonstrated a reduction of lung cancer mortality with low-dose computer tomography (LDCT) for high-risk patients over a 10-year period: men by 26% and women by 61% [1]. Previously the US Lung Cancer Screening Trial NLST in 2011 showed a possible relative risk reduction of dying of lung cancer in the risk group by 20% by performed LDCT screening procedures, which corresponds to an absolute risk reduction of 0.3% [2]. High contrast LDCT-protocols can be used for lung foci detection. Established filtered back projection (FBP) as an analytical reconstruction method generates reduced image quality with LDCT. To meet high quality standards in terms of dose reduction and image quality, manufacturers created multiple solutions: Reduction of tube voltage, automatic tube current modulation and iterative reconstruction (IR) as the most advanced techniques [3]. In this investigation the statistical, model–based IR algorithm ADMIRE (Advanced Modeled Iterative Reconstruction) was used [3, 4].

IR reconstruction may outperform traditional analytical methods as image impression alters with increasing algorithm strength. In fact, in [5] has been reported that radiologists have reservations with regard to IR reconstructed image and its possible influence on diagnostics. Furthermore, radiologists evaluated IR images rather different which may be due to professional experience and could mirror their accustoming. We address these issues in this paper and report on a statistical analysis of an inter– and intra–observer performance assessment of artificial lung foci in IR reconstructed CT images.

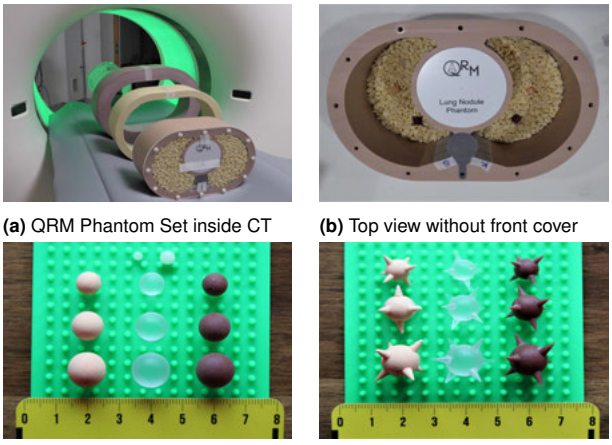
2 Materials and methods

Anthropomorphic Phantom. The commercially-available anthropomorphic QRM Lung-Nodule Phantom Set including extension rings (QRM GmbH, Möhrendorf, Germany) was

used to simulate an adult human chest, displayed in Fig.1(a). The adjustable anatomical model consists of components with calibrated Hounsfield values (HU) of human tissue. In detail, two lung lobes fill-able with lung granules, mediastinum and spine with the soft– tissue section of a chest wall. For material replacements and lung nodule positioning a front– and back–cover can be screwed. Furthermore, varying chest diameters can be simulated by fitting soft– tissue– and fat–extension rings (effective diameters = 25/ 30 cm). Replicated lung foci shaped as spheres and spicules were inserted to the phantom, see Fig. 1(b-d). We performed a total of three setups, one with the phantom body and one each with fitted soft tissue and fat ring.

CT protocol and reconstruction. The entire raw data acquisition was performed on a Somatom Force CT, further specifications in [5]. For each setup three CT dose protocols were selected: (i) standard high contrast (SHC; 120kV/51mAs), (ii) low-dose (LD; 120kV/40mAs) and (iii) ultra-low-dose (ULD; 120kV/20mAs). Raw data were acquired using FOV 430mm, rotation time 0.5s, delay 2s, beam collimation 192x0.6mm, deactivated CAREDose4D, slice thickness 1mm, increment 1mm. Preparations scans and reconstruction were performed as in [5]. The acquired raw data of each protocol were reconstructed by utilizing three kernels: BL57 is a standard for lung node detection, BR32 is a soft, and BR69 is hard kernel. The images were reconstructed using ADMIRE $S_1/S_3/S_4/S_5$ as axial slices of 5 mm thickness, 5 mm reconstruction increment with parenchymal lung window (-600/1200HU). Sixty CT image series of twenty-five slices were acquired. Subsequent test procedures used the seventeenth slice of all image series. Each CT image was anonymized. Additionally, lung sections were selected on the thirty-six images to hide varying phantom diameters.

Evaluation methods. In order to rank CT images reconstructed by ADMIRE $S_1/S_3/S_4/S_5$ according to their suitability for lung node detection, the subjective differential tests frequently used in sensor technology were implemented [6]. Specifically, two ranking test were carried out. The first test (ranktest1) contains images of the phantom body with attached fat ring, varying kernels and LDCT protocols. Pursuant to a ranking test by four investigation attributes, four (out of the thirty-six) randomly selected images were randomly placed in a four field graphic user interface (4GUI). This arrangement was presented to a radiologist, who was asked to rank the images in a descending order from rank 1 to rank 4 according to his perception. One test set contained nine 4GUIs. Six radiologists evaluated three test sets. For analyzing the determined rankings, we add up the set ranks for each investigation attribute over all 4GUIs and get the rank sums (rs). Calculated Friedman values (F-values) were compared to the approximate critical value 7.81 (which is the probability of error at level $\alpha = 5\%$ when the number attributes is $k = 4$) to verify the randomness of the ranking. To assess inter– and intra–observer reliability we computed Fleiss κ and a modification of Cohens κ_{ney}. The second ranking test (ranktest2) followed the same procedure but contained images of the phantom body only, with each ring attached, fixed BL57 and varying LDCT protocols.



(a) QRM Phantom Set inside CT **(b)** Top view without front cover

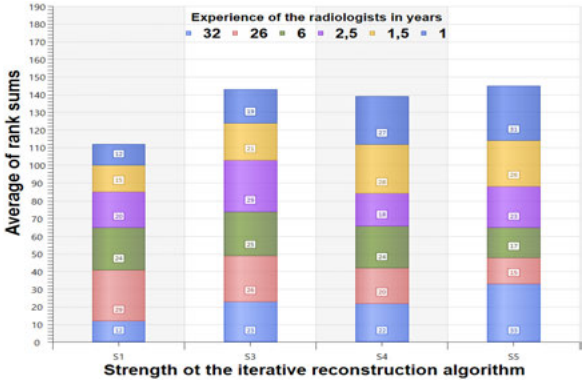
(c) Artifical lung foci, spheres **(d)** Artifical lung foci, spicules

Fig. 1: Phantom within the CT-Gantry (a), targets made from material with calibrated (at 120kV) HU values equal to -690/-50/+100 with color assignment beige/brown/opaque inside phantom on 25mm layered granules: left lung spicules; right lung spheres (b), and top view of the spherical (diameters 3/5/8/10/12mm) (c) and the speculated targets (diameters 16/20/24mm) (d).

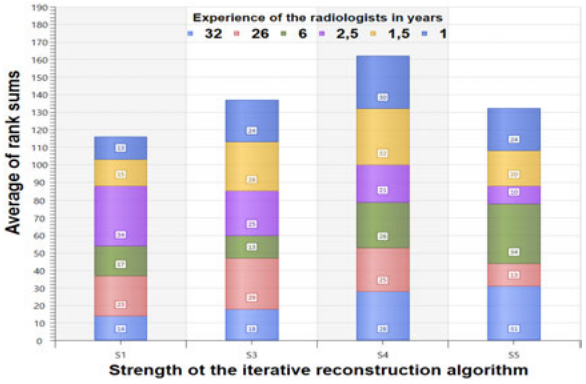
3 Results

Averaged rank sums ($\overline{rs}$) over all test sets are displayed in Fig. 2. The collective result of ranktest1 demonstrate that medium to high S ($\geq S_3$) were generally rated as higher $\overline{rs}$ than lower S ($\leq S_3$). S_5 yields with $\overline{rs} = 145$ ($> S_3$ with 143; $> S_4$ with 139) the highest collective $\overline{rs}$ and S_1 with $\overline{rs} = 112$ the lowest. The assessment of the radiologist with 26 years of professional practice (ypp) does not follow the collective result, whereas the most (32 ypp) and less experienced observers (1 ypp) rated in accordance with the overall result. Collective $\overline{rs}$ of ranktest2 shows the highest amount with $\overline{rs} = 162$ for S_4 ($> S_3$ with 137; $> S_5$ with 132;) and the lowest for S_1 with $\overline{rs} = 116$. The averaged rank sum $\overline{rs}$ of ranktest2 indicate that medium to high IR strength is considered more suitable for the detection of lung foci. Furthermore, a 2.5ypp radiologist ranking is entirely contrary to the collective result, while 1 and 1.5 ypp observers assessments are in line with the collective $\overline{rs}$ result. The most experienced observers (6 and 31 ypp)

rated similarly to the collective $\overline{rs}$ with a preference for S_5 reconstruction strength.



(a) Average of rank sums of the first test (ranktest1).



(b) Average of rank sums of the second test (ranktest2).

Fig. 2: Averaged rank sums $\overline{rs}$ based on the rank position of the order (rank) of the images order for (a) ranktest1 and (b) ranktest2. Radiologists are represented by ypp and color.

The F–values listed in Tab. 1 were calculated based on rs of each 4GUI ranking per set for every radiologist. Concerning ranktest1, F–values for radiologists of 1 and 36ypp are higher than the approximated critical value $7.81\,(\alpha = 5\%; k = 4)$. All other radiologists ranked at least once by random, in total seven of eleven F–values are less than 7.81. Fewer random rankings were found in ranktest2, except in set 1 by radiologist with 1ypp. In summary, the rankings of ranktest1 were more at random than those of ranktest2.

Fleiss κ as statistical measure for inter–observer reliability for each test set and ranking position of all radiologists is listed in Tab. 2. For the ranktest1 $\kappa < 0$, except for set 1 (Pos. 3) and set 2 (Pos. 4). Concerning ranktest2 we get $\kappa < 0$. According to [6] $\kappa < 0$ indicate poor and $\kappa = 0 - 0.20$ slight agreement. The results represent a differential observers perception in terms of the suitability of S for lung foci detection.

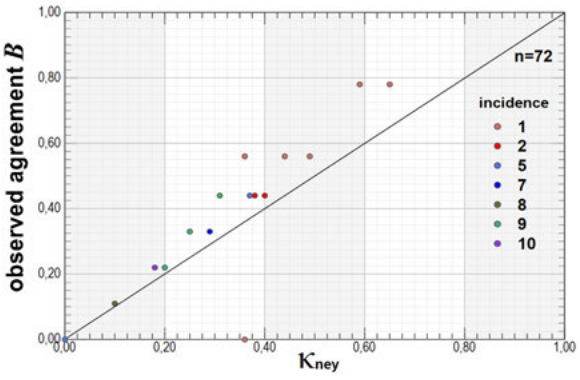
Tab. 1: Friedman values calculated of the rs of all rankings per set of each radiologist.

radiologist's experience	F–values					
	ranktest1			ranktest2		
[ypp]	Set 1	Set 2	Set 3	Set 1	Set 2	Set 3
1	16,33	14,20	14,73	5,93	14,47	13,40
1,5	10,74	13,93	4,60	11,40	23,13	21,67
2,5	7,27	3,27	8,87	24,60	18,73	16,60
6	3,93	9,40	5,40	24,33	13,67	16,60
26	3,80	8,60	13,67	12,60	12,47	15,80
36	21,93	16,20	23,27	10,33	18,73	21,40

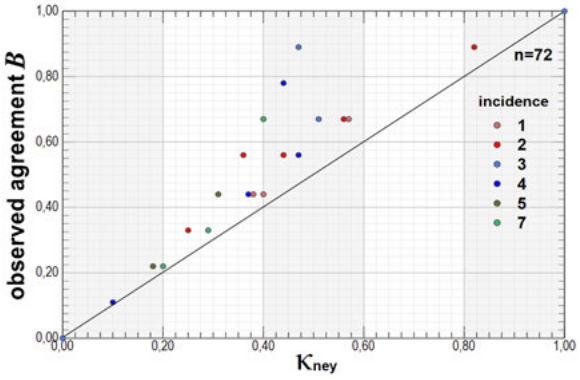
Tab. 2: Inter–observer agreement with Fleiss κ for each ranking position and set.

Ranktest	Set	Fleiss κ			
		Pos. 1	Pos. 2	Pos. 3	Pos. 4
1	1	-0,014	-0,067	0,002	-0,019
	2	-0,073	-0,057	-0,035	0,038
	3	-0,067	-0,032	-0,089	-0,052
2	1	-0,089	-0,083	-0,021	-0,019
	2	-0,119	-0,092	-0,056	-0,105
	3	-0,102	-0,089	-0,092	-0,137

In order to measure the intra–individual variability of the assessed image positions over all sets per observer at different times, we calculated a modified Cohen's κ [7]. It is well known that the interpretation of Cohen's κ does not take into consideration problems with this measure, which may lead to misleading conclusions [8]. Usually Cohen's κ is calculated by the randomly expected agreement E and the observed agreement B according to $\kappa = (B - E)/(1 - E)$. Instead of E we used its modification C and receive the modified κ_{ney}. In C only the parts of the marginal distributions of the Cohen's agreement matrix that also have a part of B are taken into consideration. In Fig. 3 B is plotted as function of κ_{ney}, i.e. $B(\kappa_{ney})$. It contains n=72 comparisons per test. In ranktest1 κ_{ney} ranges between 0 and 0.65 and $B = 0 - 0.78$ and max. incidence was 10. For ranktest2 $\kappa_{ney} = 0 - 1.00$ while $B = 0 - 0.89$ and max. incidence was 7. The agreement given by κ_{ney} results for both tests be interpreted as follows: values $0.60 - 0.79$ indicating moderat as best assessment for ranktest1, values $0 - 0.20$ as non occurs less in ranktest2 compared to ranktest1, values $0.80 - 0.90$ as strong and > 0.90 as almost perfect were the highest ratings for ranktest2, interpretation based on [8]. In total intra–observer reliability was higher for ranktest2 than for ranktest1.



(a) Observed agreement $B(\kappa_{ney})$ for ranktest1.



(b) Observed agreement $B(\kappa_{ney})$ of ranktest2.

Fig. 3: The function of the observed agreement $B(\kappa_{ney})$ for (a) ranktest1 and (b) ranktest2, including color assignment for incidence and n=72 comparisons.

4 Discussion and Conclusion

Ranking tests of iterative CT reconstructions show that radiologists consider medium to high IR strength more suitable for lung foci detection. If clinically applied parameters were used (BL57, varying phantom diameter), the experiment reveals a noticeable preference of S_4. The correlation of years of professional experience and the subjective evaluation of IR reconstructed images of various strength cannot be verified accurately. The most and less experienced observers assessed similar, differences in fields of expertise and diagnostic accustoming give further reason to test more observers in order to make a confident statement. The influence of varying kernels (fixed phantom diameter) with IR could be the reason why more random rankings were obtained, whereas more clinically conditions generated only one random ranking. It was found in [5] that radiologists rate BR69 and BR32 worse than BL57 while evaluating anonymized and randomized single images, so we expected a reliable ranking even with varying kernels. Potentially the ranking test method is more challenging for radiologists to answer than a dichotomous categorical assessment or multiple-stage categorical assessment. As suspected the inter–observer reliability shows a poor to slight agreement with regard to the ranking procedure. This is an indicator for individual subjective judgments. Negative signs according to Fleiss κ statistics may indicate systematic errors. For instance, this ranking tasks are related only to the detection of for lung foci and none strict assessment criteria were applied. Due to fields of expertise two radiologists evaluate according to their own criteria, which could have a systematic impact on Fleiss κ. Other potential errors may by due to the use of a not standardized monitor. For the intra–observer reliability the observed agreement $B(\kappa_{ney})$ has wide variability. All observers have confidently selected rank positions 1 and 4, i.e. the best and the worst image. The intra–observer reliability for the intermediate rank positions 2 and 3 was significantly lower.

In conclusion, our ranking tests suggest radiological preference of medium to high iteration strengths, especially S_4, for lung foci detection. We hypothesise, that in subjective image quality analysis the assessment habits of radiologists are dichotomous- or multiple stage–categorical. An investigation on the correlation between diagnostic experience and the subjective perception of iterative reconstructed CT images is mandatory.

Author Statement

Research funding: The author state no funding involved. Conflict of interest: Authors state no conflict of interest.

References

[1] De Koning H, Van Der Aalst C, Ten Haaf K, et al. Effects of volume CT lung cancer screening: Mortality results of the NELSON randomized-controlled population based trial. Journal of Thoracic Oncology 2018;13:185

[2] Center for Statistical Sciences, Brown University,Providence, United States of America. The National Lung Screening Trial: Overview and Study Design. Radiology 2011;258:243-253.

[3] Siemens Healthineers, Erlangen, Germany.

[4] Beister M, Kolditz D, Kalender W A. Iterative reconstruction methods in X-ray CT. Physica Medica 2012;28:94–108.

[5] König B, Guberina N, Kühl H, Zylka W. Design and first results of a phantom studyon the suitability of iterative reconstruction for lung-cancer screening with low-dose computer tomography. Biomedical Engineering 2018;5:593-596.

[6] Landis J R, Koch G G. The Measurement of Observer Agreement for Categorical Data. Biometrics 1977;33:159-174.

[7] Kutschmann M.Private communication 2020.

[8] Wirtz M, Kutchmann M. Analyse der Beurteilerübereinstimmung für Kategoriale Daten mittels Cohens Kappa und alternativer Maße. Rehabilitation 2007;46:1-8.

[9] McHugh M L. Interrater reliability: the kappa statistic. Biochemia Medica 2012;22:276-282.