BMC Medical Imaging

RESEARCH ARTICLE Open Access

Interobserver variability in quality assessment of magnetic resonance images

Rafal Obuchowicz[1], Mariusz Oszust[2] and Adam Piorkowski[3*]

Abstract

Background: The perceptual quality of magnetic resonance (MR) images influences diagnosis and may compromise the treatment. The purpose of this study was to evaluate how the image quality changes influence the interobserver variability of their assessment.

Methods: For the variability evaluation, a dataset containing distorted MRI images was prepared and then assessed by 31 experienced medical professionals (radiologists). Differences between observers were analyzed using the Fleiss' kappa. However, since the kappa evaluates the agreement among radiologists taking into account aggregated decisions, a typically employed criterion of the image quality assessment (IQA) performance was used to provide a more thorough analysis. The IQA performance of radiologists was evaluated by comparing the Spearman correlation coefficients, ρ, between individual scores with the mean opinion scores (MOS) composed of the subjective opinions of the remaining professionals.

Results: The experiments show that there is a significant agreement among radiologists ($\kappa = 0.12$; 95% confidence interval [CI]: 0.118, 0.121; $P < 0.001$) on the quality of the assessed images. The resulted κ is strongly affected by the subjectivity of the assigned scores, separately presenting close scores. Therefore, the ρ was used to identify poor performance cases and to confirm the consistency of the majority of collected scores ($\rho_{mean} = 0.5706$). The results for interns ($\rho_{mean} = 0.6868$) supports the finding that the quality assessment of MR images can be successfully taught.

Conclusions: The agreement observed among radiologists from different imaging centers confirms the subjectivity of the perception of MR images. It was shown that the image content and severity of distortions affect the IQA. Furthermore, the study highlights the importance of the psychosomatic condition of the observers and their attitude.

Keywords: Radiologists, Quality perception, Fleiss' kappa, Decision process

Background

The perception of pathologies in the displayed medical resonance (MR) images is often subjective and thus may lead to false-negative errors [22, 24]. Therefore, many clinical studies have been carried out to evaluate the radiological expertise as a part of clinical decision making [25, 32]. Consequently, factors which influence the perception became a matter of scientific discussion, resulting in the foundation of the Medical Imaging Perception Society (MIPS). The society encourages and promotes medical image perception research and education. Such research involves an investigation of physical, social, and behavioral aspects which affect decision-making performance of imaging specialists. Hence, image-dependent and independent factors which strongly influence the perception were identified [11, 38]. They are associated with image creation and processing [20]. A consensus was reached that the best possible resolution and contrast should be ensured to provide an opportunity to recognize anatom-

*Correspondence: pioro@agh.edu.pl
[3]Department of Biocybernetics and Biomedical Engineering, AGH University of Science and Technology, Mickiewicza 30, 30-059, Cracow, Poland
Full list of author information is available at the end of the article

ical and pathological structures [15, 35]. However, such conditions cannot be met in practice. Therefore, well-defined matrices used in the daily calibration of diagnostic displays are often used to address the quality of displayed image content [31, 34]. To facilitate image interpretation and improve diagnostic performance, display hardware, viewing software, and reading environment are provided in a radiology reading room [17, 41, 44].

Also, a lot of effort was put to define image-independent factors, which are semantic in nature and related to the psychosomatic and sociological aspects of the observed images [5, 21]. They also affect the performance of the cognitive tasks in presence of changes in images [4].

Since distortions are perceived by radiologists it is worth examining the degree of their agreement on the quality of assessed images and determine whether radiologists similarly perceive the quality. To the best knowledge of the authors, the interobserver variability regarding the quality assessment of MR images has not been addressed in the literature. In the existing studies, the discussion mostly covers decisions involving the risk of malignancy based on other than MR imaging methods. For example, in recent works of Pang et al. [30] and Buda et al. [3], the presence of malignancy in ultrasound images and subsequent recommendations were considered. A more developed study presented by Williams et al. [43] involved a subjective assessment of computed tomography coronary angiogram images. In that work, noisy images were used to determine the agreement among radiologists on the diagnosis of angina pectoris due to coronary heart disease for stenosis severity. Sweeney et al. [39], reviews mammographic positioning image quality criteria being the results of years of discussion on the influence of image quality on the detection of breast cancer. Such criteria have been established taking into account observer variability. Performance of radiologists in the identification of cancer cases in mammography images was studied by Rafferty et al. [33].

This study aims at the assessment of a representative group of radiologists in the quality evaluation of MR images. The considered images contain authentic distortions (i.e., they were not artificially introduced) and allow investigating the interobserver agreement among clinicians. The scores for images are also used to determine the individual performance of a clinician using the Spearman rank correlation coefficient, ρ. The ρ is typically employed to evaluate automatic methods for image quality assessment [37, 41]. This study gives important insight on the variance of the perception of image characteristics in the presence of noise of the group of experienced professionals.

Methods

Data collection

The study was performed on a group of 31 radiologists with experience in diagnostic images reading. All medical professionals completed at least 6 years of residency. They are used to work on 1.5T MRI scanners. The study took place in a controlled environment, inside of a lecture room with a limited luminance not interfering with images displayed on monitors. For displaying purposes, Eizo monitors (RadiForce 250) connected to PC computers equipped with dedicated graphics processors (Eizo Quadro) were used. Each observer was equipped with a diagnostic unit and assessed 35 cases (70 images) without interference from other radiologists using grades 1, 2, 3, 4, and 5 which correspond to 'bad,' 'poor,' 'fair,' 'good,' and 'excellent' image quality, respectively [14, 40, 42]. The scale of the grades is accepted by the Video Quality Experts Group [40] and is widely used in image quality assessment research [14, 42]. In the presented study, at the beginning of the experiment, two images of the best and worst quality were shown and the grading system was explained. The images were presented simultaneously on all monitors for one minute. Each case consisted of two images of a body structure differing in quality (the double stimulus approach [42]). The participants wrote scores on paper forms to ensure the anonymity of the answers. Then, scores were averaged to obtain the mean opinion score (MOS). The following structures were displayed in different planes: the lumbar and cervical spine (14 images), knee (14), shoulder (16), wrist (6), hip (4), pelvis (4), elbow (2), ankle (2), and brain (8).

The study protocol was designed according to the guidelines of the Declaration of Helsinki and the Good Clinical Practice Declaration Statement. Special care was taken regarding personal data safety, where all the images were anonymized before processing. Written acceptance for conducting the study was obtained from the Ethics Committee of Jagiellonian University (no. 1072.6120.15.2017). Data of 51 patients, 26 men and 25 women, in the age group of 27-41 years, were enrolled in the study. The criteria of negative selection were the image artifacts influencing the image analysis. T2-weighted sagittal sequences of selected body parts were analyzed. To routinely conduct MR studies aiming at decreasing image quality, shortened sequences were made using parallel imaging I PAT software (Siemens). The functionality was implemented using GeneRalized Autocalibrating Partially Parallel Acquisitions (GRAPPA) which resulted in 1.5 min added to the initial exam on the average. Specifically, the GRAPPA 3 was used in which 25% of the echoes were acquired with 60% signal reduction [10]. As a result of the reduced amount of the input data, reconstructed images of the tissue were degraded to lower quality.

The proposed collection was set to represent images of different fields. This is important since the perception of some of them may be different due to the specialization of radiologists in the group (e.g., neuroradiology, gastrointestinal radiology, musculoskeletal radiology,

pediatric radiology). It was assumed that the images of the head and spine are more familiar to most participants than those of the remaining parts of the body. Therefore, images of the knee, foot, or wrist were added to the dataset. This may allow determining whether the familiarity with images influences the subjective perception of their quality.

The same protocol was used to collect subjective scores of three interns. The interns were only instructed on the grading scale without any examples of degraded images. Then, the scores of interns were used for the estimation of their performance, while the scores of experienced radiologists were averaged to obtain the MOS characterizing the images in the dataset.

Exemplary image pairs of different body parts and their scores are presented in Fig. 1. It is worth noticing that the scores reflect a subjective perception of noise and its influence on the displayed body part, i.e., while images of a better quality are similarly scored, the scores of their degraded counterparts are different.

Statistical analysis

Statistical analysis was performed using Matlab [26]. The interobserver variability was assessed using the κ statistic. A Fleiss' κ [13] is related to the Cohen's κ statistic. However, it was used since it measures the consistency of the ratings obtained in tests with more than two observers. The κ of less than 0 indicated poor agreement, 0.01-0.2 slight agreement, 0.21-0.4 fair agreement, 0.41-0.6 moderate agreement, 0.61-0.8 substantial agreement, and 0.81-1 almost perfect agreement. The test statistics were approximated by a normal distribution to calculate the p-value and the 95% confidence interval (CI). Also, since the image quality assessment is considered and the kappa cannot provide a detailed analysis of the individual performance due to the employed aggregation of radiologists' decisions, the Spearman correlation coefficient, ρ, typically used in the IQA field [29, 36, 42], was employed. Subjective scores of a radiologist were compared with the mean opinion score (MOS) calculated as mean scores of the remaining observers to estimate the individual performance.

Results

For the entire dataset, the radiologists achieved a κ of 0.12 (95% CI: 0.118, 0.121; $P < 0.001$), which indicates a slight, but not accidental, agreement. The agreement can also be seen in Fig. 2 in which the number of radiologists assigning a given grade for an image is reported. Only 19 images were assigned the same grade by more than half radiologists. Interestingly, 11 images were assigned two close grades by the same number of radiologists. For example, the image shown in Fig. 1d was assigned '3' and '4' by 10 specialists (cf. no. 22 in Fig. 2). There are also some images with two close scores (e.g., Fig. 1g, image no. 15 in Fig. 2, was graded '4' and '5' by 12 and 13 radiologists, respectively).

To evaluate decisions of radiologists' from image quality perspective, they were correlated with average decisions of the other professionals (Fig. 3). Such an examination takes into account close differences between scores for images instead of aggregated totals used for the calculation of the κ. Consequently, this widely-accepted method for the evaluation of automatic IQA measures was used to provide a more detailed analysis of radiologists' performance. The obtained average, maximum, minimum, and standard deviation of the ρ are 0.5706, 0.8615, -0.4988, 0.3331, respectively. The correlation coefficients reveal a large variability among them, due to weaker or unexpected performances of several specialists. Specifically, the performance of three radiologists affected the results. The negative correlations for 16th and 29th radiologists may evidence their lack of understanding of the used grading system. However, the resulted negative correlations show that they can evaluate the images. More important is the result for the 14th radiologist who seems to assessed images disregarding their quality.

To determine the individual agreements between radiologists, in terms of the IQA, the ρ in pairs was calculated (Fig. 4). The obtained values reflect moderate to the strong correlation of scores in pairs of medical professionals. The lack of agreement of the 14th radiologists with other specialists is also highlighted in this experiment. The findings confirm the previously reported individual results and reveal that most observers' opinions are moderately (to strongly) correlated with those of other professionals.

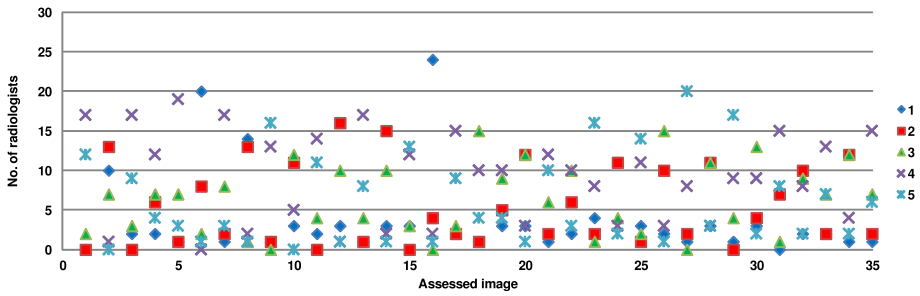
Once the performance of experienced radiologists was evaluated, the IQA performance of three interns who assessed the MR images for the first time was examined. The interns were only instructed on the grading scale. The following results, in terms of the ρ, were obtained: 0.7450, 0.6733, and 0.6419. They confirm that even an inexperienced observer can differentiate the images based on their quality.

Since the dataset contains images of different body parts, the agreements of the radiologists expressed by the κ as well as the ρ were reported (Table 1). In all experiments, the obtained agreements are slight ($\kappa \in (0; 0.2]$) and significant ($P < 0.001$). For parts of the body with two images (i.e., the ankle and elbow), some radiologists assigned them the same grades, preventing the calculation of the ρ. In such cases, the remaining values were averaged. However, mean values for images of separate body parts are close to those obtained for the entire dataset. The reported high maximum values show that the observers' opinions on the image quality were consistent, despite the opposite quality perception of several of them. The last two rows of the table show results for groups of images. Here, frequently examined parts of the body (i.e.,

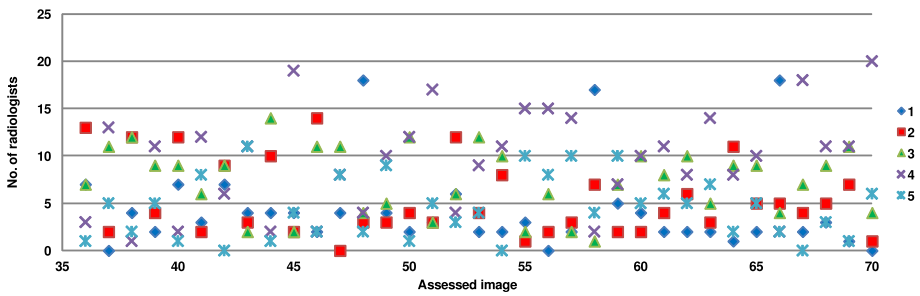
(a) 3.9032 (b) 3.9355 (c) 3.9032 (d) 3.1935

(e) 3.7097 (f) 1.5161 (g) 4.0323 (h) 1.4516

Fig. 1 T2-weighted images and their mean opinion scores. The images of the wrist and knee are of the same quality (cf. (**a**) and (**c**)), despite differences in the structure of the tissue and bones. The appearance of structures was perceived worse for the degraded knee (**e**) than shown in (**d**) due to the plane of the acquisition. Severely degraded images of different body parts were assessed similarly (**f**,**h**)



(a) 1-35



(b) 36-70

Fig. 2 Agreement among radiologists on the image quality expressed by the grades assigned to the images. Images were graded from 1 to 5 by 31 medical professionals

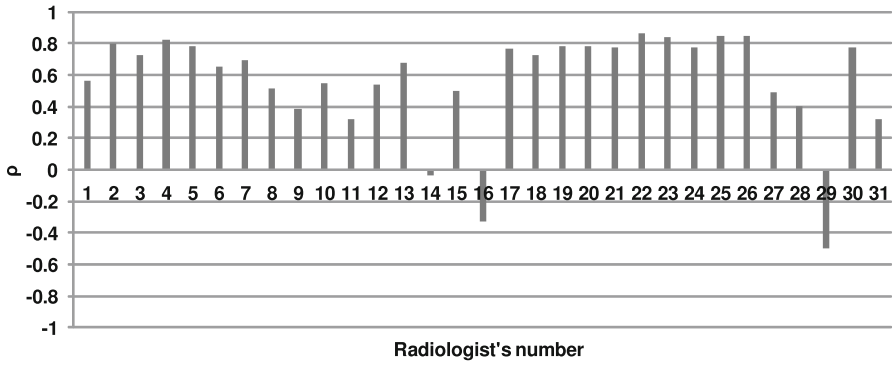


Fig. 3 Evaluation of radiologists' IQA performance in terms of Spearman correlation coefficient (ρ)

head and spine) were considered jointly. Also, the scores for the remaining images were used in calculations. As reported, the performance of radiologists assessing these two groups of images does not vary much and the quality of images was scored similarly even for rarely considered parts of the body in daily work. This is supported by the standard deviation of scores for images shown in Fig. 5 which reveal large differences of scores for some images of the knee and shoulder, closely followed by scores for images of the brain and spine.

Discussion

Perception of the image is fundamental for diagnostic imaging professionals. Radiological training is directed toward critical analysis of the possible abnormalities present in the image. Therefore, diagnostic image assessment relies on the methodological analysis of the displayed content representing human anatomy. However, to perform the analysis a plethora of possible pathological changes as well as anatomical variants should be taken into account. Since the quality perception of radiological images and its relationship with the diagnostic image assessment is seldom addressed in the literature, in this paper, the agreement among professionals on the quality of MR images is studied. The aim of the study was to determine whether the decisions on the quality of a group of radiologists are in agreement. Consequently, this may indicate that the professionals similarly perceive MR images acknowledging the severity of the observed distortions.

This study showed that decisions on the quality are in a slight agreement (κ = 0.12; 95% CI: 0.118, 0.121; P $<$ 0.001). However, due to the subjectivity of the quality assessment and range of scores assigned to the images (1-5), such a result is not surprising. Therefore, the obtained scores were further analyzed using the ρ, which is typically employed for the evaluation of the ability of the automatic image quality assessment techniques to mimic human perception and provide objective scores for images. The radiologists were separately evaluated, and the reported $\rho = 0.5706$ allows concluding that they similarly perceive distortions in MR images.

Furthermore, more detailed tests were also carried out in which familiar images were used jointly. Since the

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
1	1.00	0.68	0.49	0.63	0.51	0.50	0.48	0.28	0.48	0.29	0.13	0.21	0.44	0.06	0.28	-0.19	0.52	0.49	0.49	0.50	0.45	0.56	0.52	0.51	0.47	0.56	0.17	0.13	-0.24	0.52	0.13
2	0.68	1.00	0.64	0.79	0.73	0.66	0.65	0.48	0.42	0.46	0.18	0.43	0.62	-0.10	0.45	-0.18	0.74	0.67	0.63	0.66	0.59	0.71	0.73	0.66	0.72	0.65	0.39	0.25	-0.37	0.70	0.34
3	0.49	0.64	1.00	0.67	0.68	0.51	0.58	0.44	0.24	0.47	0.25	0.29	0.47	-0.07	0.30	-0.41	0.68	0.70	0.66	0.62	0.57	0.70	0.73	0.62	0.74	0.80	0.31	0.23	-0.56	0.67	0.35
4	0.63	0.79	0.67	1.00	0.69	0.62	0.64	0.39	0.39	0.39	0.21	0.38	0.59	-0.10	0.37	-0.21	0.77	0.73	0.63	0.64	0.64	0.78	0.75	0.71	0.74	0.74	0.41	0.30	-0.43	0.70	0.42
5	0.51	0.73	0.68	0.69	1.00	0.56	0.51	0.47	0.34	0.43	0.27	0.38	0.53	0.09	0.46	-0.19	0.69	0.59	0.63	0.68	0.67	0.71	0.69	0.63	0.67	0.65	0.28	0.29	-0.37	0.55	0.46
6	0.50	0.66	0.51	0.62	0.56	1.00	0.57	0.36	0.28	0.40	0.10	0.43	0.51	-0.21	0.30	-0.22	0.64	0.55	0.59	0.61	0.54	0.65	0.63	0.69	0.70	0.60	0.29	0.32	-0.45	0.64	0.18
7	0.48	0.65	0.58	0.64	0.51	0.57	1.00	0.42	0.27	0.40	0.11	0.44	0.51	-0.18	0.46	-0.20	0.71	0.61	0.63	0.67	0.51	0.61	0.75	0.64	0.68	0.72	0.33	0.19	-0.51	0.75	0.31
8	0.28	0.48	0.44	0.39	0.47	0.36	0.42	1.00	0.35	0.37	0.24	0.27	0.43	-0.22	0.28	-0.12	0.45	0.47	0.43	0.34	0.57	0.41	0.44	0.40	0.45	0.45	0.41	0.06	-0.26	0.32	0.24
9	0.48	0.42	0.24	0.39	0.34	0.28	0.27	0.35	1.00	0.19	0.07	0.18	0.48	0.12	0.03	-0.14	0.37	0.39	0.31	0.39	0.35	0.34	0.30	0.27	0.27	0.34	0.31	0.10	-0.09	0.27	0.11
10	0.29	0.46	0.47	0.39	0.43	0.40	0.40	0.37	0.19	1.00	0.29	0.40	0.23	-0.02	0.29	-0.33	0.45	0.43	0.58	0.41	0.42	0.51	0.51	0.48	0.49	0.39	0.43	0.23	-0.30	0.42	0.18
11	0.13	0.18	0.25	0.21	0.27	0.10	0.11	0.24	0.07	0.29	1.00	0.44	0.19	0.08	0.33	-0.20	0.09	0.29	0.45	0.21	0.33	0.27	0.26	0.14	0.21	0.26	0.42	0.08	-0.09	0.24	-0.17
12	0.21	0.43	0.29	0.38	0.38	0.43	0.44	0.27	0.18	0.40	0.44	1.00	0.42	0.00	0.31	-0.03	0.42	0.34	0.53	0.46	0.46	0.47	0.45	0.38	0.41	0.47	0.40	0.33	-0.28	0.50	-0.01
13	0.44	0.62	0.47	0.59	0.53	0.51	0.51	0.43	0.48	0.23	0.19	0.42	1.00	-0.09	0.33	-0.24	0.56	0.51	0.48	0.60	0.54	0.62	0.60	0.56	0.59	0.56	0.58	0.24	-0.26	0.59	0.18
14	0.06	-0.10	-0.07	-0.10	0.09	-0.21	-0.18	-0.22	0.12	-0.02	0.08	0.00	-0.09	1.00	0.08	0.16	-0.11	-0.14	-0.04	0.00	0.03	-0.08	-0.04	-0.11	-0.03	-0.06	-0.06	0.16	0.19	-0.04	0.05
15	0.28	0.45	0.30	0.37	0.46	0.30	0.46	0.28	0.03	0.29	0.33	0.31	0.33	0.08	1.00	-0.17	0.33	0.37	0.42	0.47	0.43	0.48	0.46	0.37	0.41	0.46	0.27	0.21	-0.23	0.46	0.12
16	-0.19	-0.18	-0.41	-0.21	-0.19	-0.22	-0.20	-0.12	-0.14	-0.33	-0.20	-0.03	-0.24	0.16	-0.17	1.00	-0.23	-0.29	-0.39	-0.36	-0.25	-0.35	-0.34	-0.24	-0.36	-0.30	-0.34	-0.07	0.44	-0.29	0.09
17	0.52	0.74	0.68	0.77	0.69	0.64	0.71	0.45	0.37	0.45	0.09	0.42	0.56	-0.11	0.33	-0.23	1.00	0.67	0.62	0.67	0.65	0.74	0.74	0.69	0.74	0.73	0.30	0.36	-0.48	0.68	0.33
18	0.49	0.67	0.70	0.73	0.59	0.55	0.61	0.47	0.39	0.43	0.29	0.34	0.51	-0.14	0.37	-0.29	0.67	1.00	0.69	0.63	0.60	0.73	0.74	0.56	0.62	0.70	0.41	0.26	-0.50	0.68	0.22
19	0.49	0.63	0.66	0.63	0.63	0.59	0.63	0.43	0.31	0.58	0.45	0.53	0.48	-0.04	0.42	-0.39	0.62	0.69	1.00	0.69	0.66	0.68	0.75	0.63	0.69	0.69	0.50	0.24	-0.52	0.67	0.20
20	0.50	0.66	0.62	0.64	0.68	0.61	0.67	0.34	0.39	0.41	0.21	0.46	0.60	0.00	0.47	-0.36	0.67	0.63	0.69	1.00	0.67	0.73	0.70	0.60	0.78	0.74	0.46	0.31	-0.48	0.71	0.31
21	0.45	0.59	0.57	0.64	0.67	0.54	0.51	0.57	0.35	0.42	0.33	0.46	0.54	0.03	0.43	-0.25	0.65	0.60	0.66	0.67	1.00	0.71	0.67	0.61	0.67	0.72	0.39	0.36	-0.31	0.56	0.26
22	0.56	0.71	0.70	0.78	0.71	0.65	0.61	0.41	0.34	0.51	0.27	0.47	0.62	-0.08	0.48	-0.35	0.74	0.73	0.68	0.73	0.71	1.00	0.80	0.75	0.80	0.81	0.47	0.40	-0.45	0.73	0.26
23	0.52	0.73	0.73	0.75	0.69	0.63	0.75	0.44	0.30	0.51	0.26	0.45	0.60	-0.04	0.46	-0.34	0.74	0.74	0.75	0.70	0.67	0.80	1.00	0.77	0.78	0.79	0.38	0.28	-0.52	0.77	0.27
24	0.51	0.66	0.62	0.71	0.63	0.69	0.64	0.40	0.27	0.48	0.14	0.38	0.56	-0.11	0.37	-0.24	0.69	0.56	0.63	0.60	0.61	0.75	0.77	1.00	0.70	0.71	0.33	0.39	-0.42	0.60	0.29
25	0.47	0.72	0.74	0.74	0.67	0.70	0.68	0.45	0.27	0.49	0.21	0.41	0.59	-0.03	0.41	-0.36	0.74	0.62	0.69	0.78	0.67	0.80	0.78	0.70	1.00	0.80	0.49	0.41	-0.52	0.77	0.35
26	0.56	0.65	0.80	0.74	0.65	0.60	0.72	0.45	0.34	0.39	0.26	0.47	0.56	-0.06	0.46	-0.30	0.73	0.70	0.69	0.74	0.72	0.81	0.79	0.71	0.80	1.00	0.36	0.41	-0.49	0.74	0.27
27	0.17	0.39	0.31	0.41	0.28	0.29	0.33	0.41	0.31	0.43	0.42	0.40	0.58	-0.06	0.27	-0.34	0.30	0.41	0.50	0.46	0.39	0.47	0.38	0.33	0.49	0.36	1.00	0.05	-0.33	0.44	0.11
28	0.13	0.25	0.23	0.30	0.29	0.32	0.19	0.06	0.10	0.23	0.08	0.33	0.24	0.16	0.21	-0.07	0.36	0.26	0.24	0.31	0.36	0.40	0.28	0.39	0.41	0.41	0.05	1.00	-0.27	0.35	0.05
29	-0.24	-0.37	-0.56	-0.43	-0.37	-0.45	-0.51	-0.26	-0.09	-0.30	-0.09	-0.28	-0.26	0.19	-0.23	0.44	-0.48	-0.50	-0.52	-0.48	-0.31	-0.45	-0.52	-0.42	-0.52	-0.49	-0.33	-0.27	1.00	-0.54	-0.12
30	0.52	0.70	0.67	0.70	0.55	0.64	0.75	0.32	0.27	0.42	0.24	0.50	0.59	-0.04	0.46	-0.29	0.68	0.68	0.67	0.71	0.56	0.73	0.77	0.60	0.77	0.74	0.44	0.35	-0.54	1.00	0.22
31	0.13	0.34	0.35	0.42	0.46	0.18	0.31	0.24	0.11	0.18	-0.17	-0.01	0.18	0.05	0.12	0.09	0.33	0.22	0.20	0.31	0.26	0.26	0.27	0.29	0.35	0.27	0.11	0.05	-0.12	0.22	1.00

Fig. 4 Spearman correlation coefficients between scores of radiologists

Table 1 Interobserver variability in the IQA of MR images of different body parts

	κ	Confidence interval		p-value	ρ_{mean}	ρ_{max}	ρ_{min}	ρ_{std}
All images	0.120	0.118	0.121		0.5706	0.8615	-0.4988	0.3331
Spine	0.106	0.103	0.110		0.5944	0.9044	-0.3188	0.3156
Knee	0.114	0.111	0.117		0.5527	0.9108	-0.5202	0.4321
Shoulder	0.128	0.125	0.131		0.5839	0.9304	-0.6490	0.4009
Wrist	0.105	0.100	0.110		0.4581	0.9258	-0.7356	0.3982
Hip	0.048	0.042	0.055	< 0.001	0.4136	0.9487	-0.9487	0.6648
Elbow	0.079	0.071	0.088		0.9231	1.0000	-1.0000	0.3922
Ankle	0.089	0.081	0.098		0.6000	1.0000	-1.0000	0.8137
Brain	0.084	0.080	0.088		0.4966	0.9698	-0.8230	0.5786
Spine ∪ Brain	0.102	0.100	0.105		0.5774	0.9052	-0.4421	0.3606
All - (Spine ∪ Brain)	0.118	0.117	0.120		0.5671	0.8951	-0.4824	0.3286

group of radiologists was far more familiar with neuroradiology than with the musculoskeletal radiology, the influence of work experience of professionals on the perceived quality could be examined. As reported in Table 1 and Fig. 5, the correlation between radiologists' scores for neuroradiology images represented by the subset of spine and brain images were similar to correlations obtained for the subset of images of different joints representing musculoskeletal radiology. To support these observations, Fig. 6 contains the ρ values for radiologists in both cases.

Also, the experiments which involved interns revealed no significant influence of professional background in the quality assessment. Their average ρ is 0.6867 and is higher than the average result for experienced radiologists ($\rho = 0.5706$), demonstrating that the correct assessment can be performed even by an inexperienced observer. This can be also seen in Fig. 7, in which mean opinion scores for images are shown separately for professionals and interns. This is in contradiction to the work of Miao et al. [28] who assumed that radiologists have an advantage in the critical analysis of the images in which quality differences are present. However, such a claim was corrected in their further study [27]. In contrary to both studies, in which only up to two radiologists took part, the findings presented in this paper are based on decisions of a much larger group of medical professionals.

Furthermore, this study reveals that the content of images strongly affects their perceived quality. As can be seen in Fig. 8, dispersion of scores for images vary much for images of medium quality. The images of the worst quality were unanimously assessed by the group since they contain visible noise or distorted contours of the displayed shapes. Consequently, images of the best quality are also characterized by a relatively small standard deviation of the scores. This indicates that the decisions of radiologists are consistent. Interestingly, as

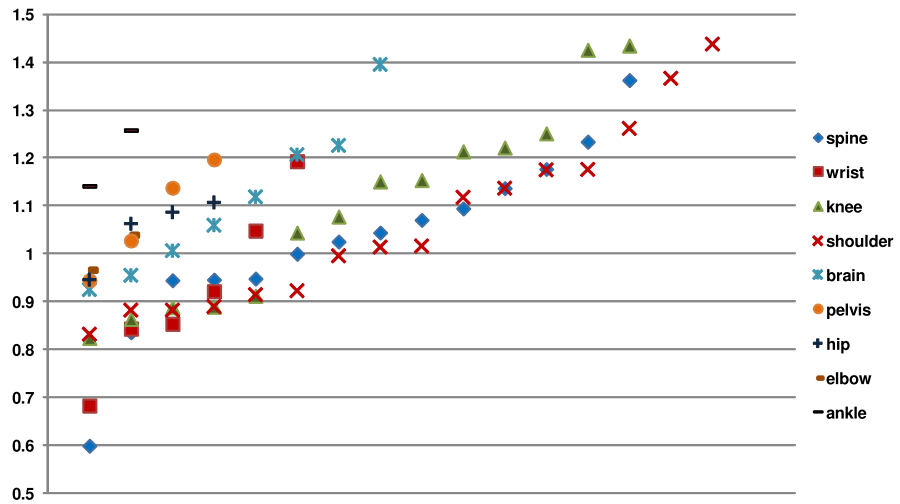


Fig. 5 Sorted standard deviations of scores for images of body parts

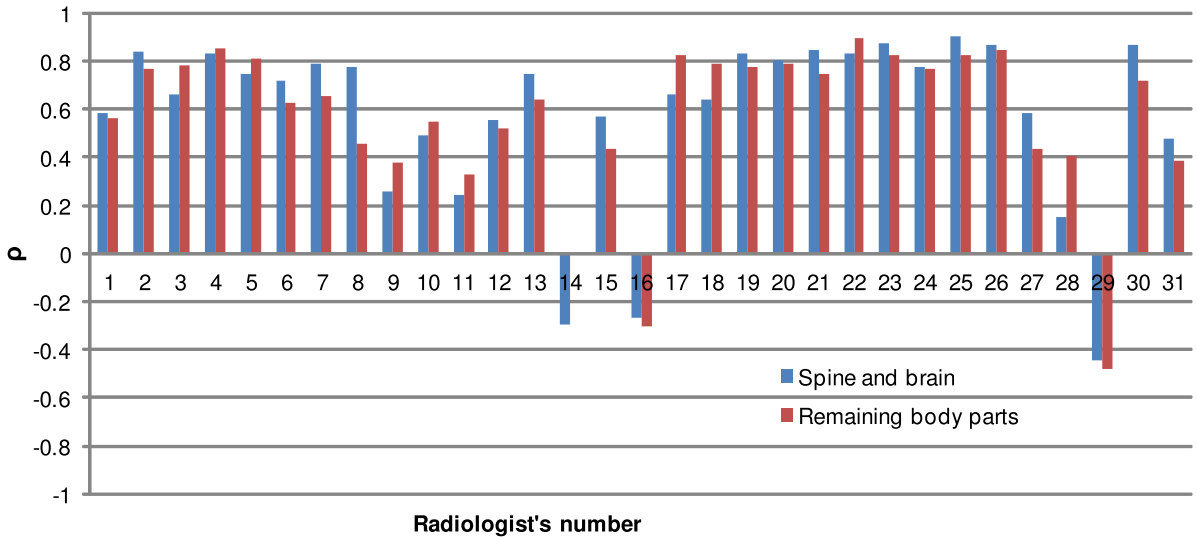


Fig. 6 Comparison of image quality assessment performances of radiologists evaluating MR images of familiar (i.e., brain and spine) and unfamiliar body parts. The performances are similar in both cases as the mean ρ for brain and spine images is 0.5774, while for the remaining images 0.5671

pointed out by Daly [8], a group of imaging professionals trained for the recognition of changes in the grayscale scene may be able to successfully use images of a low quality. To further investigate the dispersion of scores during the experiment, Fig. 9 shows their deviations for consecutive images. As revealed by the trend line, the standard deviation of scores slightly increases over time. It can be assumed that a longer duration of the test would negatively affect the performance of the group of radiologists. However, the observed trend is not strong since the experiment was fairly short to reduce the fatigue of the participants.

In the group of examined professionals, a moderate linear relationship between opinion scores was reported. This confirms the consistency of the majority of collected subjective opinions and highlights the interobserver agreement on the image quality. However, the scores of a few professionals are negatively correlated with the rest of the group which suggests that they did not use the established image grading system and assigned scores

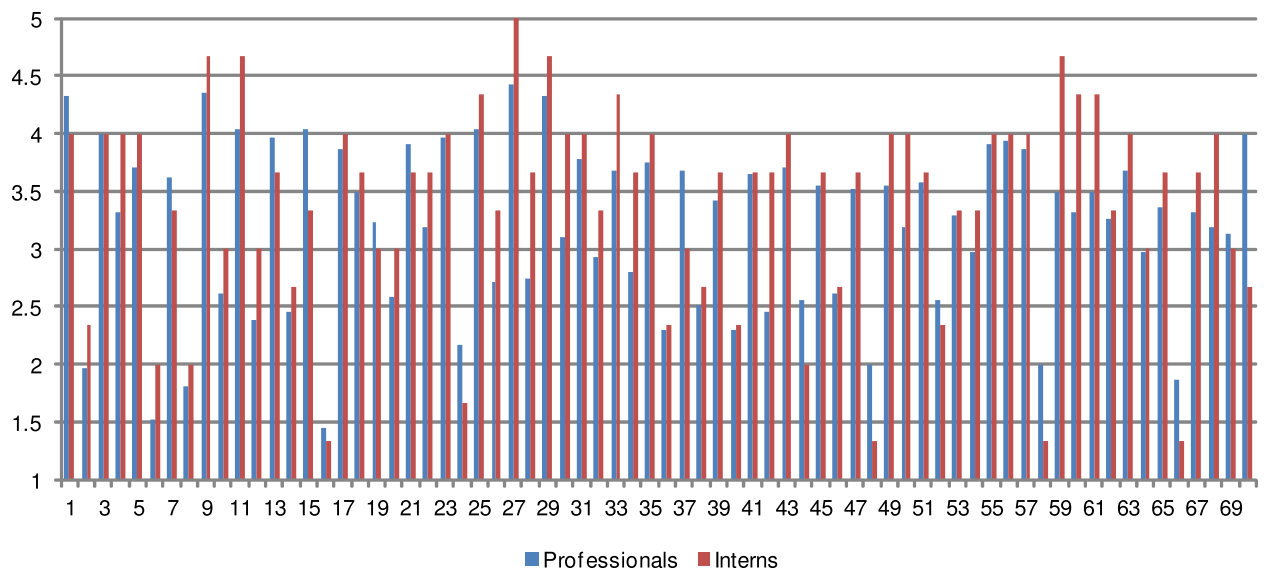


Fig. 7 Mean opinion scores of professionals and interns for images

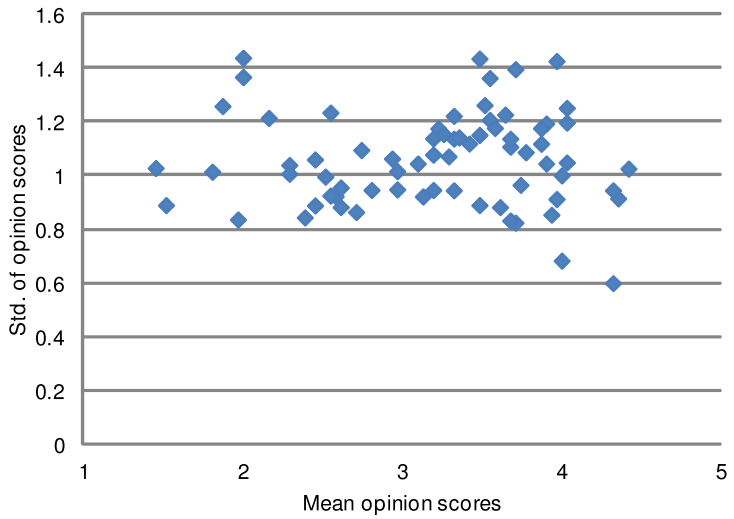


Fig. 8 Dispersion of opinion scores for images assessed by professionals. The differences between scores can be seen for images of different quality, indicating that the perception of displayed body parts and personal preferences of observers also took part in the image assessment

in the reversed scale. The correlation coefficients indicate that they were aware of the differences in the distortion severity of the assessed images as the rest of the group. The usage of the reversed scale may also show the importance of the overall attitude and psychosomatic status in the work with images [25]. In contrary to other medical professions, in radiology, there is a blind (one-way) interaction with displayed content which demands self-control and criticism.

The presented study was carried out on a representative group of radiologists and focused on the recognition of differences in the quality of MR images. The best of our knowledge such an approach is presented for the first time. Specifically, studies regarding quality in the diagnostic imaging proposed to date are directed towards the analysis of the influence of the image quality distortions on the perception of images [43]. Also, Sweeney et al. [39] and Rafferty et al. [33] presented findings on the influence of image quality on the perception of the pathology. In that work, images were artificially distorted using blur or noise. Influence of the different algorithms used for the raw image post-processing techniques on the image quality and their final perception by radiologists can be found in the literature [1, 2, 7]. Also, the analysis of the influence of image acquisition on the radiological perception of different pathologies in an various radiological modalities

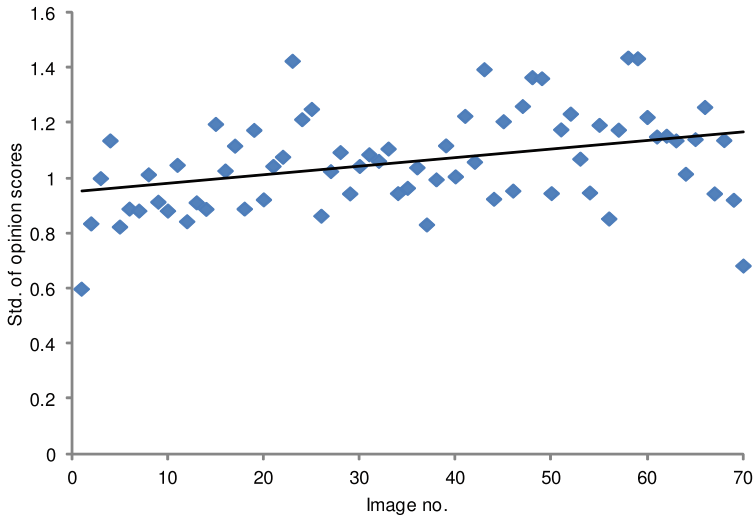


Fig. 9 Dispersion of opinion scores for images during the test

is often considered [6, 9, 12, 16, 18, 19, 23, 45]. However, these works lack an investigation of the level of the agreement among professionals assessing the quality of MR images.

The size of the group of radiologists as well as the number and diversity of the assessed images can be seen as the limitations of this study. However, to the best of our knowledge, this is the first time a large number of radiologists is involved in the assessment of the quality of images. Also, the choice of the images for the study is not accidental as they show typically examined body parts and parts with which most of the professionals are not familiar to study how their experience affects the perceived quality. Furthermore, the radiologists taking part in the study were familiar with the output of the employed 1.5T MRI scans as they work on machines of this filed strength. Consequently, assuming that the assessment of 3T MRI scans could be difficult for the professionals used to 1.5T images, the experimental setup applied in this study considers only 1.5T MRI scans to provide conditions that did not distracted participants.

Conclusions

This paper discusses the interobserver variability in the assessment of MR images. The variability was evaluated using opinion scores of the group of experienced medical professionals and interns, reflecting their assessment of a dataset of authentically distorted MR images. The observed agreement in the group of radiologists from different imaging centers confirmed that the perception of the image quality is subjective and depends on the meaning of the displayed shapes, contours, and grayscale differences responsible for the essential cognition of the image. It was determined that the quality assessment is only partially influenced by the distortion severity and is correlated neither with the knowledge on the anatomical representation of the structures nor the experiences on image perception. However, it was influenced by the psychosomatic condition and attitude of the observers.

Future work would be focused on an investigation of a group of professionals assessing medical images from different radiological modalities or an investigation of a degree of agreement among repeated examination of images in a form of intraobserver tests.

Abbreviations

CI: Confidence interval; IQA: Image quality assessment; MOS: Mean opinion scores; MRI: Magnetic resonance imaging

Acknowledgements

Not Applicable.

Authors' contributions

Conceptualization: RO, AP; Data curation: RO, MO; Formal analysis: MO ; Investigation: RO, AP, MO; Methodology: MO; Project administration: AP; Resources: RO; Software, MO; Supervision: RO, AP; Validation: MO, AP, RO; Writing - original draft: RO, MO; Writing - review & editing: RO, MO, AP. All authors have read critically and approved the manuscript.

Funding

This publication was co-funded by AGH University of Science and Technology, Faculty of Electrical Engineering, Automatics, Computer Science and Biomedical Engineering under grant number 16.16.120.773.

Availability of data and materials

The dataset supporting the conclusions of this article is available in the mriqdataset.7z repository: http://home.agh.edu.pl/pioro/mriqdata/.

Ethics approval and consent to participate

Ethical approval: The study protocol was designed according to the guidelines of the Declaration of Helsinki and the Good Clinical Practice standard. Written acceptance for conducting the study was obtained from the Ethics Committee of Jagiellonian University (no. 1072.6120.15.2017).
Ethical Statement: All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and national research committee in accordance to 1964 Helsinki declaration and its latter amandements.
Consent to participate: Written consent to participate in the study was obtained from all participants. Furthermore, written consent for publication of anonymised data was obtained from all participants.

Consent for publication

Not Applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Department of Diagnostic Imaging, Jagiellonian University Medical College, Kopernika Street 19, 31-501, Cracow, Poland. [2]Department of Computer and Control Engineering, Rzeszow University of Technology, Wincentego Pola 2, 35-959, Rzeszow, Poland. [3]Department of Biocybernetics and Biomedical Engineering, AGH University of Science and Technology, Mickiewicza 30, 30-059, Cracow, Poland.

Received: 13 April 2020 Accepted: 1 September 2020
Published online: 22 September 2020

References

1. Abbey CK, Barrett HH. Human- and model-observer performance in ramp-spectrum noise: effects of regularization and object variability. J Opt Soc Am A. 2001;18(3):473–88. https://doi.org/10.1364/JOSAA.18.000473.
2. Barrett HH, Denny JL, Wagner RF, Myers KJ. Objective assessment of image quality. ii. fisher information, fourier crosstalk, and figures of merit for task performance. J Opt Soc Am A. 1995;12(5):834–52. https://doi.org/10.1364/JOSAA.12.000834.
3. Buda M, Wildman-Tobriner B, Hoang JK, Thayer D, Tessler FN, Middleton WD, Mazurowski MA. Management of thyroid nodules seen on us images: Deep learning may match performance of radiologists. Radiology. 2019;292(3):695–701.
4. Chaddha N, Meng T. Psycho-visual based distortion measures for monochrome image and video compression. 1993841–5. https://doi.org/10.1109/ACSSC.1993.342451.
5. Chalavi S, Simmons A, Dijkstra H, Barker GJ, Reinders AA. Quantitative and qualitative assessment of structural magnetic resonance imaging data in a two-center study. BMC Med Imaging. 2012;12(27):. https://doi.org/10.1186/1471-2342-12-27.
6. Chen L, Barrett H. Optimizing lens-coupled digital radiographic imaging systems based on model observers' performance. Proc SPIE Int Soc Opt Eng. 2003;5034:63–70. https://doi.org/10.1117/12.480331.
7. Clarkson E, Barrett HH. Approximations to ideal-observer performance on signal-detection tasks. Appl Opt. 2000;39(11):1783–93. https://doi.org/10.1364/AO.39.001783.
8. Daly SJ. Visible differences predictor: an algorithm for the assessment of image fidelity. 1992. https://doi.org/10.1117/12.135952.
9. Deichmann R, Good C, Josephs O, Ashburner J, Turner R. Optimization of 3d mp-rage sequence for structural brain imaging. NeuroImage. 2000;12:112–27. https://doi.org/10.1006/nimg.2000.0601.

10. Deshmane A, Gulani V, Griswold MA, Seiberlich N. Parallel MR imaging. J Magn Reson Imaging. 2012;36(1):55–72. https://doi.org/10.1002/jmri.23639.
11. Eck BL, Fahmi R, Brown KM, Zabic S, Raihani N, Miao J, Wilson DL. Computational and human observer image quality evaluation of low dose, knowledge-based ct iterative reconstruction. Med Phys. 2015;42(10):6098–111. https://doi.org/10.1118/1.4929973.
12. Filippi M, van Waesberghe JH, Horsfield MA, Bressi S, Gasperini C, Yousry TA, Gawne-Cain ML, Morrissey SP, Rocca MA, Barkhof F, Lycklama à Nijeholt GJ, Bastianello S, Miller DH. Interscanner variation in brain mri lesion load measurements in ms: Implications for clinical trials. Neurology. 1997;49(2):371–7. https://doi.org/10.1212/WNL.49.2.371.
13. Fleiss JL. Measuring nominal scale agreement among many raters. Psychological bulletin. 1971;76(5):378.
14. Ghadiyaram D, Bovik AC. Massive online crowdsourced study of subjective and objective picture quality. IEEE T. Image Process. 2016;25(1):372–387. https://doi.org/10.1109/TIP.2015.2500021.
15. Huo D, Xu D, Liang ZP, Wilson D. Application of perceptual difference model on regularization techniques of parallel mr imaging. Magn Reson Imaging. 2006;24(2):123–32. https://doi.org/10.1016/j.mri.2005.10.018.
16. Jovicich J, Czanner S, Han X, Salat D, van der Kouwe A, Quinn B, Pacheco J, Albert M, Killiany R, Blacker D, Maguire R, Rosas H, Makris N, Gollub R, Dale A, Dickerson B, Fischl B. Mri-derived measurements of human subcortical, ventricular and intracranial brain volumes: Reliability effects of scan sessions, acquisition sequences, data analyses, scanner upgrade, scanner vendors and field strengths. NeuroImage. 2009;46:177–92. https://doi.org/10.1016/j.neuroimage.2009.02.010.
17. Kagadis GC, Walz-Flannigan A, Krupinski EA, Nagy PG, Katsanos K, Diamantopoulos A, Langer SG. Medical imaging displays and their use in image interpretation. Radiographics. 2013;33(1):275–90.
18. Kato M, Saji S, Kanematsu M, Fukada D, Miya K, Umemoto T, Kunieda K, Sugiyama Y, Takao H, Kawaguchi Y, Takagi Y, Kondo H, Hoshi H. Detection of lymph-node metastases in patients with gastric carcinoma: Comparison of three mr imaging pulse sequences. Abdom Imaging. 2000;25:25–9. https://doi.org/10.1007/s002619910004.
19. Kruggel F, Turner J, Tugan Muftuler L. Impact of scanner hardware and imaging protocol on image quality and compartment volume precision in the adni cohort. NeuroImage. 2009;49:2123–33. https://doi.org/10.1016/j.neuroimage.2009.11.006.
20. Krupa K, Bekiesinska-Figatowska M. Artifacts in magnetic resonance imaging. Pol J Radiol. 2015;80:93–106. https://doi.org/10.12659/PJR.892628.
21. Krupinski EA. Current perspectives in medical image perception. Atten Percept Psychophys. 2010;72:1205–17.
22. Kundel HL. Reader error, object recognition, and visual search; 2004. https://doi.org/10.1117/12.542717.
23. Ludwig K, Lenzen H, Kamm KF, Link T, Diederich S, Wormanns D, Heindel W. Performance of a flat-panel detector in detecting artificial bone lesions: Comparison with conventional screen-film and storage-phosphor radiography1. Radiology. 2002;222:453–9.
24. Manning DJ, Ethell SC, Donovan T. Detection or decision errors? missed lung cancer from the posteroanterior chest radiograph. Br J Radiol. 2004;77(915):231–5. https://doi.org/10.1259/bjr/28883951. PMID: 15020365.
25. Manning DJ, Gale A, Krupinski EA. Perception research in medical imaging. Br J Radiol. 2005;78(932):683–5. https://doi.org/10.1259/bjr/72087985.
26. MATLAB. version R2018b. Natick, Massachusetts: The MathWorks Inc.; 2018.
27. Miao J, Huang F, Narayan S, Wilson DL. A new perceptual difference model for diagnostically relevant quantitative image quality evaluation: A preliminary study. Magn Reson Imaging. 2013;31(4):596–603. https://doi.org/10.1016/j.mri.2012.09.009.
28. Miao J, Huo D, Wilson DL. Quantitative image quality evaluation of mr images using perceptual difference models. Med Phys. 2008;35(6Part1):2541–53. https://doi.org/10.1118/1.2903207.
29. Oszust M. No-reference quality assessment of noisy images with local features and visual saliency models. Inf Sci. 2019;482:334–49. https://doi.org/10.1016/j.ins.2019.01.034.
30. Pang Z, Margolis M, Menezes RJ, Maan H, Ghai S. Diagnostic performance of 2015 American thyroid association guidelines and inter-observer variability in assigning risk category. Eur J Radiol Open. 2019;6:122–7. https://doi.org/10.1016/j.ejro.2019.03.002.
31. Pedersen M, Hardeberg JY. A new spatial hue angle metric for perceptual image difference. In: Trémeau A, Schettini R, Tominaga S, editors. Computational Color Imaging. Berlin: Springer Berlin Heidelberg; 2009. p. 81–90.
32. R Margulis A, Dirk Sostman H. Radiologist-patient contact during the performance of cross-sectional examinations. J Am Coll Radiol JACR. 2004;1:162–3. https://doi.org/10.1016/j.jacr.2003.12.011.
33. Rafferty EA, Park JM, Philpotts LE, Poplack SP, Sumkin JH, Halpern EF, Niklason LT. Assessing radiologist performance using combined digital mammography and breast tomosynthesis compared with digital mammography alone: Results of a multicenter, multireader trial. Radiology. 2013;266(1):104–13. https://doi.org/10.1148/radiol.12120674.
34. Rajashekar U, Wang Z, Simoncelli1 EP. Quantifying color image distortions based on adaptive spatio-chromatic signal decompositions. In: 2009 16th IEEE International Conference on Image Processing (ICIP); 2009. p. 2213–6. https://doi.org/10.1109/ICIP.2009.5413889.
35. Salem KA, Lewin JS, Aschoff AJ, Duerk JL, Wilson DL. Validation of a human vision model for image quality evaluation of fast interventional magnetic resonance imaging. J Electron Imaging. 2002;11(2):224–235–12. https://doi.org/10.1117/1.1453412.
36. Sheikh HR, Bovik AC. Image information and visual quality. IEEE T. Image Process. 2006;15(2):430–44. https://doi.org/10.1109/TIP.2005.859378.
37. Sheikh HR, Sabir MF, Bovik AC. A statistical evaluation of recent full reference image quality assessment algorithms. IEEE T. Image Process. 2006;15(11):3440–51. https://doi.org/10.1109/tip.2006.881959.
38. Sinha N, Ramakrishnan A. Quality assessment in magnetic resonance images. Crit Rev Trade Biomed Eng. 2010;38(2):127–41. https://doi.org/10.1615/CritRevBiomedEng.v38.i2.20.
39. Sweeney RJI, Lewis SJ, Hogg P, McEntee MF. A review of mammographic positioning image quality criteria for the craniocaudal projection. Br J Radiol. 2018;91(1082):20170,611.
40. Video Quality ExpertsGroup. Final report from the video quality experts group on the validation of objective models of video quality assessment, phase ii (fr_tv2). 2003. https://www.itu.int/ITU-T/studygroups/com09/docs/tutorial_opavc.pdf, Accessed 07 June 2020.
41. Wang Z, Bovik AC, Lu L. Why is image quality assessment so difficult? In: 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 4; 2002. p. IV–3313–IV–3316. https://doi.org/10.1109/ICASSP.2002.5745362.
42. Wang Z, Bovik AC, Sheikh HR, Simoncelli EP. Image quality assessment: From error visibility to structural similarity. IEEE T Image Process. 2004;13(4):600–12. https://doi.org/10.1109/tip.2003.819861.
43. Williams MC, Golay SK, Hunter A, Weir-McCall JR, Mlynska L, Dweck MR, Uren NG, Reid JH, Lewis SC, Berry C, van Beek EJR, Roditi G, Newby DE, Mirsadraee S. Observer variability in the assessment of ct coronary angiography and coronary artery calcium score: substudy of the scottish computed tomography of the heart (scot-heart) trial. Open Heart. 2015;2(1):. https://doi.org/10.1136/openhrt-2014-000234.
44. Wilson R, Knutsson H. Uncertainty and inference in the visual system. IEEE Trans Syst Man Cybern. 1988;18(2):305–12. https://doi.org/10.1109/21.3468.
45. Wonderlick J, Ziegler D, Hosseini-Varnamkhasti P, Locascio J, Bakkour A, van der Kouwe A, Triantafyllou C, Corkin S, Dickerson B. Reliability of mri-derived cortical and subcortical morphometric measures: Effects of pulse sequence, voxel geometry, and parallel imaging. NeuroImage. 2008;44:1324–33. https://doi.org/10.1016/j.neuroimage.2008.10.037.

Publisher's Note