Magnetic Resonance Imaging 34 (2016) 820–831

Contents lists available at ScienceDirect

Magnetic Resonance Imaging

journal homepage: www.mrijournal.com

Original contribution

Correlation between subjective and objective assessment of magnetic resonance (MR) images☆

Li Sze Chow [a,*], Heshalini Rajagopal [a], Raveendran Paramesran [a], Alzheimer's Disease Neuroimaging Initiative [b]

[a] Department of Electrical, Faculty of Engineering, University of Malaya, Lembah Pantai, 50603 Kuala Lumpur, Malaysia
[b] ADNI Communication Office, Alzheimer's Disease Cooperative Study, University of California, San Diego, 9500 Gilman Drive, LA Jolla, CA 92093-0949

ARTICLE INFO

Article history:
Received 31 July 2015
Revised 25 February 2016
Accepted 3 March 2016

Keywords:
Difference Mean Opinion Score (DMOS)
Full Reference—Image Quality Assessment (FR-IQA)
Objective assessment
Subjective assessment

ABSTRACT

Medical Image Quality Assessment (IQA) plays an important role in assisting and evaluating the development of any new hardware, imaging sequences, pre-processing or post-processing algorithms. We have performed a quantitative analysis of the correlation between subjective and objective Full Reference - IQA (FR-IQA) on Magnetic Resonance (MR) images of the human brain, spine, knee and abdomen. We have created a MR image database that consists of 25 original reference images and 750 distorted images. The reference images were distorted with six types of distortions: Rician Noise, Gaussian White Noise, Gaussian Blur, DCT compression, JPEG compression and JPEG2000 compression, at various levels of distortion. Twenty eight subjects were chosen to evaluate the images resulting in a total of 21,700 human evaluations. The raw scores were then converted to Difference Mean Opinion Score (DMOS). Thirteen objective FR-IQA metrics were used to determine the validity of the subjective DMOS. The results indicate a high correlation between the subjective and objective assessment of the MR images. The Noise Quality Measurement (NQM) has the highest correlation with DMOS, where the mean Pearson Linear Correlation Coefficient (PLCC) and Spearman Rank Order Correlation Coefficient (SROCC) are 0.936 and 0.938 respectively. The Universal Quality Index (UQI) has the lowest correlation with DMOS, where the mean PLCC and SROCC are 0.807 and 0.815 respectively. Student's T-test was used to find the difference in performance of FR-IQA across different types of distortion. The superior IQAs tested statistically are UQI for Rician noise images, Visual Information Fidelity (VIF) for Gaussian blur images, NQM for both DCT and JPEG compressed images, Peak Signal-to-Noise Ratio (PSNR) for JPEG2000 compressed images.

1. Introduction

Image quality assessment can be categorized into two types, namely subjective and objective assessments. Subjective assessment is the ratings given by human subjects based on their judgment on the image quality. Subjective assessment is always regarded as the gold standard in the image quality assessment for MR images. Objective assessment is an alternative method defined mathematically. It can be divided into three types: Full Reference - Image Quality Assessment (FR-IQA), Reduced Reference - Image Quality Assessment (RR-IQA) and No Reference/Blind Image - Quality Assessment (NR-IQA) [1,2]. FR-IQA calculates an image quality score relative to a reference image. The reference image is usually a perfect image without any distortion. RR-IQA uses partial information from the reference image to calculate the image quality score of the distorted image. NR-IQA calculates the image quality score without using reference image. Since the reference image is usually unavailable in medical images, NR-IQA is more feasible than the other two methods.

Several researchers have performed objective FR-IQA [1,3,4] and NR-IQA [5–7] evaluation on MR images. Gulame et al. evaluated MR images distorted with speckle noise using distance based metrics [1]. They found that Manhattan, Bray–Curtis and Cosine Correlation Distance measured the MR images better than the Euclidean, Chebyshev and Canberra Distance. However, their study only focuses on MR images distorted with speckle noise. R. Kumar et al. analyzed a variety of quality metrics for MRI, X-ray and ultrasound images,

☆ Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wpcontent/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

* Corresponding author at: Department of Electrical, Faculty of Engineering, University of Malaya, Lembah Pantai, 50603 Kuala Lumpur, Malaysia.
E-mail addresses: lschow@um.edu.my, lisze.chow@gmail.com (L.S. Chow), heshalini@gmail.com (H. Rajagopal), ravee@um.edu.my (R. Paramesran).

http://dx.doi.org/10.1016/j.mri.2016.03.006
0730-725X/

which include Mean Squared Error (MSE), Structural Similarity Index (SSIM), Peak Signal-to-Noise (PSNR), Maximum difference (MD), etc. [3]. They distorted the MR images with different levels of blur, noise, compression and contrast levels. They found that SSIM can evaluate the image quality regardless of the type of distortion, and it outperformed all the other metrics analyzed in their study. However, the computational time for SSIM is large.

B. Kumar et al. performed the subjective FR-IQA on MR images. They evaluated the performance of PSNR and SSIM on compressed medical images using MOS [4]. From their study, it was found that the MOS values vary according to the type of image compression. They concluded that MOS values correlate better with PSNR than with SSIM for all compression schemes. Prieto et al. performed a study on subjective assessment based on Just Noticeable Differences (JND). They proposed JND scanning (JNDS) to evaluate reconstructed MR images [8]. They also measured the image quality using Root Mean Square Error (RMSE). The JNDS metric was validated by using subjective MOS values obtained from five observers. The result showed that JNDS has a better correlation with the subjective MOS than that with RMSE. Both of these studies used the subjective MOS as a standard benchmark to evaluate the performance of the objective assessment.

The main goal of IQA is to model an ideal objective assessment metric that is very close to the human evaluation [9]. To achieve this goal, several researchers performed experiments on the subjective assessment and produced database of natural images, namely Laboratory for Image and Video Engineering (LIVE) [9], Categorical Subjective Image Quality (CSIQ) [10], Cornell A57 [11], IVC [12], Toyoma-MICT [13], TID2008 [14], and TID2013 [15]. The most widely used database for IQA study is LIVE, which contains 779 distorted natural scene images [9]. The reference images were distorted by JPEG2000 compression, JPEG compression, White Gaussian noise (WGN), Gaussian blur (GB), or Simulated Fast Fading Rayleigh channel. All these images were evaluated by 24 human subjects and the ratings were represented with Difference Mean Opinion Score (DMOS). The rest of the databases also contain few hundreds of distorted natural scene images, which were distorted by a few types of distortion, and evaluated by a number of human subjects. The subjective scores were presented in either MOS [11–15] or DMOS [9,10].

In this work, we create a database of MR images containing six types of distortion that may possibly occur during imaging and storing. There are a total of 775 MR images consisting of 750 distorted images derived from 25 reference images. The MR images were evaluated by 28 human subjects, and the ratings were converted to DMOS. The DMOS values are compared with thirteen FR-IQA metrics: SNR, PSNR, SSIM, Multiscale SSIM (MS-SSIM), Feature SIMilarity (FSIM), Information Fidelity Criterion (IFC), Noise Quality Measurement (NQM), Weighted SNR (WSNR), Visual Information Fidelity (VIF), Pixel Visual Information Fidelity (VIFP), Universal Quality Index (UQI), Information Weighted PSNR (IW-PSNR) and Information Weighted SSIM (IW-SSIM). We used Pearson Linear Correlation Coefficient (PLCC), Spearman Rank Order Correlation Coefficient (SROCC), Kendall Rank Order Correlation Coefficient (KROCC) and RMSE to validate the correlation between the DMOS and all the FR-IQA used here.

2. Methodology

2.1. MR Images

Twenty five good quality MR images were chosen from two sources of online database: Osirix DICOM Viewer MRI database [16] and Alzheimer's Disease Neuroimaging Initiative (ADNI) MRI database (adni.loni.usc.edu) [17]. The ADNI was launched in 2003 by the National Institute on Aging (NIA), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), the Food and Drug Administration (FDA), private pharmaceutical companies and non-profit organizations, as a $60 million, 5-year public-private partnership.

Fig. 1 shows the original MR images which consist of images from the brain, abdomen (gastroenterology), spine and knee. They are T1 Weighted (T1W), T2 Weighted (T2W), or Proton Density (PD) images. These images are used as reference images in our study. All the MR images are in gray scale. They were normalized between 0 and 255 for the ease of applying the same level of distortion for all reference images. The image pixels are written below each image, respectively.

2.2. Image distortion

The reference images were distorted using six types of distortion each at five different levels as summarized in Table 1. Five different levels were used in order to predict the change in the image quality as the intensity of distortion is worsening. It is a random choice of five levels; however, it would make up a comparable amount of distorted images as performed by other similar studies using natural images [9]. Rician Noise, Gaussian White Noise, Gaussian blur, Discrete Cosine Transform (DCT), JPEG Compression and JPEG2000 Compression are chosen because they are common in majority MR images. If the SNR of the MR images is greater than 2, the images are subjected to Gaussian Noise; whereas, if the SNR is lower than 2, the images are subjected to Rician Noise [18]. MR images are subjected to Gaussian Blur when it is exposed to the atmosphere for a long time [19,20]. DCT, JPEG and JPEG2000 compressions are common techniques used to compress a wide range of MRI information [21–26]. All these distortion except the Rician noise, are commonly used by the other similar studies of subjective assessment for natural images [9–15].

2.3. Subjective evaluation

The subjective evaluation was done by following the procedures recommended by Rec. ITU-R BT.500-11 [27]. The evaluation was performed in an office environment with normal indoor illumination level. 24-in LED monitor with a resolution of 1920 × 1080 pixels was used for the subjective evaluation. We have selected twenty eight human subjects (15 male, 13 female) with normal vision to evaluate the MR images. They are research scholars from Electrical Engineering department, aged between 20 and 35 years. The subjects were screened for near visual acuity using Snellen Chart. During the vision test, the subjects were asked to sit at a distance of 760 mm from the Snellen Chart. Evaluation was done by using Simultaneous Double Stimulus for Continuous Evaluation (SDSCE) methodology [27], where two images are displayed on the monitor screen side by side. The left image is always the reference image, and the right image is the distorted image for evaluation. Each subject rates the distorted image by judging the differences between the two images on the screen. We notice that they can evaluate fairly when the reference image is displayed beside the distorted image. The subject selects Excellent, Good, Fair, Poor or Bad. The numerical scores that represent each rating are 90, 70, 50, 30 and 10 respectively. These scores were not disclosed to the subjects in order to avoid bias by the subjects [28]. Wajid et al. also used similar rating scores in their study for investigating the similarity between psychophysical experiment and LIVE image quality database [29].

Written instructions about the evaluation procedures were given to each subject prior to evaluation of the MR images. Then, a demonstration session was conducted with a few examples of distorted images corresponding to a recommended quality rating. A

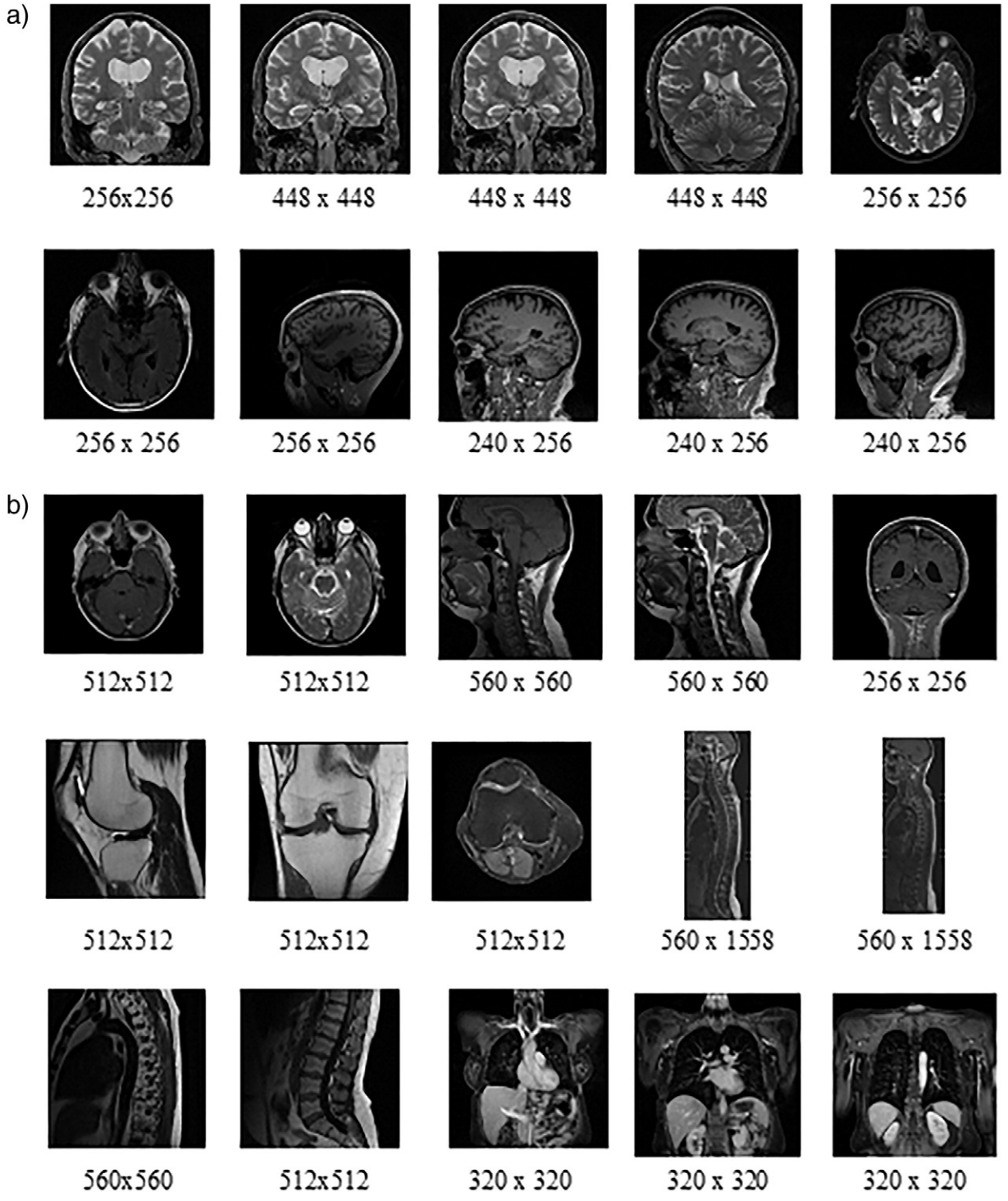

Fig. 1. Twenty five reference MR images used in this study. The image size in pixels is written below each image. The images are complimentary shared by online database from: (a) http://adni.loni.usc.edu and (b) http://www.osirix-viewer.com/datasets/.

mock test was also performed where the subjects evaluated two sets of MR images (two reference images with sixty two test images). In the case where two similar reference images were shown on the screen, if the subject did not rate them as 'Excellent' quality, this subject would not be used for our study.

The subjective evaluation period should take less than 30 min for each subject to avoid fatigue. Since there were a large number of MR images to be evaluated, the evaluation was divided into three sessions where the first two sessions contained eight sets of MR images (eight reference images with 248 distorted images), and the third session contained nine sets of MR images (nine reference images with 279 distorted images). The three sessions were conducted on three consecutive days. There was no time constraint for the subjective assessment; they took an average of 20 min for each session.

Table 1
Summary of all distortions applied to the reference MR images.

Distortion type	Description	Distortion levels
Rician Noise	Rician Noise Probability Density Function (PDF) with standard deviation, σ_R.	σ_R: 5, 15, 25, 35, 45
Gaussian White Noise	Gaussian White Noise distribution with standard deviation, σ_N.	σ_N: 4, 11, 18, 30, 50
Gaussian Blur	3σ sized square kernel window with Gaussian kernels of standard deviation, σ_{GB}.	σ_{GB}: 1.5, 3, 4.5, 6, 7.5
Discrete Cosine Transform (DCT)	Two dimensional (2-D) DCT with compression rates at bits per pixel (bpp).	bpp: 0.1, 0.8, 1.5, 2.2, 2.9
JPEG Compression	Lossy compression technique which uses 8 × 8 DCT encoded with a quality setting varies between 0 and 100. A higher quality setting produces better image quality.	Quality: 1, 7, 13, 19, 25
JPEG2000 Compression	Advanced image compression technique using wavelet transform. A higher compression ratio produces lower image quality.	Compression Ratio: 25, 50, 75, 100, 125

2.4. Data processing

The first step of data processing was to check on any unqualified subject scores using outlier detection and subject rejection algorithm based on the ITU-R BT.500-11 recommendation [27]. Mean score, $\overline{\mu_k}$, is calculated using Eq. (1):

$$\overline{\mu_k} = \frac{1}{N}\sum_{i=1}^{N}\mu_{ik} \tag{1}$$

where μ_{ik} is the score given by i^{th} observer for k^{th} image and N is the number of observers. It is recommended to present all the mean scores at 95% confidence interval represented by $[\overline{\mu_k}-\delta_k, \overline{\mu_k}+\delta_k]$ where $\delta_k = 1.96\ \frac{S_k}{\sqrt{N}}$, and S_k is the standard deviation for each image:

$$S_k = \sqrt{\sum_{i=1}^{N}\frac{(\overline{\mu_k}-\mu_{ik})^2}{(N-1)}} \tag{2}$$

β_2 test is used to verify whether the scores have a normal distribution. This test is done by calculating the kurtosis coefficient, β_{2k}, which is the ratio of the fourth order moment to the square of the second order moment:

$$\beta_{2k} = \frac{m_4}{(m_2)^2} \quad where \quad m_x = \frac{\sum_{i=1}^{N}(\mu_{ik}-\overline{\mu_{ik}})^x}{N} \tag{3}$$

If the β_2 is between 2 and 4, the distribution is considered normal. The pseudocode for the outlier detection and subject rejection algorithm is given below [27].

For each observer, i:

For each image, k:

if $2 \le \beta_{2k} \le 4$, *then:*

if $\mu_{ik} \ge \bar{\mu}_k + 2S_k$ *then* $P_i = P_i + 1$

if $\mu_{ik} \ge \bar{\mu}_k - 2S_k$ *then* $Q_i = Q_i + 1$

else:

if $\mu_{ik} \ge \bar{\mu}_k + \sqrt{20}S_k$ *then* $P_i = P_i + 1$

if $\mu_{ik} \ge \bar{\mu}_k - \sqrt{20}S_k$ *then* $Q_i = Q_i + 1$

if $\frac{P_i+Q_i}{K} > 0.05$ *and* $\left|\frac{P_i-Q_i}{P_i+Q_i}\right| < 0.3$ *then the* i^{th} *observer is rejected*

where P_i and Q_i are counters which are used to determine the rejection based on the condition given in the last line of the pseudocode.

After all the raw scores were tested with the above outlier detection and subject rejection algorithm, the qualified raw scores were used to calculate the DMOS values. DMOS measures the perceived relative quality of the degraded images based on the reference image [30]. Sheikh et al. and Thorpe et al. proved that DMOS is a good representation of raw scores obtained from the subjective evaluation [9,31]. In this study, the DMOS was calculated using [32]:

$$D = \frac{\overline{z_k} - \min(\overline{z_k})}{max(\overline{z_k}) - \min(\overline{z_k})} \times \quad 100 \tag{4}$$

where $\overline{z_k}$ is the averaged Z scores across all subjects for k^{th} image. The DMOS values range from 0 to 100, where a lower value represents higher image quality and vice versa.

Next, thirteen FR-IQA scores were calculated using the original reference image and each distorted image. The chosen FR-IQA metrics are SNR [33], PSNR [33,34], SSIM [35], MS-SSIM [35], FSIM [36], IFC [37], NQM [38], WSNR [38], VIF [39], VIFP [39], UQI [40], IW-PSNR [41] and IW-SSIM [41]]. The formulas and brief description of these FR-IQAs are given in Appendix A.

2.5. Performance metrics

We used three types of performance metrics to validate the DMOS values in this study: logistic regression, correlation coefficient and RMSE (refer Appendix B for all the formulas of these performance metrics). A nonlinear regression for the objective scores is constructed using a logistic regression function. It provides nonlinear mapping between the objective and subjective scores [42], which can be plotted on a graph for visual inspection and comparison between the subjective data points and computed objective scores.

Relationship between two datasets can be measured statistically using correlation coefficient. According to Taylor R. [43], two datasets are said to have high correlation if the correlation coefficient values are between 0.68 and 1.0. Three types of correlation coefficients (PLCC, SROCC and KROCC) were used in this study to measure the relationship between the subjective (DMOS) and objective (FR-IQA) scores. PLCC, also known as Pearson product–moment correlation coefficient, is used to evaluate the accuracy of the prediction. SROCC and KROCC are the nonparametric versions of PLCC, and do not require datasets that have been mapped nonlinearly as they operate only on the rank of the data points and ignore the relative distance between data points [36]. They evaluate the prediction monotonicity of the FR-IQA metrics. RMSE is a standard statistical metric used to evaluate the performance of a model and the consistency of the prediction.

2.6. Statistical testing

We used one sided Student's T-test at 95% confidence level to perform the statistical testing between all the FR-IQA metrics for different types of distortions and different types of anatomical images. To test for the different types of distortion, first, we calculated the residual between the DMOS and each FR-IQA scores (after logistic regression) for each type of distortions. A small residual means a small difference between the DMOS and a particular FR-IQA scores, and vice versa. Therefore, a smaller residual also represents a superior choice of FR-IQA metric. The T-test was calculated between the residual of an X_1–DMOS and another X_2–DMOS, where $X_i \in$ {FSIM, IFC, IWPSNR, IWSSIM, MSSIM, NQM, PSNR, SNR, SSIM, UQI, VIF, VIFP, WSNR}, for example, between residual of FSIM-DMOS and residual of IFC-DMOS.

The T-tests were conducted twice (left-tailed and right-tailed) for each pair of X_1–X_2 on each type of distortion. The null hypothesis is that both residuals of X_1–DMOS and X_2–DMOS have equal mean and equal variance, which means X_1 and X_2 are indistinguishable and this is recorded as symbol '-' in Table 3. The alternative hypothesis of the left-tailed T-test is that the mean of the residual of X_1–DMOS is less than the mean of the residual of X_2–DMOS. In other words, X_1 is superior to X_2. This is recorded as symbol '1' in Table 3 where X_1 is written in the row and X_2 is written in column. On the other hand, the alternative hypothesis of the right-tailed T-test is that the mean of the residual of X_1–DMOS is greater than the mean of the residual of X_2–DMOS. In other words, X_1 is inferior to X_2, which is recorded as symbol '0' in Table 3. There are six symbols for each entry of X_1–X_2 in Table 3, each symbol representing the result for different types of distortion, arranged in the following order: Rician Noise, Gaussian White Noise, Gaussian Blur, DCT, JPEG and JPEG2000.

The above T-test is repeated to investigate the efficiency of 13 FR-IQAs across different types of MR images (T1W, T2W, PD) with different field strengths (1.5 T and 3.0 T). The T-test results are recorded in Table 5, where each entry contains 6 symbols representing different types of images arrange in the following order: 1.5 T T1W, 1.5 T T2W, 1.5 T PD, 3.0 T T1W, 3.0 T T2W, and 3.0 T PD.

3. Results

According to the outlier detection and subjective rejection algorithm, no subject was rejected. In other words, all the subjects were in an alert condition during the image evaluation session over the three sessions. Hence, they were able to give a fair rating within acceptable range recommended by ITU-R BT.500-11 [27]. Therefore, all the human ratings were used for the DMOS calculation.

The scatter plots for the DMOS versus the standard deviations, σ_R, σ_N and σ_{GB} for Rician Noise, Gaussian White Noise and Gaussian Blur are shown in Fig. 2(a)–(c), respectively. The scatter plot for the DMOS versus DCT compression rate is shown in Fig. 2(d), DMOS versus JPEG quality is in Fig. 2(e), DMOS versus JPEG2000 compression ratio is in Fig. 2(f). For those images distorted with the Rician Noise, Gaussian White Noise and Gaussian Blur, the higher the standard deviation of the noise or blur, the poorer the resulted image quality, represented with higher DMOS values. The lower the DCT compression rate or JPEG compression quality, the poorer the resulting image quality. On the other hand, the higher the JPEG2000 compression rate, the poorer the resulting image quality.

Fig. 2(a)–(c) shows that the DMOS values increase with the increase in the standard deviation of the noise as expected. It means a poorer image has a higher DMOS value and this agrees with the theory. In Fig. 2(d), it is apparent that the subjects were only able to identify a low bit rate compressed DCT images. But, as the bit rate increases from 1.5 bpp onwards, majority of the subjects were not able to differentiate the image quality, resulting in a wide range of DMOS values, i.e. 38–100 at 2.9 bpp. Fig. 2(e) shows a trend of the DMOS values decrease as the compression quality increases. Yet, there is a wide range of the DMOS values when the quality level is more than 13, indicating that the subjects were not able to differentiate the images at higher compression quality levels. Fig. 2(f) shows an even wider range of the DMOS values across all JPEG2000 compression ratios.

Table 2 records the values of PLCC, SROCC, KROCC and RMSE, between DMOS and the thirteen FR-IQA metrics. As the correlation coefficient values approach 1, the closer are the subjective DMOS scores to the objective FR-IQA scores. All the PLCC and SROCC values are more than 0.68, except the one between DMOS and UQI with DCT distortion, which is 0.678 but still close to 0.68. Therefore, by referring to Taylor et al. [43], we may conclude that there is a high correlation between DMOS and the thirteen FR-IQA metrics. As shown in Table 2, most of the KROCC values between DMOS and FR-IQAs have a moderate correlation, and they are smaller than the PLCC and SROCC values. The lower KROCC values were also observed in other studies between DMOS and FR-IQA on natural images [36,41]. Therefore, the lower KROCC values could be due to its computation accuracy in showing the correlation. The RMSE values in Table 2 are in reasonable ranges for all the FR-IQAs.

The subjective DMOS versus objective FR-IQA scores were plotted for the six types of distortion: Rician Noise, Gaussian White Noise, Gaussian Blur, DCT, JPEG and JPEG2000 compression. The graphs for all the thirteen FR-IQA metrics have similar trend, hence only NQM and UQI are shown in Figs. 3 and 4 respectively. These two FR-IQA metrics were chosen because NQM has the highest mean correlation coefficient values and the lowest mean RMSE values according to Table 2. On the other hand, UQI has the lowest mean correlation coefficient values and highest mean RMSE values. The nonlinear fitting curve in Figs. 3 and 4 shows that the DMOS values decrease as the FR-IQA scores increase.

Table 3 records the T-test results which investigate statistically any difference among the 13 FR-IQA for different types of distortion. These results are further summarized in Table 4 by recording the frequencies of '1' (means the superiority) for each FR-IQA and classified according to the types of distortion. The numbers in Table 4 represent the frequencies of significant superiority of a FR-IQA metric over the other 12 FR-IQAs. The last row in Table 4 records the FR-IQA with the highest significant superiority for each type of distortion, except Gaussian White Noise which is not clearly distinguishable among several FR-IQAs. The superior IQAs are UQI for Rician noise images, VIF for Gaussian blur images, NQM for both DCT and JPEG compressed images, PSNR for JPEG2000 compressed images. Other T-test results are shown in Table 5 which investigates statistically the 13 FR-IQAs for different types of images and different field strength of MRI scanners. Table 5 is summarized in Table 6, showing the frequencies of superiority for each type of FR-IQA in each type of images and field strength. In Table 6, the superior IQAs are PSNR for 1.5 T PD images, and MSSIM for 3.0 T T1W images. The rest of the image types have no clear distinction in terms of the performance of FR-IQA.

4. Discussion

Rician Noise and Gaussian White Noise cause the low contrast object to be less visible [44], thus affecting the visual quality of the MR images. Gaussian blurring causes small objects and fine details to be less visible [45]. In DCT, JPEG and JPEG2000, the artifacts caused by the compression are not clearly seen by human eyes [25,26,46].

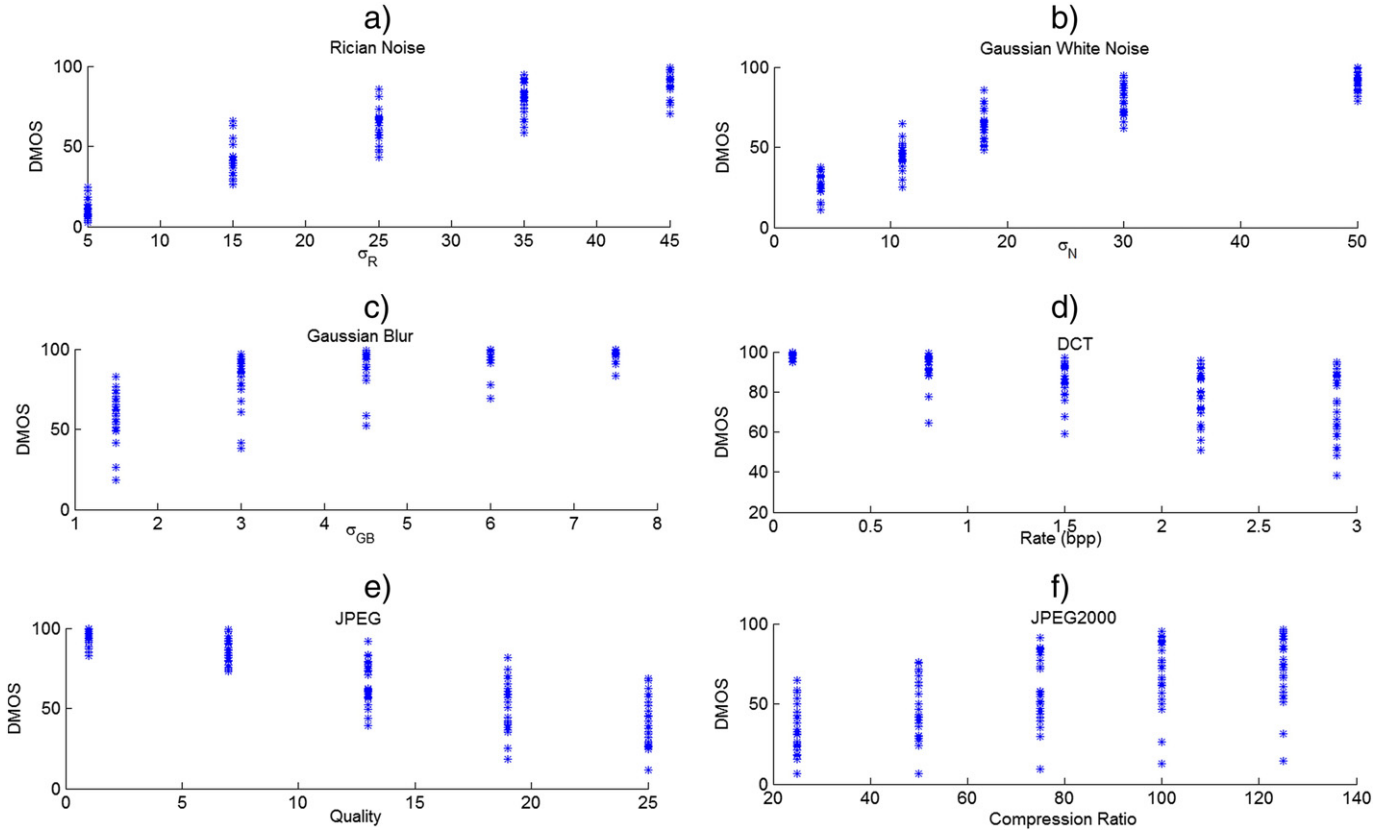

Fig. 2. DMOS values versus: (a) σ_R of Rician Noise, (b) σ_N of Gaussian White Noise, (c) σ_{GB} of Gaussian Blur, (d) compression rate (bpp) of DCT, (e) quality of JPEG, and (f) compression ratio of JPEG2000.

Table 2
PLCC, SROCC, KROCC and RMSE values between DMOS and 13 FR-IQA metrics, for six types of distortion.

	Distortion	FSIM	IFC	IWPSNR	IWSSIM	MSSIM	NQM	PSNR	SNR	SSIM	UQI	VIF	VIFP	WSNR	**Mean**
PLCC	RN	0.949	0.914	0.958	0.925	0.937	0.962	0.958	0.956	0.922	0.866	0.950	0.948	0.938	**0.937**
	GWN	0.943	0.913	0.949	0.922	0.935	0.937	0.948	0.963	0.935	0.862	0.943	0.941	0.951	**0.934**
	GB	0.898	0.913	0.875	0.881	0.895	0.911	0.839	0.821	0.856	0.857	0.891	0.907	0.826	**0.875**
	DCT	0.923	0.781	0.873	0.825	0.852	0.956	0.851	0.887	0.824	0.678	0.847	0.868	0.845	**0.847**
	JPEG	0.907	0.839	0.870	0.823	0.797	0.935	0.765	0.879	0.785	0.788	0.853	0.894	0.908	**0.849**
	JP2K	0.903	0.828	0.819	0.820	0.791	0.913	0.779	0.901	0.796	0.788	0.813	0.867	0.879	**0.838**
	Mean	**0.921**	**0.865**	**0.891**	**0.866**	**0.868**	**0.936**	**0.857**	**0.901**	**0.853**	**0.807**	**0.883**	**0.904**	**0.891**	
SROCC	RN	0.929	0.882	0.935	0.888	0.904	0.953	0.942	0.941	0.893	0.828	0.921	0.925	0.923	**0.913**
	GWN	0.941	0.904	0.927	0.908	0.925	0.934	0.926	0.960	0.926	0.853	0.923	0.935	0.984	**0.927**
	GB	0.918	0.921	0.859	0.917	0.907	0.934	0.833	0.839	0.814	0.870	0.893	0.921	0.865	**0.884**
	DCT	0.925	0.850	0.883	0.861	0.882	0.951	0.853	0.899	0.820	0.758	0.881	0.905	0.908	**0.875**
	JPEG	0.913	0.840	0.869	0.827	0.793	0.940	0.776	0.894	0.780	0.800	0.851	0.900	0.917	**0.854**
	JP2K	0.910	0.825	0.824	0.822	0.795	0.917	0.781	0.899	0.796	0.783	0.815	0.870	0.880	**0.840**
	Mean	**0.923**	**0.87**	**0.883**	**0.871**	**0.868**	**0.938**	**0.852**	**0.905**	**0.838**	**0.815**	**0.881**	**0.909**	**0.913**	
KROCC	RN	0.774	0.697	0.771	0.697	0.718	0.813	0.785	0.792	0.71	0.627	0.741	0.748	0.756	**0.741**
	GWN	0.785	0.721	0.754	0.725	0.749	0.774	0.756	0.82	0.753	0.650	0.749	0.771	0.801	**0.754**
	GB	0.769	0.759	0.670	0.755	0.740	0.781	0.654	0.645	0.626	0.691	0.725	0.765	0.668	**0.711**
	DCT	0.773	0.666	0.707	0.685	0.709	0.823	0.673	0.737	0.637	0.573	0.713	0.741	0.742	**0.706**
	JPEG	0.737	0.632	0.674	0.613	0.579	0.782	0.575	0.721	0.573	0.608	0.649	0.717	0.746	**0.662**
	JP2K	0.729	0.63	0.622	0.622	0.587	0.747	0.575	0.734	0.592	0.596	0.612	0.679	0.715	**0.649**
	Mean	**0.761**	**0.684**	**0.700**	**0.683**	**0.680**	**0.787**	**0.670**	**0.742**	**0.649**	**0.624**	**0.698**	**0.737**	**0.738**	
RMSE	RN	9.279	11.902	8.391	11.169	10.247	8.036	8.381	8.624	11.384	14.679	9.211	9.311	10.205	**10.063**
	GWN	8.109	9.901	7.653	9.432	8.598	8.470	7.702	6.560	8.597	12.303	8.053	8.223	7.516	**8.547**
	GB	7.862	7.284	8.058	8.449	7.979	7.380	9.713	10.213	9.252	9.205	8.110	7.517	10.071	**8.546**
	DCT	5.409	8.778	6.858	7.944	7.350	4.132	7.379	6.484	7.953	10.327	7.474	6.983	7.509	**7.275**
	JPEG	10.002	12.940	11.746	13.506	14.367	8.420	15.312	11.341	14.740	14.944	12.422	10.664	9.978	**12.337**
	JP2K	10.174	13.294	13.609	13.558	14.500	9.685	14.863	10.285	14.352	14.596	13.795	11.802	11.308	**12.755**
	Mean	**8.473**	**10.683**	**9.386**	**10.676**	**10.507**	**7.687**	**10.558**	**8.918**	**11.046**	**12.676**	**9.844**	**9.083**	**9.431**	

RN = Rician Noise, GWN = Gaussian White Noise, GB = Gaussian Blur, DCT = Discrete Cosine Transform, JP2K = JPEG2000.

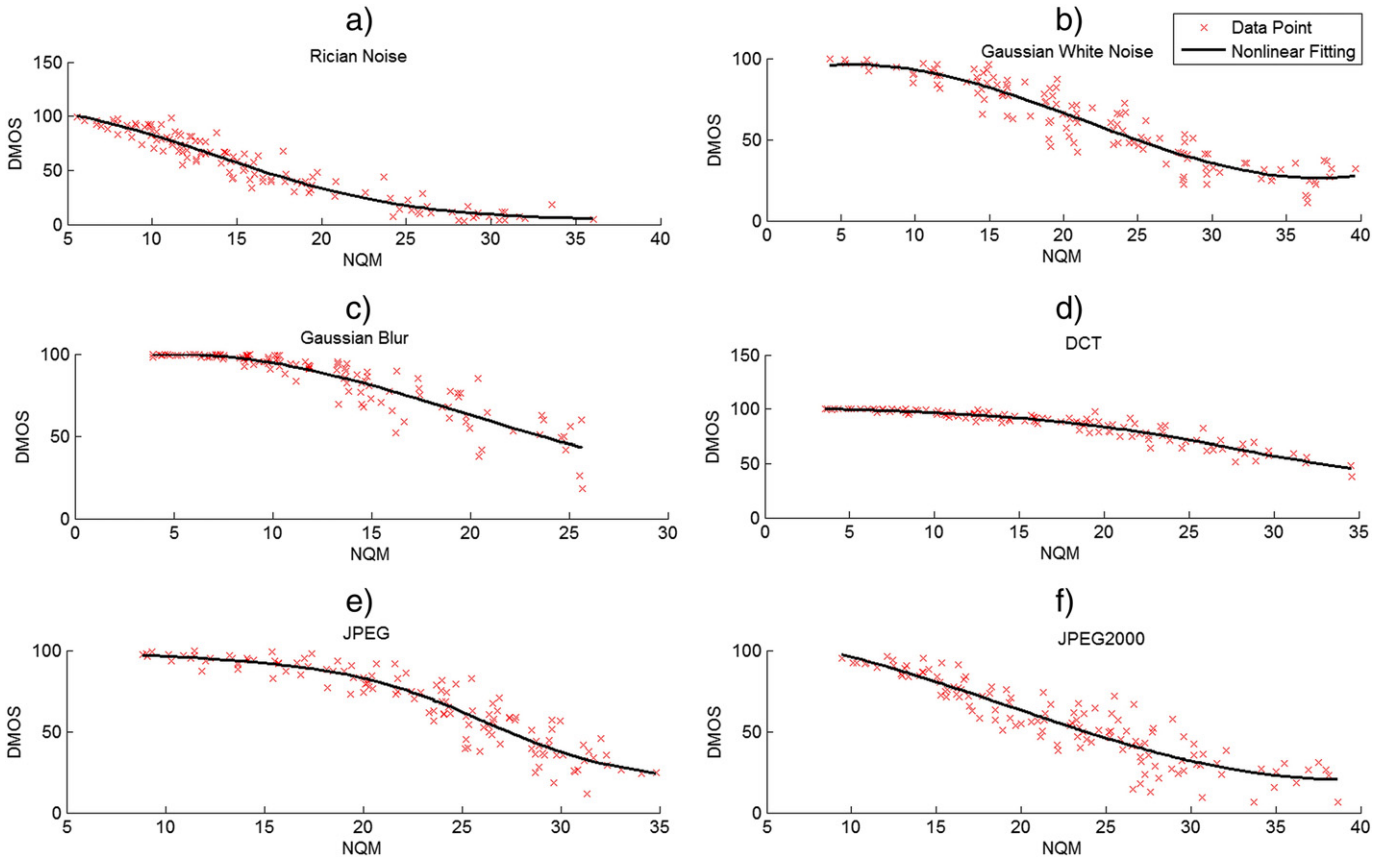

Fig. 3. DMOS versus NQM scores for (a) Rician Noise, (b) Gaussian White Noise, (c) Gaussian Blur, (d) DCT, (e) JPEG, and (f) JPEG2000.

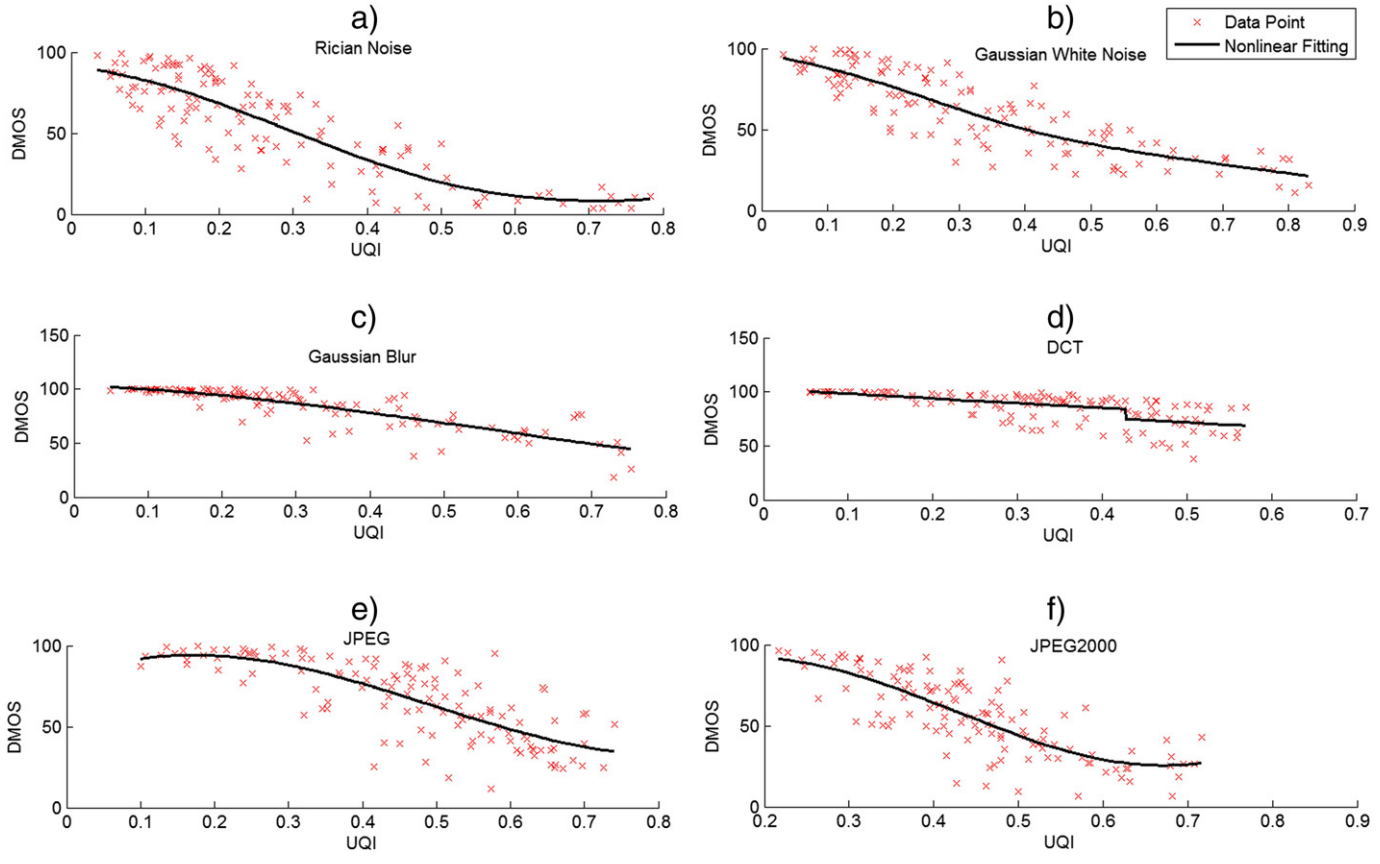

Fig. 4. DMOS versus UQI scores for (a) Rician Noise, (b) Gaussian White Noise, (c) Gaussian Blur, (d) DCT, (e) JPEG, and (f) JPEG2000.

Table 3
Statistical T-test results between the residuals of X_1–DMOS and residuals of another X_2–DMOS, where $X_i \in$ {FSIM, IFC, IWPSNR, IWSSIM, MSSIM, NQM, PSNR, SNR, SSIM, UQI, VIF, VIFP, WSNR}. Each entry contains 6 symbols which are the T-test results for each type of distortion arranged in the following order: Rician Noise, Gaussian White Noise, Gaussian Blur, DCT, JPEG and JPEG2000. Symbol '1' means that the IQA in the row is statistically superior to the IQA in the column. Symbol '0' means that the IQA in the row is statistically inferior to the IQA in the column. Symbol '-' means that both IQAs in the row and column are statistically indistinguishable.

	FSIM	IFC	IWPSNR	IWSSIM	MSSIM	NQM	PSNR	SNR	SSIM	UQI	VIF	VIFP	WSNR
FSIM	- - - - - -	- 0 0 - - -	- 0 0 - - -	- 0 - - 1 -	- 0 - - 1 -	1 - - 0 - -	- 0 - - 1 0	- - 1 1 - -	- 0 0 - - -	0 0 1 1 - -	- 0 0 – 1 -	- 0 - - - -	- - 1 1 - -
IFC	- 1 1 - - -	- - - - - -	- - - - - -	- - 1 0 1 -	- - 1 0 1 -	1 1 1 0 0 -	- - 1 – 1 0	- 1 1 1 - -	- - 0 - - -	- - 1 1 - -	- - 0 0 - -	- - - - - -	- 0 1 - - -
IWPSNR	- 1 1 - - -	- - - - - -	- - - - - -	- - 1 – 1 1	- - 1 – 1 1	1–1 0 0 1	- - 1 – 1 0	- 1 1 1 - -	- - - - - -	0–1 1–1	- - 0 - - -	- - - - - -	- - 1 1–1
IWSSIM	- 1 - - 0 -	- - 0 1 0 -	- - 0 – 0 0	- - - - - -	- - - - 1 -	1 1 - - 0 -	- - - - - 0	- 1 1 1 0 -	- - 0 - 0 -	- - 1 1 0 -	- - 0 - - 0	- - 0 - 0 -	- 1 1 1 0 -
MSSIM	- 1 - - 0 -	- - 0 1 0 -	- - 0 - 0 0	- - - - 0 -	- - - - - -	- 1 - - 0 -	- - - - 0 0	- 1 1 1 0 -	- - 0 - 0 -	0–1 1 0 -	- - 0 - 0 0	- - - - 0 -	- 1 1 1 0 -
NQM	0 - - 1 - -	0 0 0 1 1 -	0–0 1 1 0	0 0 - - 1 -	- 0 - - 1 -	- - - - - -	- 0 - - 1 0	- - 1 1 - -	0–0 - 1 -	0 0 1 1 1 -	0–0 - 1 0	- 0 – 1 - -	- - 1 1 - -
PSNR	- 1 - - 0 1	- - 0 – 0 1	- - 0 – 0 1	- - - - - 1	- - - - 1 1	- 1 - - 0 1	- - - - - -	- 1 1 1 0 1	0–0 - 0 1	0–1 1 0 1	- - 0 - - 1	- - - - 0 1	- - 1 1 0 1
SNR	- - 0 0 - -	- 0 0 0 - -	- 0 0 0 - -	- 0 0 0 1 -	- 0 0 0 1 -	- - 0 0 - -	- 0 0 0 1 0	- - - - - -	0 0 0 0 - -	0 0–1 - -	- 0 0 0 1 -	- 0 0 0 - -	- - - - - -
SSIM	- 1 1 - - -	- - - - - -	- - - - - -	- - 1 – 1 -	- - 1 – 1 -	1–1 – 0 -	1–1 - 1 0	1 1 1 1 - -	- - - - - -	- - 1 1 - -	- - - - - -	1–1 - - -	- - 1 1 - -
UQI	1 1 0 0 - -	- - 0 0 - -	1–0 0–0	- - 0 0 1 -	1–0 0 1 -	1 1 0 0 0 -	1–0 0 1 0	1 1–0 - -	- - 0 0 - -	- - - - - -	1–0 0 - -	1–0 0 - -	1 1–0 - -
VIF	- 1 1 - 0 -	- - 1 1 - -	- - 1 - - -	- - 1 - - 1	- - 1 – 1 1	1–1 – 0 1	- - 1 - - 0	- 1 1 1 0 -	- - - - - -	0–1 1 - -	- - - - - -	- - 1 - 0 -	- - 1 1 0 -
VIFP	- 1 - - - -	- - - - - -	- - - - - -	- - 1 – 1 -	- - - - 1 -	- 1 – 0 - -	- - - - 1 0	- 1 1 1 - -	0–0 - - -	0–1 1 - -	- - 0 - 1 -	- - - - - -	- - 1 1 - -
WSNR	- - 0 0 - -	- 0 0 - - -	- - 0 0–0	- 0 0 0 1 -	- 0 0 0 1 -	- - 0 0 - -	- - 0 0 1 0	- - - - - -	- - 0 0 - -	0 0–1 - -	- - 0 0 1 -	- - 0 0 - -	- - - - - -

Therefore, the differences between the reference and distorted images with Gaussian Blur, or DCT, JPEG and JPEG2000 compressions are very subtle and may not be noticed by human eyes. This fact is further proven by the scatter plots of DMOS versus DCT, JPEG and JPEG2000 compression in Fig. 2(d)–(f) where there is a wide range of DMOS values irrespective of the levels of compression. Refer to Table 2, the DCT, JPEG and JPEG2000 distortion show lower mean values of correlation coefficient, indicating a slightly bigger gap between the subjective and objective ratings on images distorted with these kinds of compression. Nevertheless, it also means that these compressions are able to produce good images unnoticeable by human even after being subjected to high compression rate. Thus, these compressions are suitable to compress a wide range of MR images[21–26]. For these three types of compressions, we recommend to use the objective assessment to evaluate the image quality because our results showed that the human eyes are not able to differentiate the compressed images.

In Table 2, we observe that NQM has the highest mean value of correlation coefficients (PLCC, SROCC and KROCC) and lowest mean value of RMSE as compared with DMOS. In other word, NQM is the closest FR-IQA metric to human judgement compared to other FR-IQA metrics used in our study. NQM evaluates the image quality based on image degradation model, and it separates the impact of frequency distortion and noise injection on Human Visual System (HVS). It was proven to be an excellent IQA to assess the image quality on natural images [47]. On the other hand, UQI has the lowest mean value of correlation coefficients (PLCC, SROCC and KROCC) and the highest mean value of RMSE with DMOS. This indicates that UQI is the least similar to the human judgment, which was also reported by Wang et al. who verified that UQI is an unsuccessful FR-IQA method that fails to correlate with all the subjective assessment [35]. Besides that, UQI is unstable when the mean and standard deviation of the image intensity are close to zero [35,47].

Table 4
Summary of the T-test results from Table 3, classified according to the types of distortion, accessed by 13 FR-IQA metrics. The numbers in each entry represent the frequencies of significant superiority of a FR-IQA metric over the other 12 FR-IQAs.

	Types of Distortion					
	RC	GWN	GB	DCT	JPEG	JP2K
FSIM	1	0	3	3	4	0
IFC	1	4	7	2	3	0
IWPSNR	1	2	7	3	3	5
IWSSIM	1	4	3	4	1	0
MSSIM	0	4	3	4	0	0
NQM	0	0	3	7	8	0
PSNR	0	3	3	3	1	12
SNR	0	0	1	1	4	0
SSIM	4	2	9	3	3	0
UQI	9	4	1	0	3	0
VIF	1	2	10	4	1	3
VIFP	0	3	4	3	4	0
WSNR	0	0	1	1	4	0
Superior IQA	**UQI**	**-**	**VIF**	**NQM**	**NQM**	**PSNR**

The results of the statistical significance testing in Table 4 classified according to the types of distortion gave a different result from the above. According to Table 4, the superior IQAs tested statistically are UQI for Rician noise images, VIF for Gaussian blur images, NQM for both DCT and JPEG compressed images, PSNR for JPEG2000 compressed images. Nevertheless, a statistically superior IQA of a certain distortion may be computationally inferior for different types of images. The statistical test may also vary with the amount of tested images. It was reported that 100 sample points will only provide 50% chance of detecting the difference between the performance of two IQAs; whereas 200 samples points will increase the chances to 75% [48]. In our first T-test to study the performance of FR-IQAs over different types of distortion, there were 125 sample points. Therefore the probability of detecting the difference between two IQAs is between 50% and 75% in our study. In the second T-test to study the performance of FR-IQAs over different types of image, the sample points vary from 30 to 200 due to unequal selection of image types. As a result, we are not able to find the best performed FR-IQA for each types of image, as summarized in Table 6.

In our study, the maximum score that represents excellent image quality is 90 but not 100 because there is no gold standard to confirm that the MR image is perfect enough to be rated as 100. The reference images used in this study might be subjected to a small degree of distortion. In fact, the FR-IQA metrics can only provide a relative measure of the image quality for various distorted images compared to the so-called 'reference image'. Apparently, FR-IQA is not the best way to evaluate MR images, but NR-IQA is more suitable for MR images. However, there are many challenges in designing a

Table 5
Statistical T-test results between the residuals of X_1–DMOS and residuals of another X_2–DMOS, grouped according to different types of images. Each entry contains 6 symbols which are the T-test results for each type of images arranged in the following order: 1.5 T T1W, 1.5 T T2W, 1.5 T PD, 3.0 T T1W, 3.0 T T2W and 3.0 T PD. Symbols '1', '0', '-' are explained in Table 3.

	FSIM	IFC	IWPSNR	IWSSIM	MSSIM	NQM	PSNR	SNR	SSIM	UQI	VIF	VIFP	WSNR
FSIM	- - - - - -	- 1 1 0 1 0	- 1 - 0 1 1	0 1 1 0 1 -	0 1–0 - -	0–1 - - 1	0 1 - - - 1	0 - - - - 1	- 1 - 0 - 1	0 1 1 0 1 1	- 1 1 0 1 1	0 1 1 0–1	0 - - 0 1 -
IFC	- 0 0 1 0 1	- - - - - -	- - 0 - 0 1	- - 0 - - -	0–0 0 0 1	0 0 0 1 0 1	0–0 1 0 1	0 0 0 1 0 1	- - 0 1 0 1	0 - - - - -	- - 0 - 0 1	- - 0 - 0 1	0 0 0 - - 1
IWPSNR	- 0 - 1 0 0	- - 1 - 1 0	- - - - - -	- - 1 0–0	0 - - 0 - 0	0 0 1 1 0 -	- - - 1 0 -	0 0–1 0 -	- - - 1 0 -	0 0 1 - - 0	- - 1 - - -	- 0 1 - - -	0 0 - - - 0
IWSSIM	1 0 0 1 0 -	- - 1 - - -	- - 0 1–1	- - - - - -	- - - - - -	- 0 - 1 0 1	- - 0 1 0 1	0 0–1 0 1	1 - - 1 0 1	0–1 - - 0	- - - 1 - 1	- - - 1 0 1	0 0–1 - -
MSSIM	1 0–1 - -	1–1 1 1 0	1 - - 1 - 1	- - - - - -	- - - - - -	- 0 - 1 0 1	- - 0 1 0 1	0 0–1 0 1	1 - - 1 0 1	0–1 - - 0	1–1 1–1	- - 1 1–1	0 0–1 1 -
NQM	1–0 - - 0	1 1 1 0 1 0	1 1 0 0 1 -	- 1 - 0 1 0	- 1 - 0 1 0	- - - - - -	- 1 0 0 - -	0 - - - - 0	1 1–0 - -	0 1 1 0 1 0	1 1–0 1 0	- 1 1 0 1 0	0 - - 0 1 0
PSNR	1 0 - - - 0	1–1 0 1 0	- - - 0 1 -	- - 1 0 1 0	- - 1 0 1 0	- 0 1 1 - -	- - - - - -	0 0 1 - - 0	1–1 - - -	0–1 0 1 0	- - 1 0 1 0	- - 1 - 1 0	0 0 1 0 1 0
SNR	1 - - - - 0	1 1 1 0 1 0	1 1–0 1 -	1 1–0 1 0	1 1–0 1 0	1 - - - - 1	1 1 0 - - 1	- - - - - -	1 1 - - - 1	- 1 1 0 1 0	1 1 1 0 1 -	1 1 1 0 - -	- - - 0 1 0
SSIM	- 0 - 1 - 0	- - 1 0 1 0	- - - 0 1 -	0 - - 0 1 0	0 - - 0 1 0	0 0–1 - -	0–0 - - -	0 0 - - - 0	- - - - - -	0 0 1 0 1 0	- - 1 0 1 0	0 0 1 - - -	0 0 - - 1 0
UQI	1 0 0 1 0 0	1 - - - - -	1 1 0 - - 1	1–0 - - 1	1–0 - - 1	1 0 0 1 0 1	1–0 1 0 1	- 0 0 1 0 1	1 1 0 1 0 1	- - - - - -	1 1 0 - - 1	1–0 1 0 1	- 0 0 - - 1
VIF	- 0 0 1 0 0	- - 1 - 1 0	- - 0 - - -	- - - 0 - 0	0–0 0–0	0 0–1 0 1	- - 0 1 0 1	0 0 0 1 0 -	- - 0 1 0 1	0 0 1 - - 0	- - - - - -	- - - - - -	0 0 - - - 0
VIFP	1 0 0 1–0	- - 1 - 1 0	- 1 0 - - -	- - - 0 1 0	- - 0 0–0	- 0 0 1 0 1	- - 0 - 0 1	0 0 0 1 - -	1 1 0 - - -	0–1 0 1 0	- - - - - -	- - - - - -	0 0 0–1 0
WSNR	1 - - 1 0 -	1 1 1 - - 0	1 1 - - - 1	1 1–0 - -	1 1–0 0 -	1 - - 1 0 1	1 1 0 1 0 1	- - - 1 0 1	1 1 - - 0 1	- 1 1 - - 0	1 1 - - - 1	1 1 1–0 1	- - - - - -

suitable NR-IQA method for medical images with complex geometrical structure of human anatomy and scanning artifacts due to various reasons.

To the best of our knowledge, we have presented the largest study of subjective assessment of MR images distorted with six types of distortion. And we have produced the MR database for IQA study, which consists of 25 original reference MR images and 750 distorted images, along with their DMOS values evaluated by 28 volunteers. This will add up to a total of 21,700 human evaluations on MR images for various anatomical human parts with various distortion types and levels. The DMOS values consist of the human judgment on various distorted images, which could be useful in calculating the image features of the MR images. Therefore, it will be used for our future study in designing a suitable NR-IQA metrics for MR images.

5. Conclusion

We have presented a database of 775 MR images inclusive of 25 reference images and 750 distorted images with six types of distortion: Rician Noise, Gaussian White Noise, Gaussian Blur, DCT compression, JPEG compression and JPEG2000 compression. The database also contains the DMOS values calculated from the raw scores obtained from the subjective evaluation on MR images. We validated the subjective DMOS with thirteen objective FR-IQA metrics with the high PLCC and SROCC values and low RMSE values. Hence, the DMOS values calculated in our study are applicable for our future study to model a new NR-IQA method.

Table 6
Summary of the T-test results from Table 5, classified according to the types of image and MRI field strength, accessed by 13 FR-IQA metrics. The numbers in each entry represent the frequencies of significant superiority of a FR-IQA metric over the other 12 FR-IQAs.

	Types of Image					
	1.5 T T1W	1.5 T T2W	1.5 T PD	3.0 T T1W	3.0 T T2W	3.0 T PD
FSIM	0	9	6	0	6	8
IFC	0	0	0	5	0	10
IWPSNR	0	0	6	5	1	0
IWSSIM	2	0	2	9	0	7
MSSIM	5	0	4	10	2	7
NQM	5	9	3	0	8	0
PSNR	3	0	10	1	8	0
SNR	10	9	4	0	7	3
SSIM	0	0	4	2	7	0
UQI	10	3	0	6	0	10
VIF	0	0	2	5	1	3
VIFP	2	2	2	3	4	2
WSNR	10	9	3	4	0	7
Superior IQA	**-**	**-**	**PSNR**	**MSSIM**	**-**	**-**

Acknowledgment

This research was funded by BKP grant (BK053-2014) from the University of Malaya. We like to thank all the volunteers involved in the subjective evaluation in this study and our colleagues Lim Chern Loon and Yu Yong Poh for their technical opinion. We also acknowledge Alzheimer's Disease Neuroimaging Initiative (ADNI) and Osirix DICOM Viewer MRI for sharing the MR images in this study. Data collection and sharing for this project were funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen Idec Inc.; Bristol-Myers Squibb Company; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd. and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; ; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Medpace, Inc.; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Synarc Inc.; and Takeda Pharmaceutical Company. The Canadian Institutes of Rev. December 5, 2013 Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

Appendix A

The table below contains all the formulas for the FR-IQA metrics used in this study. Let $r(x,y)$ represent the reference image and $t(x,y)$ represent the distorted image. n_x and n_y are the size of the image in pixels across x and y dimensions.

No	IQA Algorithm	Description
1	Signal-to-Noise Ratio (SNR) [33]	Ratio of average signal power to average noise power. $SNR = 10log_{10}\left[\frac{\sum_1^{n_x}\sum_1^{n_y}[r(x,y)]^2}{\sum_1^{n_x}\sum_1^{n_y}[r(x,y)-t(x,y)]^2}\right]$ (A.1)
2	Peak Signal-to-Noise Ratio (PSNR) [33,34]	Ratio of peak signal power to average noise power. $PSNR = 10log_{10}\left[\frac{max(r(x,y))^2}{\frac{1}{n_x.n_y}\sum_1^{n_x}\sum_1^{n_y}[r(x,y)-t(x,y)]^2}\right]$ (A.2)
3	Structural Similarity Index Metrics (SSIM) [35]	Captures the loss in the structure of the image. $SSIM = \frac{(2\mu_r\mu_t+C_1)(2\sigma_{rt}+C_2)}{(\mu_r^2+\mu_t^2+C_1)(\sigma_r^2+\sigma_t^2+C_2)}$ (A.3) where μ_r and μ_t are the mean intensity for the reference and distorted images respectively; σ_r and σ_t are the standard deviation for the reference and distorted images respectively; σ_{rt} is estimated as: $\sigma_{rt} = \frac{1}{N-1}\sum_{i=1}^{N}(r_i-\mu_r)(t_i-\mu_t)$ (A.4) where $C_1=(K_1L)^2$ and $C_2=(K_2L)^2$ where L is the dynamic range of the pixels values (i.e. 255 for 8-bit grayscale images, as in our case), $K_1 = 0.01$ and $K_2 = 0.03$.
4	Multiscale SSIM (MS-SSIM) [35]	Mean of SSIM that evaluates overall image quality by using a single overall quality. $MSSIM(r,t) = \frac{1}{M}\sum_{j=1}^{M}SSIM(r_j,t_j)$ (A.5)
5	Feature SIMilarity (FSIM) [36]	A low-level feature based image quality assessment which used two types of features: Phase Congruency (PC) and Gradient Magnitude (GM). Ω represents the whole image spatial domain. $FSIM(r,t) = \frac{\sum_{x\in\Omega}S_L(x).PC_m(x)}{\sum_{x\in\Omega}PC_m(x)}$ (A.6) where $PC_m(x) = \max(PC_r(x).PC_t(x))$ (A.7) $S_L=[S_{PC}(x)]^{\alpha}[S_G(x)]^{\beta}$ (A.8) where $S_{PC}(x) = \frac{2PC_r(x).PC_t(x)+T_1}{PC_r^2(x)+PC_t^2(x)+T_1}$ (A.9) $S_G(x) = \frac{2G_r(x).G_t(x)+T_2}{G_r^2(x)+G_t^2(x)+T_2}$ (A.10)
6	Information Fidelity Criterion (IFC) [37]	An information theoretic criterion for image fidelity where it uses the source and distortion models to compute the mutual information between the reference and the distorted images. $IFC = \sum_{k\in subbands} I((r^{N_k,k};t^{N_k,k}\|s^{N_k,k}))$ (A.11) where $r^{N_k,k}$ denotes N_k coefficients from the Random Field (RF), r_k of the k^{th} subband; and similarly for $t^{N_k,k}$ and $s^{N_k,k}$ (RF of positive scalar of reference image)
7	Noise Quality Measure (NQM) [38]	A measure of additive noise. It is designed based on Peli's contrast pyramid. $NQM(dB) = 10\log_{10}(\frac{\sum_x\sum_y O_s^2(x,y)}{\sum_x\sum_y(O_s(x,y)-I_s(x,y))^2})$ (A.12) where $O_s(x,y)$ and $I_s(x,y)$ represent the simulated versions of the model restored image and the restored images, respectively.
9	Weighted SNR (WSNR) [38]	Ratio of the average weighted signal power to the averaged weighted noise power. $WSNR = 10\log_{10}(\frac{\sum_{\omega_1}\sum_{\omega_2}\omega_2\|R(\omega_1,\omega_2)C(\omega_1,\omega_2)\|^2}{\sum_{\omega_1}\sum_{\omega_2}\|T(\omega_1,\omega_2)C(\omega_1,\omega_2)\|^2})$ (A.13) where $C(\omega_1,\omega_2)$ is the lowpass CSF, and $R(\omega_1,\omega_2)$ and $T(\omega_1,\omega_2)$ are the discrete Fourier transform of the original and noise images, respectively.
10	Visual Information Fidelity (VIF) [39]	Measures image information by computing two mutual information quantities from the reference and distorted images. $VIF = \frac{\sum_{j\in subbands}I(\vec{C}^{N,j};\vec{T}^{N,j}\|s^{N,j})}{\sum_{j\in subbands}I(\vec{C}^{N,j};\vec{R}^{N,j}\|s^{N,j})}$ (A.14) where the subbands of interest are summed over, and $\vec{T}^{N,j}$ represents the subband in the test image, $\vec{R}^{N,j}$ represents the subband in the reference image, $\vec{C}^{N,j}$ represents N elements of the RF C_j that describes the coefficient subband j, and so on.
11	Pixel Visual Information Fidelity (VIFP) [39]	Pixel domain version of VIF. It uses scalar RF model, not vector version like VIF. (A.15)
12	Universal Image Quality Index (UQI) [40]	Computes the loss of correlation, luminance distortion and contrast distortion in distorted image. $Q = \frac{4\,\sigma_{rt}\bar{r}\bar{t}}{(\sigma_r^2+\sigma_t^2)[(\bar{r})^2+(\bar{t})^2]}$ (A.16) where $\bar{r}$ and $\bar{t}$ are the means of the reference and test image, respectively, and σ_r^2 and σ_t^2 are the standard deviations of the reference and test image, respectively $\sigma_{rt} = \frac{1}{N-1}\sum_{i=1}^{N}(r_i-\bar{r})(t_i-\bar{t})$ (A.17)
13	Information Weighted PSNR (IW-PSNR) [41]	Uses the Laplacian pyramid transform domain information content weights. $IW-MSE = \prod_{j=1}^{M}[\frac{\sum_i\omega_{j,i}(r_{j,i}-t_{j,i})^2}{\sum_i\omega_{j,i}}]^{\beta_j}$ (A.18) $IW-PSNR = 10\log_{10}(\frac{max(r(x,y))^2}{IW-MSE})$ (A.19) where $\omega_{j,i}$ is the information content weight computed at the corresponding location, M is the number of scales and β_j is the weight given to the j^{th} scale.

(*continued*)

No	IQA Algorithm	Description
14	Information Weighted SSIM (IW-SSIM) [41]	Obtained by combining content weighting with MS-SSIM. $IW-SSIM = \frac{\sum_i \omega_{j,i}\, c(r_{j,i},t_{j,i}) s(r_{j,i},t_{j,i})}{\sum_i \omega_{j,i}}$ (A.20) where $c(r_{j,i},t_{j,i}) = \frac{2\sigma_r\sigma_t+C_2}{\sigma_r^2+\sigma_t^2+C_2}$ (A.21) and $s(r_{j,i},t_{j,i}) = \frac{\sigma_{rt}+C_3}{\sigma_r\sigma_t+C_3}$ (A.22) where σ_r and σ_t are the standard deviation for the reference and distorted images respectively; σ_{rt} is estimated as: $\sigma_{rt} = \frac{1}{N-1}\sum_{i=1}^{N}(r_i-\mu_r)(t_i-\mu_t)$ (A.23) where $C_1=(K_1L)^2$, $C_3=(K_2L)^2/2$ where L is the dynamic range of the pixels values (i.e. 255 for 8-bit grayscale images, as in our case), $K_1 = 0.01$ and $K_2 = 0.03$.

Appendix B

The table below contains all the formulas for the performance metrics used in this study. D represents the DMOS values, Q is the original objective scores calculated from the FR-IQA metrics, and Q_r is the objective scores after regression.

No	IQA Algorithm	Description
1	Logistic Regression [9,36,49]:	$Q_r = \beta_1\left(\frac{1}{2}-\frac{1}{1+\exp(\beta_2(Q-\beta_3))}\right)+\beta_4 Q+\beta_5$ (B.1) where β_1,β_2,β_3,β_4,β_5 are the regression model parameters. Optimal parameters, β are obtained using nonlinear least squares.
2	Pearson Linear Correlation Coefficient (PLCC) [50]	$PLCC(Q_r,D) = \frac{\sum_i^n (Q_{ri}-\overline{Q_r})\sum_i^n (D_i-\overline{D})}{\sqrt{\sum_i^n (Q_{ri}-\overline{Q_r})^2}\sqrt{\sum_i^n (D_i-\overline{D})^2}}$ (B.2) where $\overline{Q_r}$ and $\overline{D}$ are the means for dataset Q_r and D respectively.
3	Spearman Rank Order Correlation Coefficient (SROCC) [51]	$SROCC(Q,D) = 1-\frac{6\sum_{i=1}^n d_i^2}{n(n-1)}$ (B.3) where d_i is the difference between the ranks of each pair of values in Q and D; and n is the total number of data pairs.
4	Kendall Rank Order Correlation Coefficient (KROCC) [41]	$KROCC(Q,D) = \frac{N_c-N_d}{\frac{1}{2}N(N-1)}$ (B.4) where N_c and N_d represent the numbers of concordant (ordered in the same way) and discordant (ordered differently) pairs in the data sets, respectively.
5	Root Mean Square Error (RMSE) [52]	$RMSE(Q_r,D) = \sqrt{\frac{\sum_i^n (Q_{ri}-D_i)^2}{n}}$ (B.5) where n is the total number of data pairs.

References

[1] Gulame M, Joshi KR, Kamthe RS. A full reference based objective image quality assessment. Int J Adv Electr Electron Eng 2013;2:13–8.
[2] Chandler DM. Seven challenges in image quality assessment: past, present, and future research. ISRN Signal Process 2013; 2013. p. 1–53. http://dx.doi.org/10.1155/2013/905685.
[3] Kumar R, Rattan M. Analysis of various quality metrics for medical image processing. Int J Adv Res Comput Sci Softw Eng 2012;2:137–44.
[4] Kumar B, Singh SP, Mohan A, Anand A. Performance of quality metrics for compressed medical images through mean opinion score prediction. J Med Imaging Health Inform 2012;2:188–94. http://dx.doi.org/10.1166/jmihi. 2012.1083.
[5] Mortamet B. Bernstein M a., Jack CR, Gunter JL, Ward C, britson PJ, et al. Automatic quality assessment in structural brain magnetic resonance imaging. Magn Reson Med 2009;62:365–72. http://dx.doi.org/10.1002/mrm.21992.
[6] Woodard JP, Carley-Spencer MP. No-reference image quality metrics for structural MRI. Neuroinformatics 2006;4:243–62. http://dx.doi.org/10.1385/NI:4:3:243.
[7] Tisdall MD, Atkins MS. Using human and model performance to compare MRI reconstructions. IEEE Trans Med Imaging 2006;25:1510–7. http://dx.doi.org/10.1109/TMI.2006.881374.
[8] Prieto F, Guarini M, Tejos C, Irarrazaval P. Metrics for quantifying the quality of MR images. Proc. 17th Annu. Meet. ISMRM, vol. 17; 2009. p. 4696.
[9] Sheikh HR, Sabir MF, Bovik AC. A statistical evaluation of recent full reference image quality assessment algorithms. Image Process IEEE Trans 2006;15:3441–52.
[10] Larson EC, Chandler DM. Most apparent distortion: full-reference image quality assessment and the role of strategy. J Electron Imaging 2010;19. http://dx.doi.org/10.1117/12.810071 [011006–011006–21].
[11] Chandler DM, Hemami SS. VSNR : a visual signal-to-noise ratio for natural images. IEEE Trans Image Process 2007;16:2284–98.
[12] Patrick Le Callet FA. Subjective quality assessment IRCCyN/IVC database. http://www.irccyn.ec-nantes.fr/ivcdb/; 2005.
[13] Sazzad ZMP, Kawayoke Y, Horita Y. Image quality evaluation database n.d. http://mict.eng.u-toyama.ac.jp/databasetoyama/.
[14] Ponomarenko N, Lukin V, Egiazarian K, Astola J, Carli M, Battisti F. Color image database for evaluation of image quality metrics. Multimed. Signal Process2008 IEEE 10th work, IEEE; 2008. p. 403–8.
[15] Ponomarenko N, Jin L, Ieremeiev O, Lukin V, Egiazarian K, Astola J, et al. Image database TID2013: peculiarities, results and perspectives. Signal Process Image Commun 2015;30:57–77. http://dx.doi.org/10.1016/j.image.2014.10.009.
[16] MR images from Osirix DICOM viewer n.d. http://www.osirix-viewer.com/datasets/ (accessed January 20, 2015).
[17] Alzheimer's Disease Neuroimaging Initiative n.d. http://www.adni-info.org/ (accessed February 9, 2010).
[18] Gudbjartsson H, Patz S. The Rician distribution of noisy MRI data. Magn Reson Med 1995;34:910–4.
[19] Tong MY. Restoration of images in the presence of rician noise and in the presence of atmospheric turbulence. Los Angeles: University of California; 2012.
[20] Debnath A, Rai HM, Yadav C, Bhatia A. Deblurring and denoising of magnetic resonance images using blind deconvolution method. Int J Comput Appl 2013;81:7–12. http://dx.doi.org/10.5120/14046-2209.
[21] Parameswaran AP, Gaonkar M. DCT and DWT in medical image compression. Int J Adv Comput Theory Eng 2013;2:2319–526.
[22] M.E SS, Vijayakuymar VR, Anuja R. A survey on various compression methods for medical images. Int J Intell Syst Appl 2012;4:13–9. http://dx.doi.org/10.5815/ijisa.2012.03.02.
[23] Sudha MVK, Sudhakar R. Two dimensional medical image compression techniques—a survey. Int J Graph Vis Image Process 2011;11:9–20.

[24] Lustig M, Donoho DL, Santos JM, Pauly JM. Compressed sensing MRI: a look at how CS can improve on current imaging techniques. IEEE Signal Process Mag 2008;25:72–82. http://dx.doi.org/10.1109/MSP.2007.914728.
[25] Yamamoto LG. Using JPEG image compression to facilitate telemedicine. Am J Emerg Med 1995;13:55–7. http://dx.doi.org/10.1016/0735-6757(95)90244-9.
[26] Skodras A, Christopoulos C, Ebrahimi T. The JPEG 2000 still image compression standard. IEEE Signal Process Mag 2001;18:36–58. http://dx.doi.org/10.1109/79.952804.
[27] Recommendation ITURBT. 500-11. Methodology for the subjective assessment of the quality of television pictures 2000; 2002.
[28] Bindu K, Ganpati A, Sharma AK. A comparative study of image compression algorithms. Int J Res Comput Sci 2012;2:37–42. http://dx.doi.org/10.7815/ijorcs.25.2012.046.
[29] Wajid R, Bin Mansoor A, Pedersen M. A study of human perception similarity for image quality assessment. Colour Vis Comput Symp (CVCS) 2013;2013:1–6. http://dx.doi.org/10.1109/CVCS.2013.6626276.
[30] De Angelis A, Moschitta A, Russo F, Carbone P. Image quality assessment: an overview and some metrological considerations. Adv Methods Uncertain Estim Meas2007 IEEE Int. Work; 2007. p. 47–52.
[31] Thorpe L, Shelton B. Subjective test methodology: MOS vs. DMOS in evaluation of speech coding algorithms. Speech Coding Telecommun. 1993Proceedings., IEEE Work; 1993. p. 73–4.
[32] Rajagopal H, Chow LS, Paramesran R. Subjective versus objective assessment for magnetic resonance (MR) images. ICCITE 2015 17th Int. Conf. Commun. Inf. Technol. Eng., n.d.
[33] Gonzalez RC, Woods RE. Digital image processing. 3rd ed. Prentice-Hall, Inc.; 2006
[34] Wang Z, Bovik AC. Error : love it or leave it ? IEEE Signal Process Mag 2009;98–117.
[35] Wang Z, Bovik AC, Sheikh HR, Simoncelli EP. Image quality assessment: from error visibility to structural similarity. Image Process IEEE Trans 2004;13:600–12. http://dx.doi.org/10.1109/TIP.2003.819861.
[36] Zhang L, Zhang D, Mou X. FSIM: a feature similarity index for image quality assessment. Image Process IEEE Trans 2011;20:2378–86.
[37] Sheikh HR, Bovik AC, De Veciana G. An information fidelity criterion for image quality assessment using natural scene statistics. Image Process IEEE Trans 2005;14:2117–28.
[38] Damera-Venkata N, Kite TD, Geisler WS, Evans BL, Bovik AC. Image quality assessment based on a degradation model. IEEE Trans Image Process 2000;9:636–50. http://dx.doi.org/10.1109/83.841940.
[39] Sheikh HR, Bovik AC. Image information and visual quality. IEEE Trans Image Process 2006;15:430–44. http://dx.doi.org/10.1109/tip.2005.859378.
[40] Zhou W, Bovik AC. A universal image quality index. Signal Process Lett IEEE 2002;9:81–4. http://dx.doi.org/10.1109/97.995823.
[41] Wang Z, Li Q. Information content weighting for perceptual image quality assessment. IEEE Trans Image Process 2011;20:1185–98. http://dx.doi.org/10.1109/TIP.2010.2092435.
[42] Rohaly AM, Corriveau PJ, Libert JM, Webster AA, Baroncini V, Beerends J, et al. Video quality experts group: current results and future directions. SPIEVisual Commun. Image Process.; 2000. p. 742–53.
[43] Taylor R. Interpretation of the correlation coefficient: a basic review. J Diagn Med Sonogr 1990;6:35–9. http://dx.doi.org/10.1177/875647939000600106.
[44] Vibhakar A, Tiwari M, Singh J. Performance analysis for MRI Denoising using intensity averaging Gaussian blur concept and its comparison with wavelet transform method. Int J Comput Appl 2012;58:21–6.
[45] Ertas M, Yildirim I, Kamasak M, Akan A. An iterative tomosynthesis reconstruction using total variation combined with non-local means filtering. Biomed Eng Online 2014;13:65. http://dx.doi.org/10.1186/1475-925X-13-65.
[46] Watson AB. Image compression using the discrete cosine transform. Math J 1994;4:81–8. http://dx.doi.org/10.1016/0165-1684(90)90166-V.
[47] Samajdar T, Quraishi MI. Analysis and evaluation of image quality metrics. Inf. Syst. Des. Intell. Appl. India: Springer; 2015. p. 369–78. http://dx.doi.org/10.1007/978-81-322-2247-7_38.
[48] Montgomery DC, Runger GC. Applied statistics and probability for engineers. 3rd ed. John Wiley & Sons, Inc.; 2003
[49] Xue W, Zhang L, Mou X, Bovik AC. Gradient magnitude similarity deviation: a highly efficient perceptual image quality index. Image Process IEEE Trans 2014;23:684–95.
[50] Song X-K. Correlated data analysis: modeling, analytics, and applications. Springer Science & Business Media; 2007.
[51] Gautheir TD. Detecting trends using Spearman's rank correlation coefficient. Environ Forensics 2001;2:359–62. http://dx.doi.org/10.1080/713848278.
[52] Chai T, Draxler RR. Root mean square error (RMSE) or mean absolute error (MAE)?—arguments against avoiding RMSE in the literature. Geosci Model Dev 2014;7:1247–50. http://dx.doi.org/10.5194/gmd-7-1247-2014.