Comparison of Objective Image Quality Metrics to Expert Radiologists' Scoring of Diagnostic Quality of MR Images

Allister Mason, James Rioux, Sharon E. Clarke, Andreu Costa, Matthias Schmidt, Valerie Keough, Thien Huynh, and Steven Beyea

***Abstract*— Image quality metrics (IQMs) such as root mean square error (RMSE) and structural similarity index (SSIM) are commonly used in the evaluation and optimization of accelerated magnetic resonance imaging (MRI) acquisition and reconstruction strategies. However, it is unknown how well these indices relate to a radiologist's perception of diagnostic image quality. In this study, we compare the image quality scores of five radiologists with the RMSE, SSIM, and other potentially useful IQMs: peak signal to noise ratio (PSNR) multi-scale SSIM (MSSSIM), information-weighted SSIM (IWSSIM), gradient magnitude similarity deviation (GMSD), feature similarity index (FSIM), high dynamic range visible difference predictor (HDRVDP), noise quality metric (NQM), and visual information fidelity (VIF). The comparison uses a database of MR images of the brain and abdomen that have been retrospectively degraded by noise, blurring, undersampling, motion, and wavelet compression for a total of 414 degraded images. A total of 1017 subjective scores were assigned by five radiologists. IQM performance was measured via the Spearman rank order correlation coefficient (SROCC) and statistically significant differences in the residuals of the IQM scores and radiologists' scores were tested. When considering SROCC calculated from combining scores from all radiologists across all image types, RMSE and SSIM had lower SROCC than six of the other IQMs included in the study (VIF, FSIM, NQM, GMSD, IWSSIM, and HDRVDP). In no case did SSIM have a higher SROCC or significantly smaller residuals than RMSE. These results should be considered when choosing an IQM in future imaging studies.**

***Index Terms*— Evaluation and performance, image quality assessment, magnetic resonance imaging (MRI), image quality metric.**

I. Introduction

THE quality of a magnetic resonance (MR) image can be difficult to assess in a robust, objective and quantitative manner. This leads to challenges in the comparison of different image acquisition and reconstruction techniques and the validation of new ones. In practice, clinical MR images are typically viewed by an expert radiologist, so the radiologist's opinion of the diagnostic quality of the image can then be considered an appropriate measure of MR image quality. However, applying this standard on a large scale is often infeasible due to large image library sizes, limited time availability of expert radiologists, and the inherent variability in a subjective scoring technique. Objective image quality metrics (IQMs) provide an alternative to manual subjective scoring by allowing image quality to be calculated by a computer. However, the relationship between radiologists' opinion of medical image quality and IQM scores is not well explored.

Objective IQMs offer several advantages over subjective scoring by a radiologist, including the ability to be implemented in an automated pipeline or as a measure within the cost-function of a reconstruction algorithm. They also provide a consistent measure of image quality without issues of inter/intra-rater consistency and bias. Objective IQMs can be broken into three categories: no-reference, reduced reference, and full-reference IQMs [1]. No-reference IQMs such as signal to noise ratio or entropy focus criterion [2] calculate a quality score when no ground truth is known. Reduced reference IQMs make use of a partially known ground truth signal. Full-reference IQMs calculate a score for an image relative to a known ground truth reference. Common full-reference IQMs include root mean square error (RMSE), which calculates the average pixel-by-pixel difference between two images, and the structural similarity index (SSIM) [1], a more complex metric that was developed to quantify loss of structure in a degraded image compared to a reference.

The RMSE and, more recently, SSIM are two of the most popular full-reference IQMs used in the MRI literature

Manuscript received May 27, 2019; revised July 15, 2019; accepted July 16, 2019. Date of publication September 16, 2019; date of current version April 1, 2020. This work was supported in part by NSERC's Canada Graduate Scholarship and Discovery Programs, in part by Brain Canada's Platform Support Program, and in part by ACOA's Atlantic Innovation Fund. *(Corresponding author: Steven Beyea.)*

A. Mason, J. Rioux, and S. Beyea are with the Department of Physics and Atmospheric Science, Dalhousie University, Halifax, NS B3H 4R2, Canada (e-mail: allistermason@dal.ca).

S. E. Clarke, A. Costa, M. Schmidt, and T. Huynh are with the Department of Diagnostic Radiology, Dalhousie University, Halifax, NS B3H 4R2, Canada.

V. Keough was with the Department of Diagnostic Radiology, Dalhousie University, Halifax, NS B3H 4R2, Canada. She is now with the Department of Medical Imaging, Toronto General Hospital, Toronto, ON M5G 2N2, Canada.

Digital Object Identifier 10.1109/TMI.2019.2930338

for validating new image acquisition [3] and reconstruction techniques [4], including machine learning algorithms [4]–[9]. This may in part be driven by the fact that they are widely available and implemented within existing environments such as MATLAB (MathWorks, MA, USA). Many of these studies incorporate a similar workflow: acquire fully sampled data, retrospectively degrade the data, typically through undersampling, and apply the new reconstruction technique. Since a ground truth image is known from the fully sampled data, full-reference IQMs are the most logical choice for quality assessment in these studies. These IQMs can also be used when retrospectively reconstructing golden-angle sampled dynamic data, as in the GRASP technique [10], where a high quality reference image could be generated from a large subset of the collected data [11]. SSIM has also been implemented as an automated measure of MR image quality directly into new techniques. For instance, Hansen *et al.* [12] used SSIM to estimate singular-value thresholds to denoise C-13 data, and Akasaka *et al.* [13] used SSIM to guide the choice of regularization weight in compressed sensing reconstruction.

An implicit assumption in these studies is that RMSE/SSIM will correlate well with image quality in a medical setting. Taking a model observer framework [14], the quality of a medical image can be defined as how well a clinical task (e.g. diagnosis) can be performed on it [15]. In this sense, the gold standard of MR image quality would be some task-based measurement such as rating the visibility of a lesion/particular anatomical feature or measuring the diagnostic accuracy when the image is evaluated by a radiologist. However, these can be difficult to implement effectively, so it is common to use a radiologist's subjective rating of overall diagnostic quality as a surrogate. For an IQM to perform well on MR images, it should then correlate with radiologists' opinion of diagnostic image quality for a variety of image contrasts and degradations that can affect the diagnostic quality of an MR image such as noise or motion. Moreover, many other objective IQMs besides RMSE/SSIM have been developed in the image processing literature, and there may exist a more appropriate choice of IQM for assessment of MR image quality. To our knowledge, the efficacy of RMSE and SSIM has not been previously studied in this specific manner.

Some previous studies that have attempted to quantify performance of common full reference objective IQMs for MR images have used non-expert raters [16], [17]. For many non-medical IQM studies (for example, rating the quality of a television picture) the use of non-expert raters is sufficient because the quality is to be optimized for the target audience [18], which is usually a non-expert. This is not the case for medical images because medical images are designed to be viewed by expert radiologists. Through their specialized training, radiologists learn to evaluate images from a unique clinical perspective and so may have different opinions of image quality compared to non-expert raters [19]. A previous study by Renieblas *et al*. used expert raters, but only studied IQMs from the SSIM family [20]. Our work is extended to a more diverse group of IQMs.

An understanding of the performance of different objective IQMs for MR images is important because, as discussed above, these metrics are increasingly used to validate and design new imaging techniques. Therefore, it is imperative that the choice of metric reflect the goal: a high-quality diagnostic image for radiologist assessment. This study investigates correlations between RMSE, SSIM, and eight other common objective IQMs with radiologists' scoring of diagnostic MR image quality. We hypothesized that the IQMs will exhibit a broad range of correlations to radiologists' ratings of image quality. High performing IQMs should have consistently high correlations across image type, anatomical regions, and degradation types. Methods and Materials.

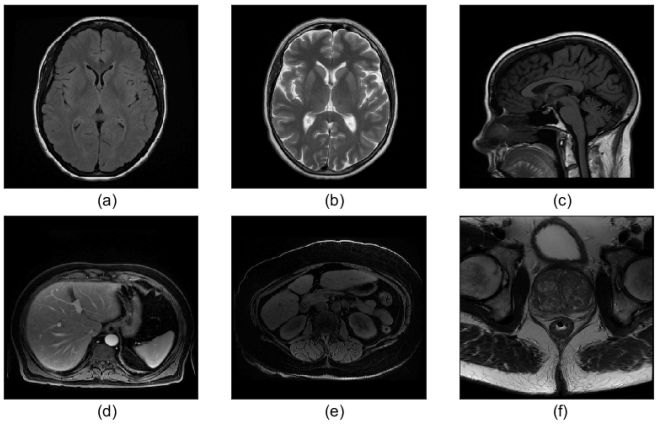

Fig. 1. Representative set of reference images. (a) T2 Flair, (b) T2 PROPELLER, (c) T1 Flair, (d) T1 LAVA-Flex (post contrast), (e) T1 Lava-Flex (pre-contrast), (f) T2 PROPELLER.

A. Generation of Image Library

Reference images were selected from our hospital's Picture Archiving and Communication System with Research Ethics Board approval. All images were anonymized before being transferred to a research server. The need for patient consent was waived. These MR images – clinically indicated and previously interpreted diagnostically as being negative for clinically relevant pathology - were chosen by a Royal College certified radiologist to have high signal to noise ratio, no visible artifacts, and no visible malignancy. The decision to use images void of pathology was made pointedly, given that intraindividual variation in pathology leads to differences in lesion conspicuity, whereas for this study the only desired differences were those introduced by the degradations. Nine reference images were selected from the abdomen (by A.C.) and from the brain (by M.S.) each. All reference images were acquired with either a GE 3T MR 750 Discovery or 1.5T Signa HDxt scanner. Of the nine abdominal images, three each were of the liver (post-contrast axial T1 LAVA-Flex), pancreas (pre-contrast axial T1 LAVA-Flex), and prostate (axial T2 PROPELLER). For the brain images, three axial T2 FLAIR, three axial T2 PROPELLER, and three sagittal T1 FLAIR images were used (Fig. 1). All reference images were of size 512×512, except for two of the axial T2 FLAIR images, which were 256×256. The reference images were originally stored as 16-bit integers with varying dynamic range. The differences in dynamic range complicated the

TABLE I
DESCRIPTION OF THE SIX DEGRADATION TYPES USED. THE STRENGTH PARAMETER CONTROLS THE DEGREE OF EACH DEGRADATION. DEGRADATION STRENGTHS FOR PARAMETERS WERE CHOSEN AT RANDOM BETWEEN THE MINIMUM/MAXIMUM VALUES

Degradation Type	Control Parameter	Minimum value	Maximum value
White Noise	Standard deviation of Gaussian	0.03	0.1
Gaussian Blur	Standard deviation of Gaussian kernel (pixels)	1	4
Motion	Percentage of frame shifted	0	10
Rician Noise	Standard deviation of Gaussian noise applied to k-space	0.02	0.05
Undersampling	Undersampling factor	2	20
Wavelet Compression	Threshold level	0.01	0.25

objective image quality assessment step, so images were first converted to 32-bit floating point and normalized to unit intensity.

One of six degradation techniques was individually applied to each image: white Gaussian noise, Gaussian blurring, Rician noise, undersampling of k-space data, wavelet compression, and motion artifacts. All degradation techniques were retrospectively applied in varying strengths to each reference image, with the exception of motion, which was only applied to the brain reference images. Retrospectively degrading the images allowed for controlled and consistent degradation of the images. Further, a ground truth image is available with this approach, which is how full reference IQMs are used in practice. The degradation techniques were chosen for their commonality in MR images and use in other similar imaging studies [21]. For each degradation type, a control parameter was determined that controlled the strength of the degradation. Each degradation was applied to each reference at four different strengths. This yielded a total image library of 414 images including the reference images. Table I provides a summary of degradation methods and control parameter ranges, and a representative set of the degradations is shown in Fig. 2.

The Gaussian white noise and Rician noise images were generated by adding noise of a Gaussian distribution either directly to the image, or to the real and imaginary components of the Fourier transforms of the image, respectively. Gaussian blurred images were generated by convolving the image with a 2D Gaussian smoothing kernel of specified standard deviation, which defines the strength of the degradation. To implement motion artifacts, the 2D image was repeatedly Fourier transformed as it was translated horizontally across the frame to simulate motion during k-space acquisition, similar to Braun *et al*. [22]. After each Fourier transform, four lines of k-space were kept. Once the k-space was filled, it was Fourier transformed back to the image domain. Undersampling was introduced by retrospectively removing components from the Fourier transform of the images. While any pattern of under-sampling could have been arbitrarily used (e.g., Poisson disc, radial, etc.) in this work we used the CIRcular Cartesian UnderSampling pattern [23]. Undersampled images were reconstructed with the BART toolbox [24] wavelet regularization with a regularization weight of 0.01. Finally, wavelet compressed images were generated by applying a global threshold of a specified value to the wavelet transform of the reference image. Wavelet transforms were generated for four levels with *sym8* type wavelets.

B. Objective IQMs

Ten full-reference objective IQMs were included in this study: RMSE, peak signal to noise ratio (PSNR), SSIM, multi-scale SSIM (MSSSIM [25]), information-weighted SSIM (IWSSIM [26]), gradient magnitude similarity deviation (GMSD [27]), feature similarity index (FSIM [28]), high dynamic range visible difference predictor (HDRVDP [29]), noise quality metric (NQM [30]), and visual information fidelity (VIF [31]). Other IQMs have been proposed in the literature, but these metrics were chosen for their prevalence, performance, and ease of implementation for diagnostic images.

The RMSE is a measure of the voxel-by-voxel difference between the reference and degraded image. PSNR is a transform of the RMSE (PSNR = 20*log(max/RMSE), where max is the maximum pixel value in the image). SSIM is a measure of the similarity in luminance, contrast, and structural content of the two images. MSSSIM and IWSSIM extend SSIM by calculating it on multiple scales and incorporating a more advanced pooling strategy of the local SSIM map by considering local information content, respectively. GMSD is the standard deviation of the gradient magnitude similarity map of the two images. FSIM considers low level features such as the phase congruency as well as the gradient magnitude. HDRVDP is a human visual system-based metric that provides a score based on the probability a human would detect a difference between two images. Although originally based on a visual difference predictor, it has been used as a quality metric for CT images in the past [32]. Like SSIM, NQM accounts for differences in luminance and contrast, but considers how these effects are affected by spatial frequencies, distance, and contrast masking effects. Finally, VIF uses natural scene statistics to measure how much information is shared between the reference and degraded image. A higher IQM value corresponds to a higher quality image for all metrics except GMSD and RMSE, where 0 denotes perfect agreement. The reader is directed to the reference papers of each metric for more detailed descriptions.

C. Radiologist Image Quality Assessment

Three body radiologists and two neuro radiologists were involved in the study. The radiologists scored only images

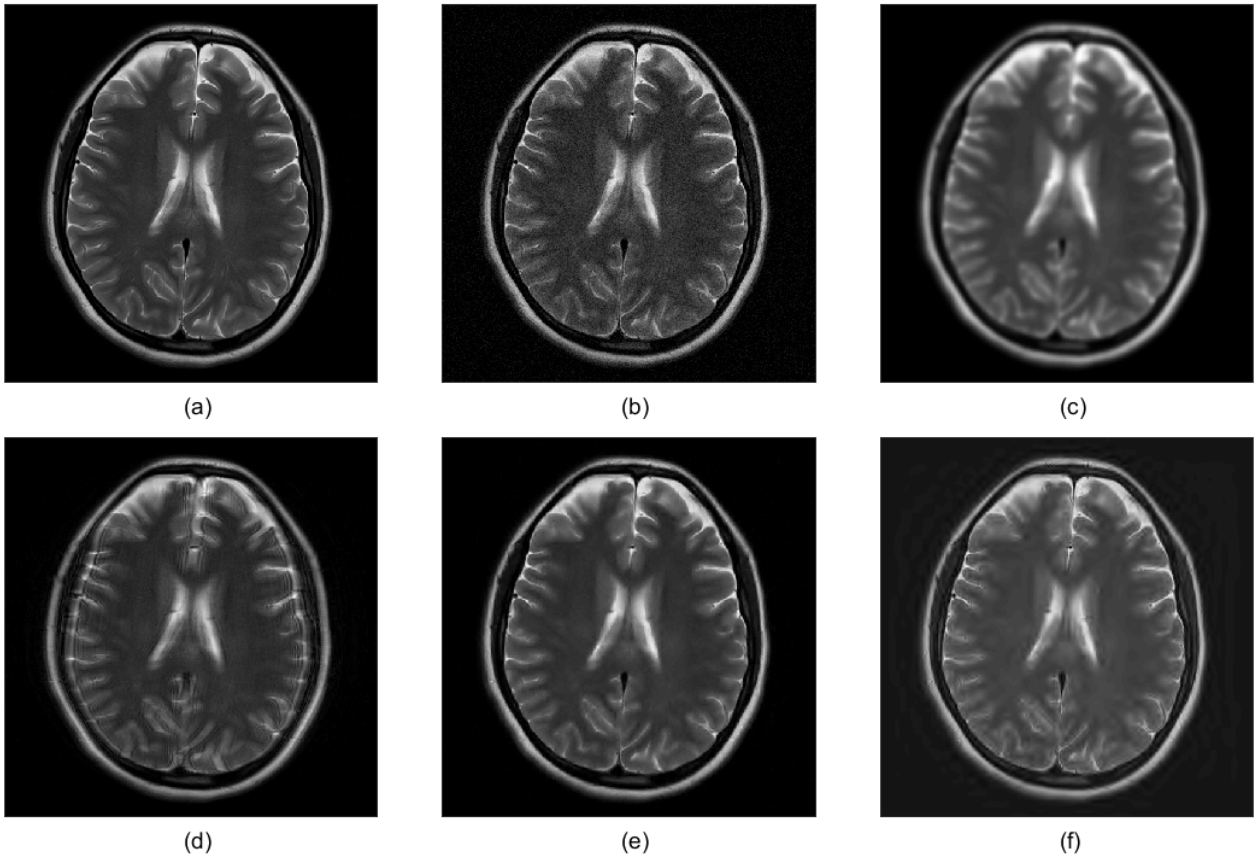

Fig. 2. Five degradations applied to a 512 × 512 T2 PROPELLER image of the brain. (a) Reference image. (b) White noise ($\sigma = 0.04$). (c) Gaussian blur ($\sigma = 3$). (d) Motion (percent shift = 6). (e) Undersampling (R = 8) and CS reconstruction. (f) Wavelet compression (threshold = 0.2). Rician noise is omitted from the figure due to its visual similarity to Gaussian noise at this degradation level.

from their subspecialties, using a 1-5 Likert scale. The radiologists were asked to rate the diagnostic quality of the images with respect to delineation of relevant anatomy and ability to confidently exclude a lesion on a 5-point scale as follows: excellent diagnostic quality (5), good diagnostic quality (4), fair diagnostic quality (3), poor diagnostic quality (2), and non-diagnostic (1). This scale was calibrated by consensus of radiologists in each subspecialty based on a training set consisting of three reference MR images (different from those in the testing set) and degraded images generated by applying each degradation technique to each reference at two representative strengths. All judgments of quality were made in their opinion as diagnostic radiologists (e.g. their ability to discriminate relevant tissues, their confidence in using the image to detect, or in this case not detect, pathology, etc.). All subsequent scoring was performed individually.

D. Data Analysis

The diagnostic image quality scores for the radiologists were not evaluated in their raw form [33]. To account for potential differences in quality of each reference image, raw scores were converted first to a difference score:

$$D_{mn} = s_{m,ref} - s_{mn}, \tag{1}$$

where D_{mn} is the difference score for the m^{th} radiologist on the n^{th} degraded image, $s_{m,ref}$ is the raw score of the m^{th} radiologist for the reference image corresponding to the n^{th} degraded image, and s_{mn} is the raw score of the m^{th} radiologist on the n^{th} degraded image. These scores were then converted to a z-score to account for differences in mean and standard deviation for each radiologist:

$$z_{mn} = (D_{mn} - \mu_m)/\sigma_m, \tag{2}$$

where μ_m and σ_m are the mean and standard deviation of the difference scores of the m^{th} radiologist. This converts all the scores to a zero mean, unit standard deviation distribution. The z-scores from each radiologist were then averaged and rescaled from 0-100.

The Spearman rank order correlation coefficient (SROCC) was calculated between the transformed radiologist scores and each of the IQM scores. The SROCC is equivalent to a linear correlation coefficient on the rank order of the data. A higher SROCC would then correspond to a better performing IQM. This metric was used because of the nonlinear relationship between subjective scores and objective IQM scores [33] (visible in Fig. 3).

Correlations were calculated under three scenarios: when scores were divided by individual radiologists, by image type, and by degradation type. For the first division, SROCCs were calculated between each individual radiologist's scores and the IQM scores of the images they scored. This division also includes the "combined" group, which combines all

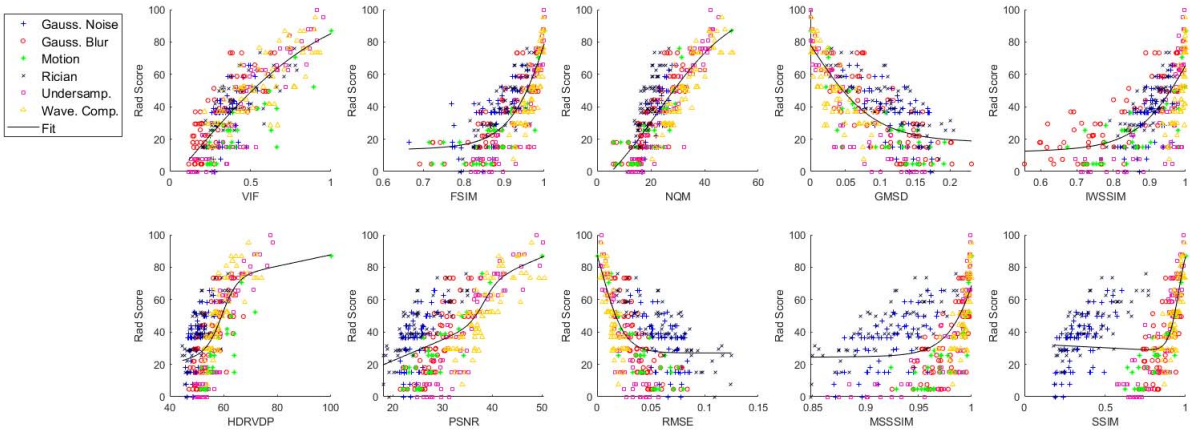

Fig. 3. Relationship between subjective radiologist score and IQMs for the full image library (414 images). Images are sorted by degradation type. The fit is calculated on all images using the non-linear regression in (3).

TABLE II

INFORMATION ABOUT EACH GROUP OF IMAGES IN THE TESTING SET FOR EACH OF THE THREE SUBDIVISIONS USED FOR ANALYSIS. N IS THE NUMBER OF IMAGES IN EACH GROUP. FOR GROUPS IN WHICH SOME IMAGES RECEIVED DIFFERENT NUMBERS OF RATINGS, THE NUMBER IN PARENTHESIS IN THE RATINGS PER IMAGE COLUMN IS THE NUMBER OF IMAGES RATED BY THE ACCOMPANYING NUMBER OF RADIOLOGISTS, I.E. 36 IMAGES DEGRADED BY WHITE NOISE WERE RATED BY 3 RADIOLOGISTS AND THE OTHER 36 WERE ONLY RATED BY 2 RADIOLOGISTS

By radiologist			By degradation type			By image type		
Image group	N	Ratings per image	Image group	N	Ratings per image	Image group	N	Ratings per image
Combined	414	3(189) or 2(225)	White Noise	72	3(36) or 2(36)	T1 Flair	75	2
Body Rad 1	189	1	Gaussian Blur	72	3(36) or 2(36)	T1 GRE (post)	63	3
Body Rad 2	189	1	Motion	36	2	T1 GRE (pre)	63	3
Body Rad 3	189	1	Rician Noise	72	3(36) or 2(36)	T2 Flair	75	2
Neuro Rad 1	225	1	Undersample	72	3(36) or 2(36)	T2 PROP (brain)	75	2
Neuro Rad 2	225	1	Wavelet	72	3(36) or 2(36)	T2 PROP (body)	63	3

radiologists' scores for all images in the study. The image type division presents SROCCs for each group of reference image. The radiologists' scores for images in each group are averaged and the SROCC is calculated with the corresponding IQM scores. Finally, images are grouped by degradation type where all images degraded by a particular technique are grouped and the SROCC is calculated. The sizes of each group are described in Table II.

A variance-based hypothesis test was performed to measure statistical significance in the difference in the performance of the IQMs. First, a non-linear regression was performed on the IQM scores according to the equation [33]:

$$Q_p = \beta_1 \left(\frac{1}{2} - \frac{1}{1 + \exp\left(\beta_2 \left(Q - \beta_3\right)\right)} \right) + \beta_4 Q + \beta_5, \quad (3)$$

where Q are the original IQM scores and β are the model parameters. The residuals between the IQM scores after the regression and the radiologists scores were calculated and Gaussianity was confirmed by measuring the kurtosis of the residuals. Residuals with a kurtosis between 2 and 4 was taken to be Gaussian (97% were found to be Gaussian). To test for statistical differences in the variance of residuals an F-test of equality of variances was performed, with the null hypothesis being that the residuals of two IQMs come from distributions of equal variance (with 95% confidence). Since each IQM was compared to all nine other IQMs, a Benjamini-Hochberg correction for false discovery rate controlling was performed [34]. An IQM performed statistically better than another IQM if the variance of its residuals is statistically less than the variance of the residuals of the other IQMs.

E. IQM Calculation Times

The time required by various IQM algorithms to calculate a quality score is also of interest to researchers looking to adopt these metrics. To measure this, we repeatedly calculated (50 times) all IQMs for all body images (512×512, N = 189). Timing calculations were performed in MATLAB 2017b running on a 24 CPU Linux research server (Intel X5650, 2.67GHz).

II. RESULTS

Fig. 3 shows the radiologists' scores versus each IQM for the "combined" subgroup, which is the combination of all images and all radiologist scores. The data is sub-divided by degradation type. The IQMs are ordered by decreasing average SROCC when the data is broken up by each radiologist. This order is kept throughout all results for consistency. As shown in Fig. 3, all IQM scores demonstrated a trend to improve (decrease for RMSE and GMSD, increase for all others) as radiologists' image quality scores increase. However, a varying strength of this trend is seen among IQMs. The sensitivities

TABLE III
WEIGHTED COHEN'S KAPPA FOR SCORES BETWEEN RADIOLOGISTS RATING THE SAME SET OF IMAGES

	Body Rad 1.	Body Rad. 2	Body Rad. 3	Neuro Rad. 1	Neuro Rad. 2
Body Rad. 1	1.00	0.850	0.655	--	--
Body Rad. 2	0.850	1.00	0.780	--	--
Body Rad. 3	0.655	0.780	1.00	--	--
Neuro Rad. 1	--	--	--	1.00	0.615
Neuro Rad. 1	--	--	--	0.615	1.00

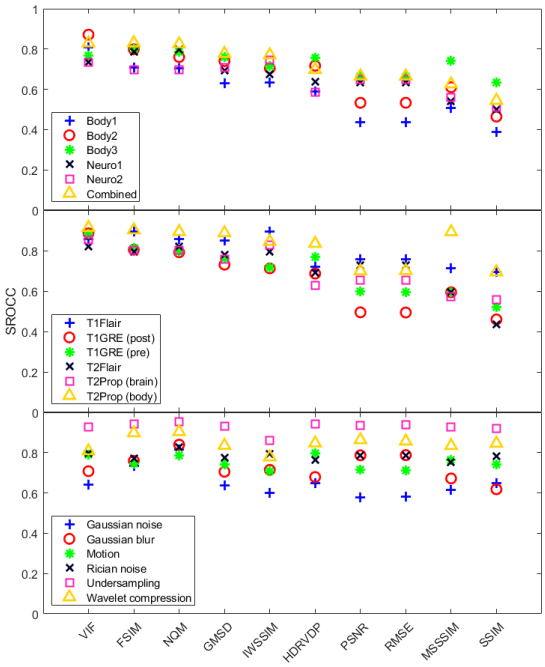

Fig. 4. Spearman rank order correlation coefficient for each IQM when data is divided by radiologist (top), reference type (middle), or degradation type (bottom).

of each of the IQMs to different degradation types can also be clearly seen. The weighted kappa [35] between radiologists who rated the same set of images is presented in Table III. Moderate to substantial agreement was found [36].

The SROCC of each IQM with each radiologist's scores are shown in Fig. 4. This ordering of the IQMs show the decreasing correlations of the IQMs with the radiologist scores in each subgroup. Note that RMSE and SSIM are among the metrics with the lowest SROCC. Overall, VIF had the highest SROCC values. Results of the hypothesis testing on the variance in the residuals for each IQM in for each radiologist are shown in Table IV. The sorted IQMs show that for many of the radiologists in the study, VIF, FSIM, and NQM perform statistically better than the other IQMs included in the study. SSIM did not perform statistically better than any another IQM including RMSE. When dividing data by reference image type, similar results were obtained (Fig. 4). The same metrics demonstrated higher correlation with radiologists' scores. The statistical differences in performance based on the variance in the residuals of IQMs also demonstrated a similar behaviour (Table V). The fewer significant differences is likely due to fewer images in each subgroup. When dividing the data by degradation type, the variation in SROCC between IQMs is less clear (Fig. 4). This lack of variation is shown by the small number of statistically different performance of the IQMs (Table VI). NQM appears to perform particularly well for images degraded by noise as it has a statistically better performance than all other IQMs except FSIM for these images. IWSSIM performed poorly for images degraded by undersampling artifacts, showing statistically larger residuals compared to all other IQMs for these images.

The calculation times of the IQMs are shown in Table VII. The simple and rapid algorithms of RMSE, PSNR, and GMSD demonstrate short calculation times (all less than 2 seconds). SSIM and MSSSIM have slightly longer calculation times (less than 20 seconds) and FSIM, NQM, and IWSSIM are longer still (less than 45 seconds). The VIF (241 ± 6seconds) and HDRVDP (403 ± 5seconds) IQMs are shown to have the longest calculation times of the metrics in this study.

III. DISCUSSION

The results of this study have important implications for researchers who are developing MR image acquisition and reconstruction techniques and using objective IQMs to test, validate and/or optimize their techniques. Recently, SSIM has gained popularity as a surrogate for RMSE as the IQM of choice, with the underlying assumption that it provides a more accurate measure of image quality. However, the results of this study demonstrate that, in the retrospectively degraded images used, SSIM does not show a significantly stronger correlation with radiologist opinion of diagnostic image quality than RMSE, and that there are other objective IQMs that perform better. This does not imply that previous studies that use RMSE or SSIM are invalid, since RMSE and SSIM were still seen to correlate with radiologists' scores, but that there exist other metrics that may provide a more accurate measure of diagnostic image quality.

When considering the trends in Fig. 3 and the bottom of Fig. 4, it appears as if the factor that most affects IQM performance is how uniformly the IQM quantifies the quality of images with different degradations. For instance, in Fig. 3, VIF shows substantial overlap of all degradation techniques, but as one progresses through the metrics, the distributions of degradations in the Rad Score-IQM plane become much more distinguishable. One can clearly discern the distributions of different degradation techniques for IQMs such as PSNR or RMSE. In the extreme case of SSIM, a bimodal distribution appears between the noise and other degradations. As seen in Fig. 4, when the images are divided by degradation type, each IQM has a similar SROCC with the radiologists

TABLE IV

STATISTICAL SIGNIFICANCE IN RESIDUALS OF IQM SCORES AFTER REGRESSION AND RADIOLOGISTS SCORES (SIGNIFICANCE LEVEL = 0.05 WITH BENJAMINI-HOCHBERG CORRECTION FOR MULTIPLE COMPARISONS) WHEN DATA IS BROKEN UP BY RADIOLOGIST. A '1' MEANS THE IQM PERFORMED STATISTICALLY BETTER THAN THE IQM OF THE COLUMN. A '0' MEANS IT WAS STATISTICALLY WORSE. A '-' MEANS NO SIGNIFICANT DIFFERENCE. THE ORDER OF THE SUB-ELEMENTS IS: COMBINED, BODY RADIOLOGIST 1, BODY RADIOLOGIST 2, BODY RADIOLOGIST 3, NEURORADIOLOGIST 1, AND NEURORADIOLOGIST 2

	VIF	FSIM	NQM	GMSD	IWSSIM	HDRVDP	PSNR	RMSE	MSSSIM	SSIM
VIF	------	-11---	--1---	111---	111---	111---	111---	111---	111-11	111---
FSIM	--0---	------	------	1-----	1-1-1-	1-----	1-11--	1111--	111-1-	1-11--
NQM	--0---	------	------	1---1-	1-1-1-	1---1-	1-111-	11111-	111-1-	1-111-
GMSD	000---	------	0---0-	------	------	------	------	--1---	1-1--1	1-----
IWSSIM	000---	0-0-0-	0-0-0-	------	------	------	------	------	------	------
HDRVDP	0-0---	0-----	0---0-	------	------	------	--1---	--1---	1-1---	--1---
PSNR	000---	0-0---	0-000-	--0---	------	--0---	------	------	1----1	------
RMSE	000---	0000--	00000-	--0---	------	--0---	------	------	1-----	------
MSSSIM	000-00	000-00	000-00	0-0--0	0----0	000---	0----0	0----0	------	------
SSIM	000---	0-00--	0-000-	------	------	--0---	------	------	------	------

TABLE V

STATISTICAL SIGNIFICANCE IN RESIDUALS OF IQM SCORES AFTER REGRESSION AND RADIOLOGISTS SCORES (SIGNIFICANCE LEVEL = 0.05 WITH BENJAMINI-HOCHBERG CORRECTION FOR MULTIPLE COMPARISONS) WHEN DATA IS BROKEN UP BY REFERENCE IMAGE TYPE. A DESCRIPTION OF THE TABLE FORMAT IS PROVIDED IN TABLE IV. THE ORDER OF THE SUB-ELEMENTS IS: T1 FLAIR (BRAIN), T1 LAVA-FLEX (POST-CONTRAST, LIVER), T1 LAVA-FLEX (PRE-CONTRAST, PANCREAS), T2 FLAIR (BRAIN), T2 PROPELLER (BRAIN), T2 PROPELLER (PROSTATE)

	VIF	FSIM	NQM	GMSD	IWSSIM	HDRVDP	PSNR	RMSE	MSSSIM	SSIM
VIF	------	------	-1----	-1----	-11--1	-1--11	-11--1	-11--1	11111-	-11-11
FSIM	------	------	------	------	------	------	-----1	-----1	1-111-	---1-1
NQM	------	------	------	------	--1---	------	--1--1	--1--1	--111-	--11-1
GMSD	------	------	------	------	------	------	------	------	----1-	------
IWSSIM	-0---0	------	------	------	------	------	------	------	------	------
HDRVDP	----0-	------	------	------	------	------	------	------	------	------
PSNR	-00--0	-----0	--0--0	-----0	------	------	------	------	----1-	------
RMSE	-00--0	-----0	--0--0	-----0	------	------	------	------	------	------
MSSSIM	00000-	0-000-	0-000-	0---0-	0---0-	--0---	----0-	----0-	------	----0-
SSIM	-000-0	---0-0	--00-0	------	------	------	------	------	------	------

score. It is only when the different degradations are grouped together that the differences in performance of IQMs arise. This is important to notice because, as discussed in the Introduction, an IQM should correlate with radiologists' opinion over a range of degradations. This also shows how choice of degradation types and strengths can affect the results of studies of this nature.

After normalizing each radiologist's score and combining scores across all images, we found that VIF exhibited the highest SROCC of all the metrics evaluated in this study. These results suggest that VIF provides the most accurate surrogate measure of subjective image quality scores of a radiologist of the IQMs included in this study. VIF is unique among IQMs in this study in that it generates a quality score based on shared information between the reference image and the distorted image, instead of generating a score from an algorithm based on some definition of signal fidelity. In VIF, the information in the reference image is calculated from natural scene statistics. The distorted image is modelled as the reference image passed through a distortion channel. The VIF is found from the information remaining in the degraded image from the reference image. It should be noted that VIF is designed with natural scene statistics, not medical image statistics, indicating an area of potential future research.

FSIM and NQM also consistently demonstrated high correlations with the radiologist scores. Indeed, NQM had performed statistically better than VIF for images degraded with Gaussian noise, undersampling, or wavelet compression. This is consistent with other similar studies of MR images

TABLE VI

STATISTICAL SIGNIFICANCE IN RESIDUALS OF IQM SCORES AFTER REGRESSION AND RADIOLOGISTS SCORES (SIGNIFICANCE LEVEL = 0.05 WITH BENJAMINI-HOCHBERG CORRECTION FOR MULTIPLE COMPARISONS) WHEN DATA IS BROKEN UP BY DEGRADATION TYPE. A DESCRIPTION OF THE TABLE FORMAT IS PROVIDED IN TABLE IV. THE ORDER OF THE SUB-ELEMENTS IS: GAUSSIAN NOISE, GAUSSIAN BLUR, MOTION, RICIAN NOISE, UNDERSAMPLING, AND WAVELET COMPRESSION

	VIF	FSIM	NQM	GMSD	IWSSIM	HDRVDP	PSNR	RMSE	MSSSIM	SSIM
VIF	------	------	------	------	------	------	------	------	------	------
FSIM	------	------	------	------	----11	------	------	------	-----1	------
NQM	1---11	------	------	1---1-	1---11	1-----	1-----	1-----	1---11	1-----
GMSD	------	------	------	------	------	------	------	------	------	------
IWSSIM	----0-	----00	----00	----0-	------	----0-	----0-	----0-	------	----0-
HDRVDP	------	------	------	------	----1-	------	------	------	------	------
PSNR	------	------	------	------	----1-	------	------	------	------	------
RMSE	------	------	0-----	------	----1-	------	------	------	------	------
MSSSIM	------	-----0	----00	------	------	------	------	------	------	------
SSIM	------	------	------	------	------	------	------	------	------	------

TABLE VII

AVERAGE FOR CALCULATION TIMES IN SECONDS OF EACH METRIC FOR ALL BODY IMAGES (RESOLUTION: 512 × 512, N = 189)

	VIF	FSIM	NQM	GMSD	IWSSIM	HDRVDP	PSNR	RMSE	MSSSIM	SSIM
mean	241.38	31.11	39.21	1.77	42.01	403.37	0.32	0.23	14.98	6.07
Standard error	6.52	0.16	0.91	0.03	0.36	5.39	0	0	0.41	0.03

[16], [17]; however, a key distinction between these studies and ours is that we correlated with the image quality scores of expert radiologists instead of non-experts. For Renieblas *et al.* [20], who did use expert raters, moderate agreement in SROCC was found from IQMs used in both our study and theirs (SSIM: 0.54 versus 0.44; MSSSIM: 0.63 versus 0.60). Differences are likely due to differences in degradation techniques/strengths, as well as variability in the subjective scoring by experts.

For all IQMs, there is significant variation in the radiologists scores for a particular value of the IQM (Fig. 3). For example, in images that had an averaged scaled radiologist score of 50, the VIF ranged from 0.30-0.68. Similar trends can be seen in the results of Chow *et al.* [17]. This highlights the difficulties in predicting subjective scoring with objective IQMs, since there will always be variability in the subjective scoring and the IQMs and radiologists may have different sensitivities or preferences for different forms of degradation. As machine learning algorithms advance, it is possible they may be able to learn this sensitivity and preferences in ways objective IQMs cannot.

There are some aspects of our study which may limit the generalizability of our results. First, we limited the scope of this study to ten objective IQMs. Since many IQMs exist, including all of them is not feasible. It is possible that a metric not included in this study could demonstrate stronger correlation with radiologists' opinion of image quality than any of the IQMs we evaluated. We also limited our choice of IQM to full-reference IQMs because these are most commonly used for the validation of imaging techniques. Full-reference IQMs allowed us to use retrospectively degraded images, which adds more control to the study, but may add artificiality to the degraded images and limit the generalizability of the results in practice. A similar study with no-reference IQMs may also be considered, particularly for techniques that wish to assess the diagnostic quality of MR images on the scanner as part of a built-in quality assurance system. The present study only included data from brain images and body images, but presented both independently and together, which allows for the interpretation of the data for specific applications. However, our results may not be generalizable to other MRI systems, anatomical regions, or even different MRI sequences. Finally, it should be noted that the scoring of diagnostic quality in a clinically normal MR image is strongly related to but not necessarily equivalent to scoring of an image containing pathology. Our current work focused on clinically normal images, a critical first step in determining if IQMs developed for non-experts rating natural images would also correlate with radiologists rating diagnostic quality of MR images .Future studies will examine whether these same IQMs correlate with other task based measures such as lesion conspicuity scores in images containing pathology or diagnostic accuracy.

IV. CONCLUSION

We measured the correlations between 10 full-reference objective IQMs and the subjective image quality score of five subspecialty radiologists. When considering images divided by reference location or combining all images in the study, SSIM

and RMSE demonstrated statistically worse performances than other metrics evaluated, suggesting that SSIM and RMSE are potentially not ideal surrogate measures of MR image quality as determined by radiologist evaluation. The VIF, FSIM, and NQM demonstrated the highest correlation with radiologists' opinions of MR image quality. However, these metrics come at the cost of longer calculation times, which may influence their use in future research. Differences in the performance of the IQMs was also largely lost when images are divided by degradation type. Both the IQM SROCC values and calculation times presented in this study should be considered in future imaging studies applying an objective IQM to assess the quality of an MR image, for example in studies evaluating novel image reconstruction algorithms. These data also highlight the importance of the ongoing development of techniques for automatic and objective assessment of image quality.

REFERENCES

[1] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[2] D. Atkinson, D. L. G. Hill, P. N. R. Stoyle, P. E. Summers, and S. F. Keevil, "Automatic correction of motion artifacts in magnetic resonance images using an entropy focus criterion," *IEEE Trans. Med. Imag.*, vol. 16, no. 6, pp. 903–910, Dec. 1997.

[3] S. Kojima, S. Shinohara, H. Hashimoto, and T. Suzuki, "Undersampling patterns in k-space for compressed sensing MRI using two-dimensional Cartesian sampling," *Radiological Phys. Technol.*, vol. 11, no. 3, pp. 303–319, Sep. 2018.

[4] H. Zheng *et al.*, "Multi-contrast brain MRI image super-resolution with gradient-guided edge enhancement," *IEEE Access*, vol. 6, pp. 57856–57867, 2018.

[5] K. Hammernik *et al.*, "Learning a variational network for reconstruction of accelerated MRI data," *Magn. Reson. Med.*, vol. 79, no. 6, pp. 3055–3071, 2017.

[6] B. Gözcü *et al.*, "Learning-based compressive MRI," *IEEE Trans. Med. Imag.*, vol. 37, no. 6, pp. 1394–1406, Jun. 2018.

[7] B. A. Duffy, "Retrospective correction of motion artifact affected structural MRI images using deep learning of simulated motion," in *Proc. 1st Conf. Med. Imag. Deep Learn.*, 2018, pp. 1–8.

[8] H. Jeelani, J. Martin, F. Vasquez, M. Salerno, and D. S. Weller, "Image quality affects deep learning reconstruction of MRI," in *Proc. IEEE 15th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2018, pp. 357–360.

[9] A. S. Chaudhari *et al.*, "Super-resolution musculoskeletal MRI using deep learning," *Magn. Reson. Med.*, vol. 80, no. 5, pp. 2139–2154, Nov. 2018.

[10] L. Feng *et al.*, "Golden-angle radial sparse parallel MRI: Combination of compressed sensing, parallel imaging, and golden-angle radial sampling for fast and flexible dynamic volumetric MRI," *Magn. Reson. Med.*, vol. 72, no. 3, pp. 707–717, Sep. 2014.

[11] A. Mason, N. Murtha, J. Rioux, S. Clarke, C. Bowen, and S. Beyea, "Pharmocokinetic parameter accuracy correlates with image quality metrics in flexible temporal resolution DCE-MRI," *Proc. Int. Soc. Mag. Reson. Med.*, vol. 27, p. 5035, May 2019.

[12] R. Hansen *et al.*, "Multichannel hyperpolarized 13C MRI in a patient with liver metastases using multi- slice EPI and an alternating projection method for denoising," *Proc. Int. Soc. Mag. Reson. Med*, vol. 26, p. 3560, Jun. 2018.

[13] T. Akasaka *et al.*, "Optimization of regularization parameters in compressed sensing of magnetic resonance angiography: Can statistical image metrics mimic radiologists' perception?" *PLoS ONE*, vol. 11, no. 1, 2016, Art. no. e0146548.

[14] H. H. Barrett, J. Yao, J. P. Rolland, and K. J. Myers, "Model observers for assessment of image quality," *Proc. Nat. Acad. Sci. USA*, vol. 90, pp. 9758–9765, Nov. 1993.

[15] X. He and S. Park, "Model observers in medical imaging research," *Theranostics*, vol. 3, no. 10, pp. 774–786, 2013.

[16] H. Rajagopal, L. S. Chow, and R. Paramesran, "Subjective versus objective assessment for magnetic resonance images," in *Proc. 17th Int. Conf. Commun. Inf. Technol. Eng.*, Oct. 2015, vol. 9, no. 12, pp. 2426–2431.

[17] L. S. Chow, H. Rajagopal, and R. Paramesran, "Correlation between subjective and objective assessment of magnetic resonance (MR) images," *Magn. Reson. Imag.*, vol. 34, no. 6, pp. 820–831, 2016.

[18] *Methodology for the Subjective Assessment of the Quality of Television Pictures*, document BT.500, 2009.

[19] A. Keshavan, J. Yeatman, and A. Rokem, "Combining citizen science and deep learning to amplify expertise in neuroimaging," *Frontiers Neuroinform.*, vol. 13, p. 29, May 2019. doi: 10.3389/fninf.2019.00029.

[20] G. P. Renieblas, A. T. Nogués, A. M. González, N. Gómez-Leon, and E. G. Del Castillo, "Structural similarity index family for image quality assessment in radiological images," *J. Med. Imag.*, vol. 4, no. 3, 2017, Art. no. 035501.

[21] L. S. Chow and R. Paramesran, "Review of medical image quality assessment," *Biomed. Signal Process. Control*, vol. 27, no. 1, pp. 145–154, 2016.

[22] S. Braun, C. Xiao, B. Odry, B. Mailhe, and M. Nadar, "Motion detection and quality assessment of mr images with deep convolutional densenets," *Proc. Int. Soc. Mag. Reson. Med.*, vol. 26, p. 2715, Jun. 2018.

[23] J. Liu and D. Saloner, "Accelerated MRI with circular Cartesian undersampling (CIRCUS): A variable density Cartesian sampling strategy for compressed sensing and parallel imaging," *Quant. Imag. Med. Surg.*, vol. 4, no. 1, pp. 57–67, Feb. 2014.

[24] M. Uecker *et al.*, "Berkeley advanced reconstruction toolbox," *Proc. Int. Soc. Mag. Reson. Med.*, vol. 23, p. 2486, Jun. 2015.

[25] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *Proc. 37th Asilomar Conf. Signals, Syst. Comput.*, vol. 2, Nov. 2003, pp. 9–13.

[26] Z. Wang and Q. Li, "Information content weighting for perceptual image quality assessment," *IEEE Trans. Image Process.*, vol. 20, no. 5, pp. 1185–1198, May 2011.

[27] W. Xue, L. Zhang, X. Mou, and A. C. Bovik, "Gradient magnitude similarity deviation: A highly efficient perceptual image quality index," *IEEE Trans. Image Process.*, vol. 23, no. 2, pp. 668–695, Feb. 2014.

[28] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "FSIM: A feature similarity index for image quality assessment," *IEEE Trans. Image Process.*, vol. 20, no. 8, pp. 2378–2386, Aug. 2011.

[29] R. Mantiuk, K. J. Kim, A. G. Rempel, and W. Heidrich, "HDR-VDP-2: A calibrated visual metric for visibility and quality predictions in all luminance conditions," *ACM Trans. Graph.*, vol. 30, no. 4, Jul. 2011, Art. no. 40.

[30] N. Damera-Venkata, T. D. Kite, W. S. Geisler, B. L. Evans, and A. C. Bovik, "Image quality assessment based on a degradation model," *IEEE Trans. Image Process.*, vol. 9, no. 4, pp. 636–650, Apr. 2000.

[31] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Trans. Image Process.*, vol. 15, no. 2, pp. 430–444, Feb. 2006.

[32] B. Kim *et al.*, "Prediction of perceptible artifacts in JPEG2000 compressed abdomen CT images using a perceptual image quality metric," *Acad. Radiol.*, vol. 15, no. 3, pp. 314–325, Mar. 2008.

[33] H. R. Sheikh, M. F. Sabir, and A. C. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE Trans. Image Process.*, vol. 15, no. 11, pp. 3441–3452, Nov. 2006.

[34] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: A practical and powerful approach to multiple testing," *J. Roy. Stat. Soc. Ser. B (Methodol.)*, vol. 57, no. 1, pp. 289–300, 2019.

[35] J. Cohen, "Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit," *Psych. Bull.*, vol. 70, no. 4, pp. 213–220, Oct. 1968.

[36] A. J. Viera and J. M. Garrett, "Understanding interobserver agreement: The kappa statistic," *Fam. Med.*, vol. 37, no. 5, pp. 360–363, May 2005.