The British Journal of Radiology, 83 (2010), 604–611

A method to analyse observer disagreement in visual grading studies: example of assessed image quality in paediatric cerebral multidetector CT images

[1]K LEDENIUS, MSc, [2]E SVENSSON, PhD, [3]F STÅLHAMMAR, MD, [3]L-M WIKLUND, MD, PhD and [1,4]A THILANDER-KLANG, PhD

[1]*Department of Radiation Physics, Sahlgrenska Academy at Göteborg University, Sahlgrenska University Hospital, SE-413 45 Göteborg, Sweden,* [2]*Department of Statistics, Örebro University, SE-701 82, Örebro, Sweden,* [3]*Department of Paediatric Radiology and Physiology, The Queen Silvia Children's Hospital, SE-416 85 Göteborg, Sweden, and* [4]*Department of Medical Physics and Biomedical Engineering, Sahlgrenska University Hospital, SE-413 45 Göteborg, Sweden*

ABSTRACT. The purpose was to demonstrate a non-parametric statistical method that can identify and explain the components of observer disagreement in terms of systematic disagreement as well as additional individual variability, in visual grading studies. As an example, the method was applied to a study where the effect of reduced tube current on diagnostic image quality in paediatric cerebral multidetector CT (MDCT) images was investigated.

Quantum noise, representing dose reductions equivalent to steps of 20 mA, was artificially added to the raw data of 25 retrospectively selected paediatric cerebral MDCT examinations. Three radiologists, blindly and randomly, assessed the resulting images from two different levels of the brain with regard to the reproduction of high- and low-contrast structures and overall image quality. Images from three patients were assessed twice for the analysis of intra-observer disagreement.

The intra-observer disagreement in test-retest assessments could mainly be explained by a systematic change towards lower image quality the second time the image was reviewed. The inter-observer comparisons showed that the paediatric radiologist was more critical of the overall image quality, while the neuroradiologists were more critical of the reproduction of the basal ganglia. Differences between the radiologists regarding the extent to which they used the whole classification scale were also found.

The statistical method used was able to identify and separately measure a presence of bias apart from additional individual variability within and between the radiologists which is, at the time of writing, not attainable by any other statistical approach suitable for paired, ordinal data.

Received 22 January 2009
Revised 14 August 2009
Accepted 18 August 2009

DOI: 10.1259/bjr/26723788

Optimisation of radiological examinations can be performed using many different approaches. One way of optimising multidetector CT (MDCT) examinations is to adjust the tube current and study the effect of image noise on diagnostic image quality in order to find a balance between radiation dose and indication. Receiver operating characteristic (ROC) methodology may be used to determine the appropriate settings for a specific diagnosis if a sufficient number of patients are involved in the study [1]. However, the ROC method is not very practicable when optimising a general protocol intended to be used for a broad variety of diagnoses, some of which may occur only a few times per year. An alternative method is visual grading, where the visibility of organs and structures is evaluated. The benefit of visual grading is that it is not limited to a specific diagnosis. The disadvantage is that it is a subjective method of evaluating the quality of an image, and not a measure of the ability to make the correct diagnosis. However, the method is very similar to the clinical situation faced by radiologists when determining whether the image quality is sufficient with regard to the indications and possible diagnoses.

An optimised MDCT protocol must ensure an image quality that is considered adequate by all the radiologists at the radiology department. Poor image quality can result in an additional radiation dose to the patient if the examination has to be repeated. It is thus advisable not to apply a mean value of the lowest possible dose, but to respect the different requirements of all radiologists, within reasonable limits. Although this should be taken into account, too large a deviation should be investigated, and measures taken to protect the patient from excessive exposure. Identifying how and why radiologists differ in accepting a certain image quality is to quality assure the department, and the information may well serve as a basis for training. The evaluation of

Address correspondence to: K Ledenius, Department of Radiation Physics, Sahlgrenska Academy at Göteborg University, Sahlgrenska University Hospital, SE-413 45 Göteborg, Sweden. E-mail: kerstin.ledenius@vgregion.se

Table 1. Patient and protocol characteristics

	0–5 months	6–11 months	1–5 years	6–10 years	11–14 years	>14 years
No. of patients	3	1	5	8	5	3
Tube voltage (kV)	120	120	120	120	120	120
Scan FOV	Paed.head	Paed.head	Head	Head	Head	Head
Configuration (mm)	4 × 5	4 × 5	4 × 5	4 × 5	4 × 5	4 × 5
Rotation time (s)	0.8	0.8	1	1	1	1
Recon. algorithm	Soft	Soft	Soft	Soft	Soft	Soft
Upper level						
Tube current (mA)	110	130	180	200	230	240
Focal spot	Small	Small	Small	Small	Large	Large
$CTDI_{vol}$ (mGy)	15	17	30	33	41	43
Lower level						
Tube current (mA)	110	130	200	220	250	260
Focal spot	Small	Small	Small	Large	Large	Large
$CTDI_{vol}$ (mGy)	15	17	33	39	44	46

Age-based scanning protocols used for routine paediatric cerebral MDCT examination.
$CTDI_{vol}$, volume CT dose index; FOV, field of view.

observer disagreement in MDCT image quality assessment using a visual grading approach has been sparsely investigated; we have found only one other published study [2], thus indicating a lack of research within this subject. One reason for this could perhaps be the lack of statistical methods that are both appropriate for qualitative data and give valuable information.

The aim of this study was to demonstrate a non-parametric statistical method that can identify and explain the components of observer disagreement, in terms of systematic disagreement, as well as additional individual variability in visual grading studies. As an example, the method was applied to a study where the effect of reduced tube current on diagnostic image quality in paediatric cerebral MDCT images was investigated [3].

Methods and materials

Simulated tube current reduction in paediatric cerebral MDCT images

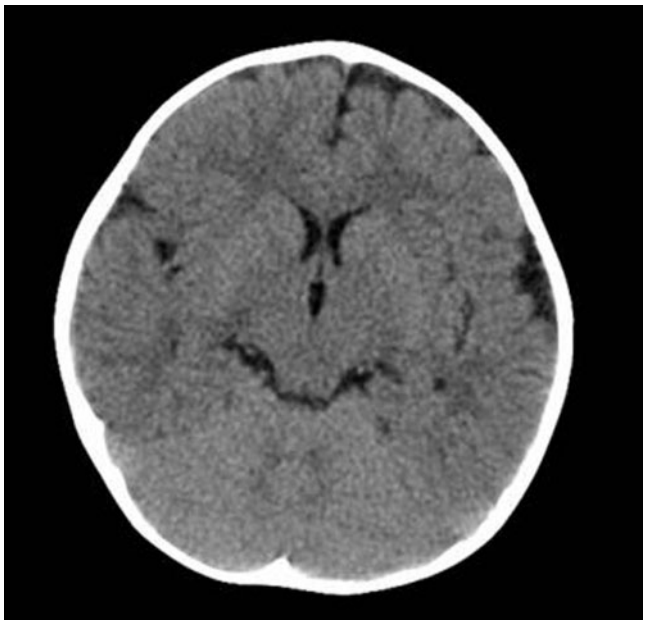
Original digital scanning data (raw data) were retrospectively selected from 25 routine paediatric cerebral MDCT examinations (14 male/11 female). Exclusion criteria were pathological findings that interfered with or overlapped structures of interest, or contrast medium enhancement, in order to avoid variation in the premises for structure visibility. All examinations had been performed using axial scanning on an eight-slice MDCT scanner (LightSpeed Ultra, GE Healthcare, Milwaukee, WI). Scanning parameters had been chosen from standard head protocols suitable for the patients' age and size (Table 1). No automatic tube current modulation had been used, since the technique had not been implemented in clinical routine at the time of the study. Patient identification information was removed from all examinations. Images representing 2 different levels in the brain (upper and lower) were selected from each of the 25 individuals (Figure 1). The two different levels were chosen for their difference in composition. Both levels represent important areas for diagnostics and contain both high- and low-contrast structures.

A noise-simulating program developed by GE Healthcare was installed on a separate research CT console. The software adds a random Gaussian noise distribution, corresponding to the size of the dose reduction, to the raw data, thus including it in the filtering and reconstruction of the image. A more thorough description of the noise simulation software and its validation can be found in Frush et al [4]. Simulations of the tube current were performed at intervals of 20 mA from the clinically used tube currents down to 30/50 mA (upper/lower levels of the brain) for patients younger than 1 year 40/60 mA for patients 1–10-years-old and older than 14 years, and 50/70 mA for patients 11–14-years-old (Table 1). According to the Institutional Review Board, this approach was not subject to ethical review or informed consent.

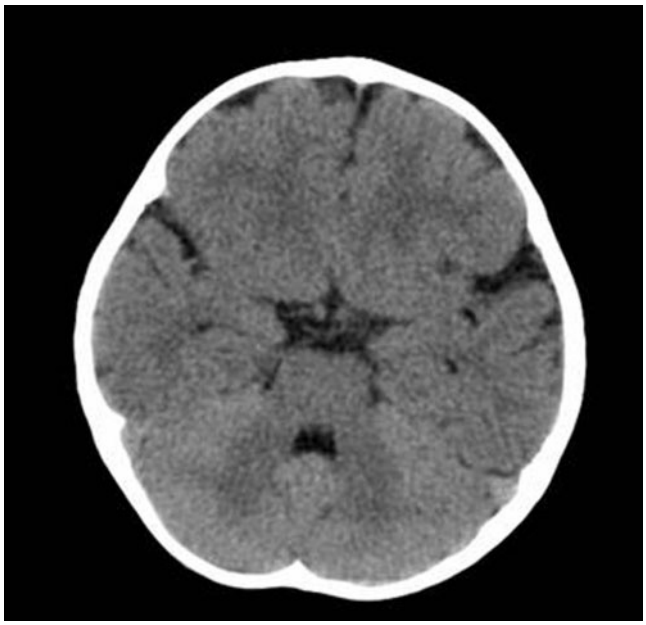
Evaluation of image quality

The image quality of 496 images was blindly evaluated, in random order, on an 8-bit greyscale, 1280 × 1024 CRT monitor with max./min. luminance level of 385/3.58 cd m^{-2} (Siemens Simomed, Munich, Germany). The monitor was calibrated according to Digital Imaging and Communications in Medicine (DICOM) part 14 [5], and the display window width and window level were set to 65 and 35 HU, respectively; however, the settings were adjustable. The 496 images consisted of simulated dose-reduced images (n=386), original dose images (n=50) and duplicated images from three patients (n=60). The radiologists were unaware of the duplicated images.

The images were assessed using verbal rating scales where questions 1 to 5 (Q1–Q5) refer to the reproduction of structures, and question 6 (Q6) refers to the overall image quality considering the indication (Table 2). The following classifications were used for the reproduction of structures: ''Clearly'', the structure had a completely distinct shape; ''Acceptably'', the structure was moderately reproduced; ''Poorly'', the structure was vaguely reproduced; and ''Not at all'', the structure could not be discerned. The choice of structures was based on the structures defined in the European Guidelines on Quality Criteria for Computed Tomography [6], which



(a)



(b)

Figure 1. Multidetector CT images representing the levels of the brain used for diagnostic image quality assessment: (a) the upper level of the brain, showing the lateral ventricles and the basal ganglia, and (b) the lower level of the brain, including the posterior fossa at the level of the fourth ventricle. The patient is female, 15 months old, and was scanned with the parameters according to Table 1.

provides guidelines for image quality criteria for adult CT examinations. Guidelines for paediatric patients were not available at the time of the study. Overall image quality was classified with regard to suspected pathology, using the following verbally defined scale categories: ''High-resolution diagnostics'' allows analysis of low-contrast targets such as cancer; ''Standard diagnostics'' allows analysis of mixed targets of varying contrast such as trauma; ''Low-resolution diagnostics'' allows analysis of high-contrast targets such as the ventricles;

Table 2. Classification scale

1	How well can you differentiate white and grey matter?
2	How well can you visualise the basal ganglia?
3	How well is the ventricular system delineated?
4	How well is the cerebrospinal fluid space around the mesencephalon delineated?
5	How well is the cerebrospinal fluid space around the brain delineated?
A	Clearly
B	Acceptably
C	Poorly
D	Not at all
E	Not applicable
6	For what diagnostic situation is this image quality sufficient?
F	For high-resolution diagnostics
G	For standard diagnostics
H	For low-resolution diagnostics
I	Not diagnostically useful

Questions 1–5 concern the reproduction of high- and low-contrast objects with possible responses A–E, and question 6 concerns overall image quality regarding indication, with possible responses F–I.

and ''Not diagnostically useful'' when the image quality is of no diagnostic value.

The images were evaluated using the computer software ViewDEX (viewer for digital evaluation of X-ray images) [7, 8]. ViewDEX is a Java program developed to present images in a random order, without patient or scanning data, with the facility of answering the related questions directly on the screen. Each radiologist had a personal login ID so that the images could be assessed over a period of several weeks. The radiologists were not allowed to discuss their findings with each other. The assessments were stored in text files that were imported into Microsoft Excel®. The effect of the dose reductions on the diagnostic image quality has been investigated in a separate article [3].

Experience of the radiologists

At the time of the study, Observer 1 had 25 years of experience as a radiologist, 20 of which as a neuroradiologist and 19 as a paediatric radiologist. Observer 1 also had considerable experience of visually grading image quality for research purposes. Observer 2 had 13 years of experience as a radiologist, 8 of which as a paediatric radiologist. Observer 2 also had experience of visual grading studies, but not to the same extent as Observer 1. Observer 3 had 33 years of experience as a radiologist, 25 of which as a neuroradiologist. Observer 3 reported having no experience of similar studies, but had experience of visually grading different concentrations of contrast media.

Statistical method

Assessing images using a verbal rating scale produces ordered categorical data, also known as ordinal data. The scale assessments indicate only an ordered structure and not a numerical value in a mathematical sense. Statistical

evaluations of ordinal data must take into account their so-called rank-invariant properties, which means that the methods must be unaffected by a relabelling of the scale categories. Hence, rank-based statistical methods must be used [9–11]. The statistical approach [12, 13] used in this study takes these properties into account and allows comprehensive analysis of paired ordinal data, which identifies and measures systematic disagreement (bias) separately from the individual variations in paired assessments. The method has been shown to be valuable in various studies of reliability [14], validity [15–18] and change [19–24]. The frequency distribution of the assessments of one radiologist on two different occasions (intra-observer comparison) and of two different radiologists (interobserver comparison) was displayed in cross-classification tables where the main diagonal, representing agreement, is orientated from the lower-left to the upper-right corner (Figure 2). The agreement was expressed as percentage agreement (PA).

The presence of systematic disagreement is indicated by different marginal frequency distributions in the two assessments. Two measures of systematic disagreement were calculated: the relative position (RP) and the relative concentration (RC) with possible values ranging from –1 to 1. In the interobserver analysis the RP expresses the difference between the proportions of overestimated and of underestimated scale assessments made by Observer X compared with the assessments made by Observer Y, and estimates the difference between corresponding probabilities: $p(X<Y)-p(Y<X)$. Figure 2a shows the paired assessments made by Observers 1 (X) and 2 (Y). The positive RP (0.33) indicates that Observer 2 systematically used higher scale categories (which in our study means lower levels of image quality) than Observer 1. RP=0.33 means that 33% more images are being classified to higher categories than to lower by Observer 2 when compared with the classifications made by Observer 1. This means that Observer 2 was more likely to assess an image as being of a poorer quality than Observer 1. In the intra-observer analysis, X denotes the first and Y the second review. The measure of systematic disagreement in concentration, *i.e.* the RC, provides an estimate of the difference between the probability that Observer X concentrates the assessments on the scale classifications more than does Observer Y, and vice versa. It can be seen in Figure 2b that Observer 1 (X) tends to have a higher proportion of assessments in the central classification levels than Observer 3 (Y), which is apparent from the different marginal distributions, hence a negative value of RC. In intra-observer disagreement, the value of RC is negative when a higher proportion of the assessments have central classifications at the first review than at the second. Values of RP and RC of zero indicate a lack of systematic disagreement in position and in concentration on the scale, respectively. The 95% confidence intervals (CI) were estimated by means of the bootstrap technique. The relative rank variance (RV) is a measure of additional variability in assessments that cannot be explained by systematic disagreement. Possible values of RV range from zero to 1, where non-zero RV indicates the presence of random disagreement, and the higher the value of RV the more heterogeneous are the paired assessments of the same image. For details regarding the calculations of RP, RV and RC, the reader is referred to Svensson [12, 13, 25].

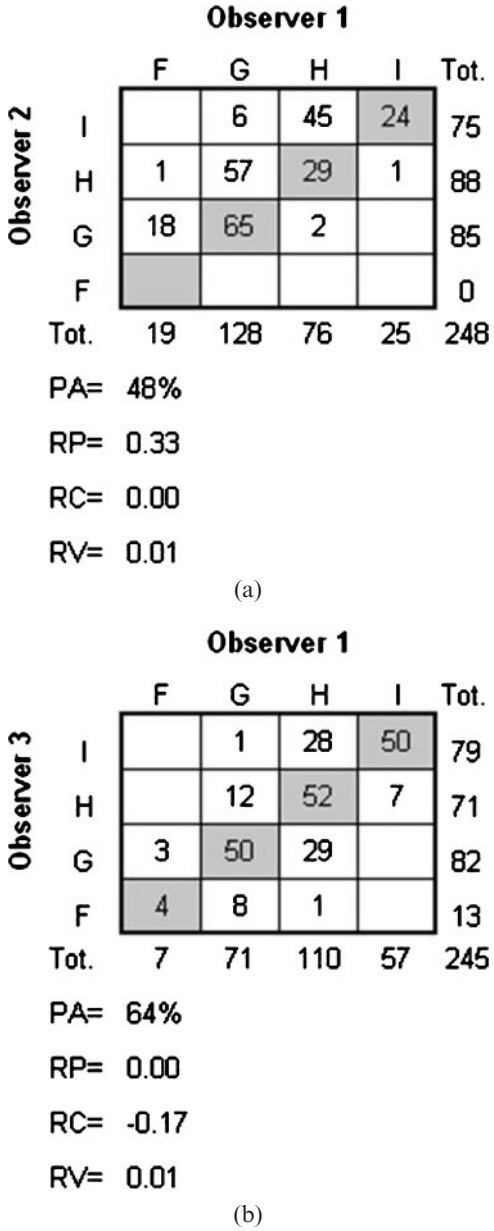


Figure 2. The frequency distribution of interobserver assessments with corresponding values of systematic disagreement in relative position (RP) and in relative concentration (RC), the measure of additional individual variability in assessments (RV) and percentage agreement (PA). (a) The distribution of assessments made by Observer 1 and Observer 2 regarding Q6 in the upper level of the brain (n=248), and (b) the distribution of assessments made by Observer 1 and Observer 3 regarding Q2 in the upper level of the brain (n=245).

Results

Intra-observer disagreement

The number of images assessed in the test–retest by each radiologist was 30 for each level of the brain. When assessing the lower level of the brain, Q2 (Table 2) was generally not applicable, and was thus excluded at this level.

The PA values (representing agreement) for Observer 1 ranged from 50% to 83% (Table 3). The results of the evaluation of the disagreement between the test–retest assessments of visual grading are given in Table 4. The

Table 3. Observer agreement

	Interobserver agreement				Intra-observer agreement		
	1 *vs* 2	1 *vs* 3	2 *vs* 3		1	2	3
Upper level of the brain							
Q1	66 (164/247)	65 (162/248)	55 (138/248)	Q1	83 (25/30)	77 (23/30)	70 (21/30)
Q2	62 (152/247)	64 (156/245)	53 (131/246)	Q2	53 (16/30)	67 (20/30)	63 (19/30)
Q3	61 (151/248)	66 (163/248)	57 (142/248)	Q3	77 (23/30)	83 (25/30)	63 (19/30)
Q4	67 (165/248)	68 (168/248)	61 (152/248)	Q4	83 (25/30)	77 (23/30)	60 (18/30)
Q5	63 (156/248)	60 (149/248)	49 (121/248)	Q5	73 (22/30)	73 (22/30)	87 (26/30)
Q6	48 (118/248)	70 (173/248)	44 (109/248)	Q6	77 (23/30)	63 (19/30)	60 (18/30)
Lower level of the brain							
Q1	58 (143/248)	59 (147/248)	60 (149/248)	Q1	67 (20/30)	67 (20/30)	80 (24/30)
Q3	67 (165/248)	57 (141/248)	57 (140/248)	Q3	77 (23/30)	77 (23/30)	60 (18/30)
Q4	52 (129/248)	64 (158/248)	51 (127/248)	Q4	70 (21/30)	63 (19/30)	67 (20/30)
Q5	65 (160/248)	48 (119/247)	41 (102/247)	Q5	50 (15/30)	63 (19/30)	77 (23/30)
Q6	52 (130/248)	61 (152/248)	44 (108/248)	Q6	80 (24/30)	63 (18/30)	70 (21/30)

The values of percentage agreement (PA) (%) and numerator/denominator within brackets for each observer (1, 2 and 3) and question (see Table 2) in the upper and lower levels of the brain.

RV values are negligibly small, except for Q5 in the lower level of the brain, which means that the observed disagreements are mainly owing to small systematic disagreements in the two assessments of the same images, with the following exceptions. The positive RP values for Q3 and Q6 in the upper level of the brain reveal that Observer 1 was more likely to assess an image as being of a poorer image quality on the second occasion than on the first; regarding Q5, the observer concentrated the assessments more on the second occasion than on the first (RC, 0.14). In the lower level of the brain, the negative RP value for Q4 indicated that Observer 1 was more likely to assess an image as being of a higher image quality on the second occasion than on the first.

The PA ranged from 63% to 83% for Observer 2 and from 60% to 87% for Observer 3 (Table 3). The main explanation of the observed disagreements between the test–retest assessments is systematic disagreement in the two assessments (Table 4). Both radiologists were more likely to assess an image as being of a poorer image quality on the second occasion.

Interobserver disagreement

When a question was not applicable to an image it was removed from the evaluation. The number of paired data ranged between 245 and 248. Table 3 gives the percentage agreement and Table 5 the values of RP, RV and RC for the interobserver disagreement analyses of the assessments regarding the upper and lower levels of the brain.

Observer 1 vs *Observer 2*

The percentage agreement in the assessments ranged from 48% to 67%, and the disagreement is mainly explained by systematic disagreement in the assessments made by the two observers as the RV values were small. For the upper level of the brain, only one of the 95% confidence intervals of the RP and RC values covered the zero value, which indicates a statistically significant bias between the observers. RP was positive for all questions except Q2, *i.e.* Observer 2 systematically graded the images as being of a poorer quality than did Observer 1. Figure 2a shows the disagreement pattern for Q6. The disagreeing pairs are situated above the diagonal, indicating that Observer 2 systematically used higher categories representing lower image quality than Observer 1. The significant RP value (0.33; 95% CI 0.29–0.38) and the negligible RV value confirm that the disagreement is explained by the systematic disagreement in relative position on the scale. RC was positive for all questions except Q6, *i.e.* Observer 2 systematically concentrated the assessments more than Observer 1. For the lower level of the brain, RC remains positive (except for Q6) whereas RP, in contrast to the upper level of the brain, was negative for all questions except Q6.

Observer 1 vs *Observer 3*

The percentage agreement of the assessments ranged from 48% to 70% and the disagreement is mainly explained by interobserver bias as the RV values are small. For the upper level of the brain, RC is negative for all questions except Q6, which means that Observer 1 most likely concentrated the assessments more than Observer 3. Figure 2b shows the paired assessments of Q2. The assessments made by Observer 1 are more concentrated to the central classifications than the assessments made by Observer 3, as is evident from the significant RC value (−0.17; 95% CI −0.23 to −0.10) and the negligible RV value. The RP values differ between the questions, but indicate significant interobserver bias in position for Q1, Q4, Q5 (upper level) and Q1, Q3, Q5 (lower level).

Observer 2 vs *Observer 3*

The percentage agreement in the assessments ranged from 41% to 61%. Observer 2 systematically concentrated the assessments more than Observer 3 for all questions except Q6, in both the upper and lower levels of the brain. Systematic disagreement in position on the scales was found for Q3, Q4 and Q6 (upper level), in that Observer 2 was more likely to classify the images as being of a poorer quality than Observer 3; the opposite was found for Q2 and Q5 (upper level). Corresponding results were found in the assessments of the lower level of the brain (Table 5) with the exception of Q4. The RV values were significant but negligible.

Table 4. Intra-observer disagreement

	Observer 1	Observer 2	Observer 3
Upper level of the brain			
RP (95% CI)			
Q1	−0.01 (−0.11 to 0.09)	0.15 (0.00 to 0.29)	0.12 (0.00 to 0.23)
Q2	0.00 (−0.15 to 0.16)	0.14 (−0.02 to 0.30)	0.04 (−0.09 to 0.18)
Q3	0.13 (0.00 to 0.26)	0.03 (−0.10 to 0.15)	0.19 (0.03 to 0.35)
Q4	0.02 (−0.09 to 0.14)	0.15 (0.01 to 0.28)	0.25 (0.11 to 0.40)
Q5	−0.03 (−0.17 to 0.11)	0.16 (0.03 to 0.28)	0.09 (0.01 to 0.17)
Q6	0.10 (−0.01 to 0.20)	0.16 (0.02 to 0.30)	0.13 (0.00 to 0.26)
RC (95% CI)			
Q1	0.04 (−0.11 to 0.20)	0.01 (−0.09 to 0.12)	−0.02 (−0.21 to 0.18)
Q2	0.01 (−0.21 to 0.23)	−0.06 (−0.19 to 0.08)	−0.05 (−0.27 to 0.16)
Q3	−0.06 (−0.21 to 0.08)	−0.03 (−0.14 to 0.09)	0.10 (−0.12 to 0.31)
Q4	−0.07 (−0.19 to 0.04)	−0.03 (−0.20 to 0.14)	0.06 (−0.18 to 0.30)
Q5	0.14 (−0.02 to 0.29)	0.00 (−0.19 to 0.19)	−0.07 (−0.21 to 0.08)
Q6	−0.10 (−0.28 to 0.07)	0.10 (−0.13 to 0.34)	−0.15 (−0.35 to 0.05)
RV (95% CI)			
Q1	0.00 (0.00 to 0.00)	0.01 (0.00 to 0.02)	0.00 (0.00 to 0.01)
Q2	0.02 (0.00 to 0.05)	0.02 (0.00 to 0.05)	0.01 (0.00 to 0.02)
Q3	0.00 (0.00 to 0.01)	0.00 (0.00 to 0.01)	0.01 (0.00 to 0.04)
Q4	0.00 (0.00 to 0.01)	0.00 (0.00 to 0.01)	0.01 (0.00 to 0.04)
Q5	0.01 (0.00 to 0.02)	0.00 (0.00 to 0.01)	0.00 (0.00 to 0.00)
Q6	0.00 (0.00 to 0.00)	0.01 (0.00 to 0.03)	0.00 (0.00 to 0.01)
Lower level of the brain			
RP (95% CI)			
Q1	−0.09 (−0.23 to 0.05)	0.16 (−0.01 to 0.32)	0.09 (−0.02 to 0.19)
Q3	−0.03 (−0.20 to 0.14)	0.14 (0.01 to 0.28)	0.19 (0.01 to 0.38)
Q4	−0.20 (−0.34 to −0.05)	0.25 (0.13 to 0.38)	0.20 (0.06 to 0.34)
Q5	0.03 (−0.18 to 0.24)	0.17 (0.02 to 0.33)	−0.03 (−0.20 to 0.14)
Q6	−0.01 (−0.13 to 0.12)	0.18 (0.03 to 0.32)	0.27 (0.12 to 0.42)
RC (95% CI)			
Q1	0.18 (−0.01 to 0.37)	0.03 (−0.10 to 0.15)	0.01 (−0.10 to 0.11)
Q3	0.00 (0.00 to 0.00)	0.02 (−0.12 to 0.15)	0.17 (−0.05 to 0.38)
Q4	−0.07 (−0.27 to 0.12)	−0.08 (−0.24 to 0.08)	0.17 (0.00 to 0.34)
Q5	−0.02 (−0.20 to 0.17)	−0.12 (−0.32 to 0.08)	0.00 (0.00 to 0.00)
Q6	−0.06 (−0.19 to 0.08)	−0.09 (−0.31 to 0.14)	0.03 (−0.13 to 0.19)
RV (95% CI)			
Q1	0.01 (0.00 to 0.03)	0.02 (0.00 to 0.06)	0.00 (0.00 to 0.01)
Q3	0.02 (0.00 to 0.05)	0.00 (0.00 to 0.02)	0.02 (0.00 to 0.05)
Q4	0.00 (0.00 to 0.00)	0.00 (0.00 to 0.00)	0.00 (0.00 to 0.00)
Q5	0.08 (0.00 to 0.17)	0.01 (0.00 to 0.03)	0.02 (0.00 to 0.05)
Q6	0.00 (0.00 to 0.01)	0.03 (0.00 to 0.06)	0.00 (0.00 to 0.00)

The table presents the measures of intra-observer disagreement in the different questions (Table 2) in the upper and lower levels of the brain. Values of systematic disagreement in position (RP) and in concentration (RC), and the measure of additional individual variability in assessments (RV) are given, together with the 95% confidence intervals (CI) of the measures.

Discussion

As mentioned in the introduction, evaluation of observer disagreement in MDCT image quality assessment has been sparsely investigated. Mayo et al [2] investigated the effect of dose reduction on intra-observer disagreement using McNemar's test for paired data, yielding a χ^2 statistic with one degree of freedom. Using this statistical analysis, they were able to show that a reduction in dose increased the intra-observer disagreement. Further investigations using the statistical approach demonstrated in this study may have determined whether the different disagreements were caused by the same proportions of systematic and random disagreement, or if they were caused by an increase in the random disagreement for example. Otherwise, the κ coefficient is often used as a measure of observer agreement adjusted for the chance-expected agreement. The κ coefficient is a single measure of agreement and does not explain the different sources of an observed disagreement and, therefore, is not a very informative measure when carrying out a thorough investigation of observer differences. The κ value also assumes unbiased pairs of assessment, which means identical marginal distributions, which is rarely the case in agreement studies. Our study has shown that the observed disagreements could mainly be explained by the systematic disagreement (bias) between and within the observers. The intra- and interobserver disagreement in the paired comparisons were comprehensively analysed, and the systematic and the occasional sources of disagreement were identified and measured by RP, RC and RV. The presence of systematic disagreement can be adjusted for by training the observers or by specifying the classifications further. Large individual variability is a sign of poor-quality scales or unstable examination situations. The study example is also an absolute visual grading study, *i.e.* all the images were graded separately

Table 5. Interobserver disagreement

	Observer 1 *vs* 2	Observer 1 *vs* 3	Observer 2 *vs* 3
Upper level of the brain			
RP (95% CI)			
Q1	0.06 (0.01 to 0.11)	0.10 (0.06 to 0.15)	0.06 (0.00 to 0.12)
Q2	−0.19 (−0.24 to −0.13)	0.00 (−0.05 to 0.05)	0.17 (0.11 to 0.23)
Q3	0.16 (0.10 to 0.22)	−0.05 (−0.11 to 0.00)	−0.21 (−0.27 to −0.15)
Q4	0.07 (0.02 to 0.13)	−0.08 (−0.13 to −0.03)	−0.16 (−0.22 to −0.10)
Q5	0.07 (0.01 to 0.13)	0.16 (0.11 to 0.22)	0.12 (0.05 to 0.20)
Q6	0.33 (0.29 to 0.38)	−0.02 (−0.07 to 0.02)	−0.36 (−0.41 to −0.31)
RC (95% CI)			
Q1	0.21 (0.15 to 0.27)	−0.10 (−0.17 to −0.03)	−0.30 (−0.36 to −0.24)
Q2	0.17 (0.10 to 0.24)	−0.17 (−0.23 to −0.10)	−0.34 (−0.41 to −0.27)
Q3	0.12 (0.05 to 0.19)	−0.01 (−0.08 to 0.05)	−0.14 (−0.22 to −0.07)
Q4	0.13 (0.07 to 0.20)	−0.02 (−0.09 to 0.04)	−0.17 (−0.24 to −0.10)
Q5	0.17 (0.11 to 0.23)	−0.18 (−0.25 to −0.11)	−0.34 (−0.41 to −0.27)
Q6	0.00 (−0.09 to 0.09)	0.02 (−0.04 to 0.07)	0.06 (−0.03 to 0.16)
RV (95% CI)			
Q1	0.00 (0.00 to 0.01)	0.01 (0.00 to 0.01)	0.01 (0.00 to 0.02)
Q2	0.01 (0.00 to 0.02)	0.01 (0.00 to 0.03)	0.01 (0.00 to 0.02)
Q3	0.03 (0.01 to 0.05)	0.02 (0.01 to 0.04)	0.02 (0.01 to 0.04)
Q4	0.02 (0.00 to 0.03)	0.01 (0.00 to 0.02)	0.02 (0.01 to 0.03)
Q5	0.02 (0.01 to 0.04)	0.02 (0.01 to 0.03)	0.06 (0.03 to 0.09)
Q6	0.01 (0.00 to 0.02)	0.01 (0.00 to 0.01)	0.01 (0.00 to 0.02)
Lower level of the brain			
RP (95% CI)			
Q1	−0.18 (−0.24 to −0.12)	−0.11 (−0.16 to −0.05)	0.07 (0.01 to 0.13)
Q3	−0.05 (−0.11 to 0.01)	−0.30 (−0.36 to −0.24)	−0.26 (−0.33 to −0.20)
Q4	−0.20 (−0.26 to −0.14)	−0.05 (−0.10 to 0.01)	0.15 (0.08 to 0.22)
Q5	−0.03 (−0.09 to 0.02)	0.30 (0.24 to 0.36)	0.34 (0.27 to 0.41)
Q6	0.25 (0.19 to 0.30)	−0.03 (−0.09 to 0.03)	−0.27 (−0.36 to −0.19)
RC (95% CI)			
Q1	0.24 (0.18 to 0.31)	0.05 (−0.02 to 0.12)	−0.18 (−0.24 to −0.12)
Q3	0.04 (−0.02 to 0.10)	−0.05 (−0.14 to 0.04)	−0.09 (−0.18 to −0.01)
Q4	0.18 (0.11 to 0.26)	0.00 (−0.06 to 0.06)	−0.20 (−0.27 to −0.12)
Q5	0.09 (0.03 to 0.15)	−0.26 (−0.36 to −0.17)	−0.37 (−0.47 to −0.28)
Q6	−0.05 (−0.13 to 0.03)	−0.01 (−0.07 to 0.04)	0.08 (0.00 to 0.16)
RV (95% CI)			
Q1	0.02 (0.01 to 0.04)	0.04 (0.02 to 0.07)	0.01 (0.00 to 0.02)
Q3	0.01 (0.01 to 0.02)	0.03 (0.01 to 0.05)	0.04 (0.01 to 0.07)
Q4	0.03 (0.01 to 0.05)	0.03 (0.01 to 0.06)	0.07 (0.04 to 0.12)
Q5	0.02 (0.01 to 0.04)	0.03 (0.02 to 0.05)	0.05 (0.03 to 0.08)
Q6	0.02 (0.01 to 0.03)	0.02 (0.01 to 0.04)	0.05 (0.01 to 0.09)

The table presents the measures of interobserver disagreement for each observer and question (Table 2) in the upper and lower levels of the brain. Values of systematic disagreement in position (RP) and in concentration (RC), and the measure of additional individual variability (RV) are given, together with the 95% confidence intervals (CI) of the measures.

in a random order. Radiologists are influenced by their experience of image quality and by the images graded previously in the study. It is therefore up to the observer to be as objective as possible and to consider how the image quality relates to the image quality criteria defined by the verbal rating scale. Other aspects on what could have affected the results of the reviewer were the viewing conditions. All radiologists used the same viewing station but were free to review whenever they had time in their daily schedule, thus representing a daily work basis. This resulted in the radiologists reviewing in varying time slots and at different times of the day. However, with the ViewDEX, all the reviewers' logins and logouts were registered and nothing out of the ordinary was noted. The viewing environment was constant as it took place in a quiet image archive where temperature and light were kept constant. It also is possible that the observers could have looked at different positions of the image when assessing; however, the instructions for the observers were to consider the reproduction of the entire structure.

When optimising MDCT examinations with a visual grading approach, the visual assessment of image quality is performed by several radiologists. Considering the various backgrounds of the radiologists working at a radiology department, it is preferable that the reviewing radiologists are representative of the cohort. The systematic disagreement between the radiologists in this study could perhaps be explained by their previous experience. Our study showed that the two radiologists specialised in neurology (Observers 1 and 3) were significantly more critical of the reproduction of the basal ganglia than Observer 2. In contrast, Observer 2 was significantly more critical of the overall image quality than the neuroradiologists. Observer 2 also concentrated the assessments more than the other observers. The intra-observer evaluation showed little systematic disagreement regarding the concentration of scale classifications; however, for the two radiologists with

least experience of visually grading image quality for research purposes (Observers 2 and 3), the disagreement found consisted mainly of a systematic downgrading of the image quality the second time the image was reviewed. With only three radiologists in our study, we have too few observers to be able to state that the results are related to their experience, but the method offers the opportunity to investigate a possible relationship. The findings in this study did not result in any further dose adjustments above those concluded in the separate article about the study example [3]. However, it did give a further dimension of information that was used as a basis for thorough discussions at the department regarding what should be considered a diagnostically useful image.

In conclusion, we have obtained information on the intra- and interobserver disagreement in a paediatric cerebral MDCT visual grading study which is, at the time of writing, not attainable by any other statistical approach suitable for paired, ordinal data. The statistical approach used enables the identification of systematic bias and level of additional individual variability. This provides information that will help us to gain a better understanding of the difference in radiologists' needs for diagnostic image quality.

Acknowledgments

We would like to thank Siw Johansson, Marianne Gustavsson, Stig Holtås, Sune Svensson, Sara Börjesson, Angelica Svalkvist and Markus Håkansson for all their help and expertise; and Magnus Båth and Sara Zachrisson for their help and expertise, and for critically reviewing this paper. GE Healthcare is thanked for the use of the noise simulation software, and the King Gustav V Jubilee Clinic Cancer Research Foundation and the Swedish Radiation Safety Authority for financial support.

References

1. Obuchowski NA. Sample size tables for receiver operating characteristic studies. AJR Am J Roentgenol 2000;175: 603–8.
2. Mayo JR, Kim KI, MacDonald SL, Johkoh T, Kavanagh P, Coxson HO, et al. Reduced radiation dose helical chest CT: effect on reader evaluation of structures and lung findings. Radiology 2004;232:749–56.
3. Ledenius K, Gustavsson M, Johansson S, Stalhammar F, Wiklund LM, Thilander-Klang A. Effect of tube current on diagnostic image quality in paediatric cerebral multidetector CT images. Br J Radiol 2009;82:313–20.
4. Frush DP, Slack CC, Hollingsworth CL, et al. Computer-simulated radiation dose reduction for abdominal multidetector CT of pediatric patients. AJR Am J Roentgenol 2002;179:1107–13.
5. NEMA. Digital Imaging and Communications in Medicine (DICOM) part 14: Grayscale standard display function. NEMA PS 3.14-2004. National Electrical Manufacturers Association, Rosslyn, VA, 2004.
6. European Commission. European Guidelines on quality criteria for computed tomography. Publication EUR 16262 EN. Luxembourg: Office for Official Publications of the European Communities, 1999.
7. Håkansson M, Svensson S, Zachrisson S, Svalkvist A, Båth M and Månsson LG. ViewDEX: an efficient and easy-to-use software for observer performance studies. Radiat Prot Dosimetry doi: 10.1093/rpd/ncq057(2010).
8. Håkansson M, Svensson S, Zachrisson S, Svalkvist A, Båth M, Månsson LG. ViewDEX 2.0: a Java-based DICOM-compatible software for observer performance studies. Proc. SPIE 2009;7263:72631G1–G10.
9. Dybkaer R, Jorgensen K. Measurement, value, and scale. Scand J Clin Lab Invest Suppl 1989;194:69–76.
10. Merbitz C, Morris J, Grip JC. Ordinal scales and foundations of misinference. Arch Phys Med Rehabil 1989;70:308–12.
11. Svensson E. Guidelines to statistical evaluation of data from rating scales and questionnaires. J Rehabil Med 2001;33: 47–8.
12. Svensson E. A coefficient of agreement adjusted for bias in paired ordered categorical data. Biometric J 1997;39:643–57.
13. Svensson E. Application of a rank-invariant method to evaluate reliability of ordered categorical assessments. J Epidemiol Biostat 1998;3:403–9.
14. Svensson ES, Ekholm J-E, von Essen S, Johansson C, A. Analysis of interobserver disagreement in the assessment of subarachnoid blood and acute hydrocephalus on CT scans. Neurolog Res 1996;18:487–94.
15. Reuterskiold MH, Lasson A, Svensson E, Kilander A, Stotzer PO, Hellstrom M. Diagnostic performance of computed tomography colonography in symptomatic patients and in patients with increased risk for colorectal disease. Acta Radiol 2006;47:888–98.
16. Svensson MH, Svensson E, Lasson A, Hellstrom M. Patient acceptance of CT colonography and conventional colonoscopy: prospective comparative study in patients with or suspected of having colorectal disease. Radiology 2002;222: 337–45.
17. Svensson E. Concordance between ratings using different scales for the same variable. Stat Med 2000;19:3483–96.
18. Lund I, Lundeberg T, Sandberg L, Budh CN, Kowalski J, Svensson E. Lack of interchangeability between visual analogue and verbal rating pain scales: a cross sectional description of pain etiology groups. BMC Med Res Methodol 2005;5:31.
19. Dahlin-Ivanoff S, Sonn U, Svensson E. Development of an ADL instrument targeting elderly persons with age-related macular degeneration. Disabil Rehabil 2001;23: 69–79.
20. Svensson E. Construction of a single global scale for multi-item assessments of the same variable. Stat Med 2001;20: 3831–46.
21. Sonn U, Svensson E. Measures of individual and group changes in ordered categorical data: application to the ADL staircase. Scand J Rehabil Med 1997;29:233–42.
22. Svensson E, Starmark JE. Evaluation of individual and group changes in social outcome after aneurysmal subarachnoid haemorrhage: a long-term follow-up study. J Rehabil Med 2002;34:251–9.
23. Lund I, Lundeberg T, Lonnberg L, Svensson E. Decrease of pregnant women's pelvic pain after acupuncture: a randomized controlled single-blind study. Acta Obstet Gynecol Scand 2006;85:12–19.
24. Engman E, Andersson-Roswall L, Svensson E, Malmgren K. Non-parametric evaluation of memory changes at group and individual level following temporal lobe resection for pharmaco-resistant partial epilepsy. J Clin Exp Neuropsychol 2004;26:943–54.
25. Svensson E. Svensson's method: professor in biostatistics, Sweden [cited 2009 January 22]. Available from: http://www.oru.se/templates/oruExtNormal____9746.aspx.