FULL PAPER

Visual grading evaluation of commercially available metal artefact reduction techniques in hip prosthesis computed tomography

[1,2]KARIN M ANDERSSON, MSc, [1]EVA NORRMAN, PhD, [2,3]HÅKAN GEIJER, MD, PhD, [2,3]WOLFGANG KRAUSS, MD, [4,5]YANG CAO, PhD, [2,3]JOHAN JENDEBERG, MD, [3,6]MATS GEIJER, MD, PhD, [3]MATS LIDÉN, MD, PhD and [1,2]PER THUNBERG, PhD

[1]Department of Medical Physics, Faculty of Medicine and Health, Örebro University, Örebro, Sweden
[2]School of Health and Medical Sciences, Örebro University, Örebro, Sweden
[3]Department of Radiology, Faculty of Medicine and Health, Örebro University, Örebro, Sweden
[4]Clinical Epidemiology and Biostatistics, School of Medical Sciences, Örebro University, Örebro, Sweden
[5]Unit of Biostatistics, Institute of Environmental Medicine, Karolinska Institutet, Stockholm, Sweden
[6]Department of Clinical Sciences, Lund University, Lund, Sweden

Address correspondence to: **Karin M Andersson**
E-mail: *karin.andersson@regionorebrolan.se*

Objective: To evaluate metal artefact reduction (MAR) techniques from four CT vendors in hip prosthesis imaging.

Methods: Bilateral hip prosthesis phantom images, obtained by using MAR algorithms for single-energy CT data or dual-energy CT (DECT) data and by monoenergetic reconstructions of DECT data, were visually graded by five radiologists using 10 image quality criteria. Comparisons between the MAR images and a reference image were performed for each scanner separately. Ordinal probit regression analysis was used.

Results: The MAR algorithms in general improved the image quality based on the majority of the criteria (up to between 8/10 and 10/10) with a statistical improvement in overall image quality ($p < 0.001$). However, degradation of image quality, such as new artefacts, was seen in some cases. A few monoenergetic reconstruction series improved the image quality ($p < 0.004$) for one of the DECT scanners, but it was only improved for some of the criteria (up to 5/10). Monoenergetic reconstructions resulted in worse image quality for the majority of the criteria (up to 7/10) for the other DECT scanner.

Conclusion: The MAR algorithms improved the image quality of the hip prosthesis CT images. However, since additional artefacts and degradation of image quality were seen in some cases, all algorithms should be carefully evaluated for every clinical situation. Monoenergetic reconstructions were in general concluded to be insufficient for reducing metal artefacts.

Advances in knowledge: Qualitative evaluation of the usefulness of several MAR techniques from different vendors in CT imaging of hip prosthesis.

INTRODUCTION

Metallic implants lead to creation of artefacts in CT images. The artefacts are caused by photon starvation and beam hardening effects and may limit the diagnostic value of the CT images, both close to the implant and in the surrounding tissues.[1] Hip prostheses cause severe metal artefacts, which limit possibility to recognize implant loosening and fractures, to diagnose inflammation or haematoma in the surrounding tissues or to diagnose pathology in the pelvic organs.

Over the years, several different strategies for reducing metal artefacts have been used clinically. Simple approaches such as increasing the tube current or tube potential have been shown to be insufficient in most cases.[1–3] These strategies also have drawbacks, *e.g.* the radiation dose to the patient is increased. In recent years, however, commercial metal artefact reduction (MAR) software, working on raw projection data, has been introduced by several CT vendors. Projection–interpolation methods are often used in these applications, which for some of the algorithms are implemented in an iterative process.[4–9] These MAR algorithms have previously been shown to reduce metal artefacts for several types of metallic objects, ranging from smaller implants, such as dental fillings, to larger orthopaedic devices.[5–18]

Images reconstructed from dual-energy CT (DECT) data have also been used as a way to enhance the diagnostic

value of metal artefact-degraded CT images.[18–23] DECT imaging enables image reconstruction from two energy sources, which allows creation of virtual monoenergetic images. This means that images are generated as though they would have been acquired with a high-energy beam. The user can choose for which kiloelectronvolt (keV) level the image should be generated. This theoretical image reconstruction at a high keV level makes it possible to reduce beam hardening artefacts, without increasing the actual tube voltage and thereby increasing the radiation dose to the patient.

Commercial MAR techniques have previously been evaluated for numerous applications, but only a few studies have evaluated several commercial MAR techniques in the same way.[5,12] These studies have been based on quantitative measures only, such as CT number accuracy and noise, hence comparative studies including visual grading evaluation are still lacking in literature.

Therefore, the aim of this study was to qualitatively evaluate several MAR techniques in CT imaging of metallic hip prostheses. The main objective was to evaluate the visualization of bone close to hip implants.

METHODS AND MATERIALS
Phantom
A phantom simulating a patient with bilateral hip implants was used in the study.[5] The phantom consisted of two chromium–cobalt hip prostheses inserted into the hip and femur bones of a calf by an orthopaedic surgeon. Almost no soft tissue was left. Based on CT imaging of the phantom, the bones were judged to be an adequate simulation of the human anatomy.

The bones with the inserted prostheses were placed in a water-filled rectangular-shaped polymethyl methacrylate box. The cross-sectional area of the phantom was $20 \times 40\,cm^2$, and the

Table 1. CT scan parameters used in the study, for the four different CT scanners

Parameter	CT scan parameters			
	Philips Healthcare (Best, Netherlands)	Toshiba Medical Systems (Otawara, Japan)	GE Healthcare (Milwaukee, WI)	Siemens Healthcare (Forchheim, Germany)
Scanner type	Philips Ingenuity Core (SE)	Toshiba Aquilion ONE™ ViSION Edition (SE)	GE Discovery™ 750HD (DE by fast kV switching)	Siemens SOMATOM® Definition Flash (DE with dual sources)
CT protocol	Helical (pitch 0.5)	Volume (SEMAR not compatible with helical scanning)	Helical (pitch 0.5)	Helical (pitch 0.5)
Tube potential (kVp)	120	120	SE: 120 DE: 80/140	SE: 120 DE: Sn140/100 (140 spectrum hardened by 0.1 mm tin)
$CTDI_{vol32}$ (mGy)	28	28	28	28
Collimation (mm)	64×0.625	280×0.5	64×0.625	128×0.6
Slice thickness (mm)	2 (1 mm increment)	2 (1 mm increment)	2 (1 mm increment)	2 (1 mm increment)
Reconstruction FOV (mm)	420	420	420	420
MAR technique	MAR algorithm (O-MAR)	MAR algorithm (SEMAR)	Monoenergetic reconstruction (110 keV) MAR algorithm (MARS)	Monoenergetic reconstruction (110 keV) DE-composition reconstruction (weight of −0.3)
IR	iDose Level 3 (range: 1–5)	AIDR 3D Level standard (range: mid, standard, strong)	ASIR Level 50% (range: 0–100%)	SAFIRE Level 3 (range: 1–5)
Soft kernel	B	FC08	Standard	D34 (FBP) Q30 (IR)
Sharper kernel	YB	FC30	Detail	D45 (FBP) Q50 (IR)

$CTDI_{vol32}$, volume CT dose index; FBP, filtered backprojection; FOV, field of view; IR, iterative reconstruction; MAR, metal artefact reduction; SE, single energy.

implants were placed about 20 cm apart. The bones with implants were centred in the phantom with plastic slabs and rods. The use of rods made it possible to place the bones with implants in approximately the same position for every CT scan.

Image acquisition

The hip prosthesis phantom was imaged with MAR techniques on four different CT scanners: Philips Ingenuity Core (Philips Healthcare, Best, Netherlands); Toshiba Aquilion ONE™ ViSION Edition (Toshiba Medical Systems, Otawara, Japan); GE Discovery™ 750HD (GE Healthcare, Milwaukee, WI) and SOMATOM® Definition Flash (Siemens Healthcare, Forchheim, Germany). In addition to acquiring CT images with the specific MAR technique of the scanner, 120-kVp CT images without using any MAR technique were also acquired for every CT scanner. The images acquired without any MAR technique are hereafter called uncorrected images.

The CT scan parameters for each CT scanner are summarized in Table 1. In addition to evaluating the MAR techniques, the effects of reconstruction kernel and iterative reconstruction (IR) on the metal artefacts were studied. Both uncorrected and MAR images were reconstructed with IR and filtered backprojection (FBP), and with a soft and sharper kernel. The choice of which kernels to use was based on recommendations from the application specialists for each CT scanner and on the kernels that are commonly used in the hospital's clinic for CT examinations in the pelvic area. The soft kernel is commonly used for examination of soft tissues in the pelvic area and the sharper kernel for depiction of bone. An intermediate level of the IR algorithm installed on the CT scanner in question was used in the evaluation.

The volume CT dose index ($CTDI_{vol32}$) was kept constant during all scans. The scan parameters, including the $CTDI_{vol32}$, were chosen with the purpose of being as similar as possible for the four different scanners.

Metal artefact reduction techniques

The evaluated single-energy CT scanners (Philips Ingenuity Core and Toshiba Aquilion ONE ViSION Edition) use MAR algorithm software to reduce the artefacts; O-MAR (metal artefact reduction for orthopaedic implants) (Philips Healthcare, Best, Netherlands)[4–6,9] and SEMAR (single-energy metal artefact reduction) (Toshiba Medical Systems, Otawara, Japan).[8] The evaluated DECT scanners (GE Discovery 750HD and Siemens SOMATOM Definition Flash) use monoenergetic reconstructions of DECT data to reduce metal artefacts.

The GE CT scanner, a single-source DECT with fast kilovoltage switching between 80 and 140 kVp in 0.25 ms, in addition uses a MAR software called MARS.[7] This algorithm is combined with the monoenergetic reconstructions. Monoenergetic reconstructions with and without the MARS software were evaluated.

The Siemens CT scanner, a dual-source DECT scanner with the capability of reconstructing monoenergetic images, uses no additional MAR algorithm. However, besides the monoenergetic reconstructions, the scanner uses an application called DE-composition. The DE-composition images are, according to the vendor, reconstructed based on similar principles as the monoenergetic reconstructions but uses an additional noise-reduction filter. The DE-composition setting is manually chosen by the user, on a scale from −1.0 to 1.0 which correlates to the weighting of the 140-kVp spectrum data and the 100-kVp spectrum data. A DE-composition value of −0.3 is recommended by the vendor for hip prosthesis imaging and was therefore used in this study.

For both the GE and the Siemens scanners, the energy level (keV level) used for the monoenergetic reconstructions is chosen by the operator. In this study, 110 keV was used for all monoenergetic reconstructions, which is in agreement with the results of a study by Meinel et al,[20] in which a DECT protocol was optimized for imaging of hip prostheses.

Visual grading evaluation

The MAR images and uncorrected images with modified scan parameters were evaluated by visual grading. Five radiologists independently evaluated axial CT images of the hip prosthesis phantom, blinded to the settings. The acquired images were compared with a reference image from the same CT scanner. The reference image was chosen to be the uncorrected 120-kVp CT image, reconstructed with a soft kernel and FBP. The visual inspection of the images was performed on dedicated picture archiving and communication system (PACS) reporting workstations by displaying the reference image on one-half of a medical-grade colour monitor with 1600 × 1200 pixel resolution and cycling through the test images on the other half. The monitors used were clinical systems which are regularly calibrated according to clinical routine. All comparisons were performed for each scanner separately; hence no images from different CT scanners were compared with each other.

Image quality was graded as much worse (−2), worse (−1), equal (0), better (1) or much better (2) compared with the reference image based on 10 image quality criteria (Figure 1). The radiologists used the default zoom and a standardized bone window (width/level of 2500/500) when evaluating the images, except for Criterion 10, where a soft-tissue window (width/level of 400/50) was used. The bone window was used since the main aim of the study was to evaluate the reproduction of the bone in the presence of metallic implants. The overall image quality was, however, also evaluated with a soft-tissue window for evaluating the water area (corresponding to the soft-tissue area in a patient). The radiologists had access to the entire image stack.

Image quality Criteria 1–2 of this study are from the pelvis section of the European Guidelines for Multislice CT.[24] Since no other existing image quality criteria were considered applicable for the phantom used, the remaining criteria were in-house developed together with a consultant radiologist to fit the purpose of this study.

The phantom was designed to make it possible to evaluate how the MAR techniques affect the image quality of bone surrounding hip prostheses, which is of importance when recognizing implant loosening and fractures (Criteria 5–6 and 8).

Figure 1. The image quality criteria used in the visual grading evaluation of the hip prosthesis phantom images. The areas corresponding to the different criteria are marked in the CT images. The images shown here were included in the instructions given to the radiologists.

Image quality criteria

1. Visually sharp reproduction of cortical bone in the pelvis, cranial to the prostheses without metal in the image.

2. Visually sharp reproduction of trabecular bone in the pelvis, cranial to the prostheses.

3. Reproduction of water in the middle of the pelvis in a slice transecting the centre of the left femoral head.

4. Reproduction of water ventral to the centre of the left femoral head.

5. Visually sharp reproduction of cortical and trabecular bone medial to the head of the left femoral prosthesis.

6. Visually sharp reproduction of cortical and trabecular bone ventral medial to the head of the left femoral prosthesis.

7. Reproduction of the head and cup of the right femoral prosthesis.

8. Visually sharp reproduction of cortical and trabecular bone 2 cm cranial to the tip of the stem of the right femoral prosthesis.

9. Overall image quality in the images containing metal.

10. Overall image quality in the images containing metal with soft tissue window (width/level of 400/50).
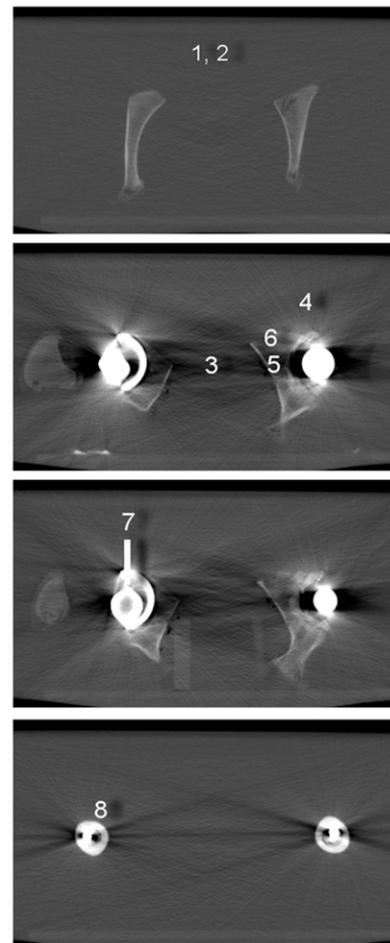
However, image quality criteria concerning the water area surrounding the bones and overall image quality were also included to evaluate the overall change in image quality for a certain setting (Criteria 3–4 and 9–10). The image quality of bones in image slices without any metal present was also evaluated (Criteria 1–2).

In order to evaluate the depiction of the implant itself, the radiologists graded the reproduction of the head and cup of the prosthesis (Criterion 7) and also measured the thickness of the cup. The radiologists were instructed to measure the thickness of the ventral part of the metallic right acetabular prosthesis at the mid-level of the head. The thickness of the cup was measured with a vernier caliper for comparison.

Statistical analysis

The median value of the radiologists' scores was calculated for each criterion and displayed in bar diagrams. The median scores for the different criteria were marked by colours in the diagrams, which makes it possible to see for which criterion the image quality was improved/worsened. If the image quality was considered improved based on a criterion, a positive value is shown

in the diagram (+0.1 for better image quality and +0.2 for much better). Likewise, a negative value is shown when the image quality was graded as worse based on a criterion (−0.1 for worse image quality and −0.2 for much worse). This means that the total score is at maximum 2 (corresponding to all of the 10 criteria considered much better) or at minimum −2 (corresponding to all of the 10 criteria considered much worse). When the score for a certain criterion is not seen in the diagram, the median score was zero (meaning equal image quality to the reference image).

To determine if the complete image quality (composed of the 10 criteria) was statistically different compared with the reference image quality, regression analysis was used. This approach is an established way to statistically analyse data from visual grading experiments. Smedby and Fredrikson[25] described this way of handling dependent variables defined on an ordinal scale. In their article, the ordinal logistic regression (OLR) model was applied to the visual grading data. An equivalent approach, the ordinal probit regression (OPR) model was used in our analysis. The OPR model uses the normal distribution for the latent variable, instead of the logistic distribution. OLR and OPR models are essentially the same, except for that OPR model

approaches the probabilities of 0 and 1 quicker than the OLR model.[26] Since the scores from our visual grading experiment were extremely centralized (Scores 3 and 4 account for >85% of all gradings), the OPR model was used to increase the distinguishability of the model around the median score.

Coefficients of the probability unit were estimated from the OPR model, by adjusting for image quality criterion and radiologist. Scores for each image quality criterion and from all radiologists were included in these calculations. In the analysis, every criterion was considered to be of equal importance for the image quality. A positive coefficient indicates that the acquired image has a greater probability of receiving a higher image quality score than the reference image. A negative coefficient means that the acquired image has a smaller probability of receiving a higher score than the reference image. In such cases where the images were judged to be equal to the reference image, based on every image quality criterion, the coefficient and confidence interval could not be estimated using the OPR model.

When estimating coefficients of the OPR model, sandwich robust estimator was used to estimate standard error when variance was abnormally large,[27] and the jackknife resampling method was used to estimate coefficient and corresponding credible interval when the OPR had convergence problem in using maximum likelihood estimation.[28]

Adjusted p-values for multiple comparisons were calculated using Holm's sequential Bonferroni method.[29] The adjustment considered all the 36 comparisons of the 4 CT vendors. A p-value of <0.05 was considered statistically significant. STATA® v. 14.1 (StataCorp LP, College Station, TX) was used to perform the statistical analysis.

RESULTS
In Figures 2–5, representative images are shown of the hip prosthesis phantom acquired with the Philips CT (Figure 2), Toshiba CT (Figure 3), GE CT (Figure 4) and Siemens CT (Figure 5) scanners, with and without MAR techniques. The median values of the radiologists' scores from the visual grading of the phantom images are shown in Figure 6.

In Table 2 the results of the statistical analysis are presented. The coefficients of acquired or reference image variable and corresponding confidence intervals are shown, together with the adjusted p-value.

Philips
In the visual grading of the images from the Philips CT scanner, every O-MAR image series received significantly higher scores ($p = 0.021$ for one series and $p < 0.001$ for the rest) than the reference image (Figure 6a, Table 2). In Table 2, it can be seen that the OPR model coefficients for the O-MAR images were all positive (ranging from 0.83 to 18.90).

The uncorrected series reconstructed with a sharper kernel and IR also received significantly higher scores ($p = 0.003$) than the reference image, but the total score was not as high as for the O-MAR image series. The O-MAR images reconstructed with IR in general resulted in higher scores.

Toshiba
The Toshiba SEMAR CT images reconstructed with a soft kernel received significantly higher scores ($p < 0.001$) than the reference image (Figure 6b, Table 2). For the SEMAR images combined with IR, the image quality of the head and the cup of the prosthesis (Criterion 7) was, however, considered much worse than the reference image. This degradation in image quality appeared in the form of high-density streaks adjacent to the implant.

One of the uncorrected Toshiba image series showed significantly improved image quality ($p < 0.001$). The SEMAR images, however, received higher scores for the image quality adjacent to the stem (Criterion 8) than the uncorrected images. The Toshiba CT images reconstructed with a sharper kernel and FBP were scored as worse or much worse for almost every criterion.

GE
The results of the visual grading of the uncorrected GE images showed that the use of sharper reconstruction kernel or IR did not change the perceived image quality (Figure 6c, Table 2). In general, the monoenergetic images showed improvement in only the reproduction of the cup of the prosthesis (Criterion 7) compared with the GE reference image and showed degraded image quality based on a number of other criteria.

When MARS was combined with the monoenergetic reconstructions, the image quality was improved for several criteria. The image quality adjacent to the stem (Criterion 8) was considered better or much better in the monoenergetic reconstructions combined with MARS. An overall significant improvement in image quality ($p < 0.001$) was only seen for the MARS image reconstructed with FBP and a soft kernel.

Figure 2. Uncorrected (a) and O-MAR (b) philips CT (Best, Netherlands) images acquired with 120 kVp, soft kernel and filtered backprojection.
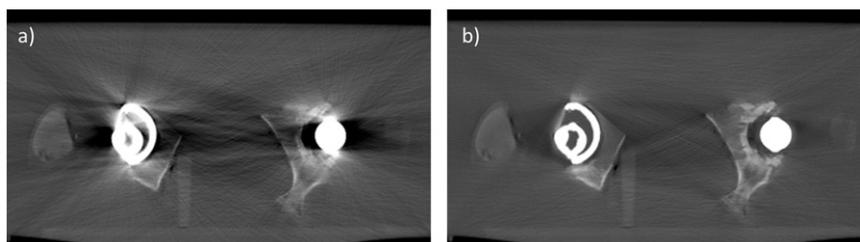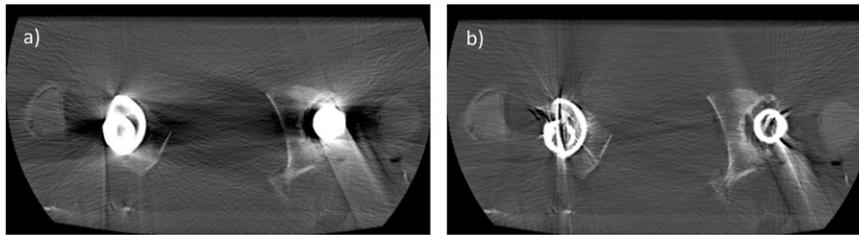
Figure 3. Uncorrected (a) and SEMAR (b) Toshiba CT (Otawara, Japan) images acquired with 120 kVp, soft kernel and filtered backprojection.

The image quality of bone in images without any metal (Criteria 1–2) was considered worse than in the reference image for all monoenergetic reconstructed series, both with and without the use of MARS.

Siemens

In the visual grading evaluation of the Siemens CT images, three of the monoenergetic reconstruction series showed a significantly improved image quality ($p < 0.004$) compared with the Siemens reference image (Figure 6d, Table 2). However, none of the series received a higher total score than 0.5. No series reconstructed with the DE-composition application resulted in a significant image quality improvement.

Overall, the monoenergetic reconstructions and the DE-composition images showed improved image quality adjacent to the head, cup and stem (Criteria 7–8) and improved overall image quality in some cases (Criteria 9–10). The uncorrected images acquired with a sharper kernel improved the reproduction of bone in image slices containing no metal (Criteria 1–2) but resulted in worse image quality based on several other criteria.

Cup thickness

Figure 7 shows the mean values of the radiologists' measurements of the cup. The cup thickness was physically measured to approximately 7 mm. The results show that monoenergetic reconstructions combined with the MARS algorithm of the GE CT scanner depicted the cup as being as thin as 2 mm. In the uncorrected Toshiba images reconstructed with IR, the cup was measured to up to 9 mm.

DISCUSSION

This study evaluates the use of MAR techniques in hip prosthesis images for four CT vendors in a consistent way. However, no images from different CT scanners are compared with each other and therefore this evaluation could not be used as a direct comparison between the different scanners.

The O-MAR algorithm of the Philips CT scanner generally improved the image quality for the majority of the criteria, and every O-MAR series showed significantly improved image quality. The Toshiba SEMAR series reconstructed with a soft kernel and only one of the GE MARS series were shown to be of significantly higher image quality than the corresponding reference images. The lack of significance was due to the fact that the image quality was considered to be worse based on a couple of criteria, even though the image quality was scored to be improved for the majority of the criteria.

The images obtained with monoenergetic reconstruction alone, acquired with the GE CT and the Siemens CT scanners, only resulted in improved image quality based on a few criteria. The

Figure 4. A 120-kVp CT image from the GE CT (Milwaukee, WI) (a) shown together with GE DECT images reconstructed with a monoenergetic level of 110 keV without (b) and with the metal artefact reduction algorithm MARS (c). A soft kernel and filtered backprojection are used for all images shown.
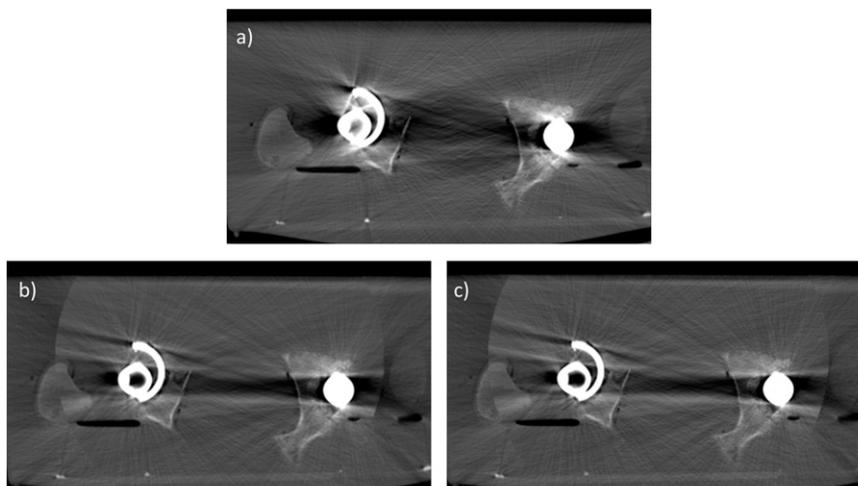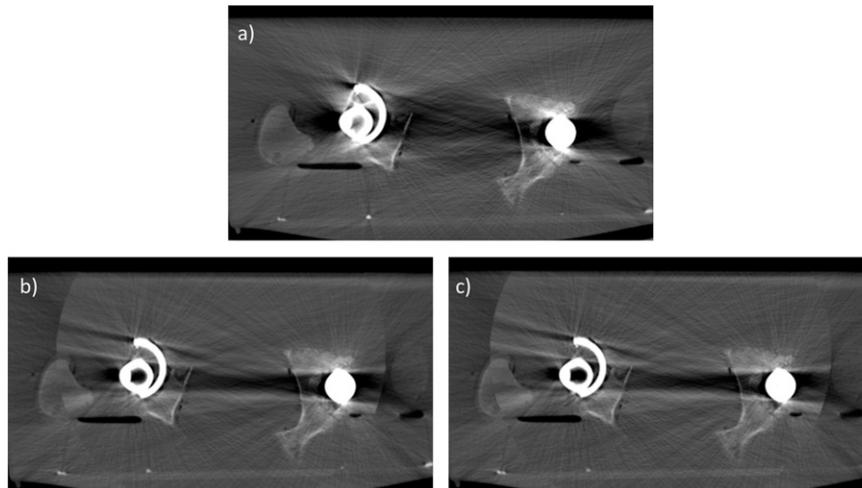
Figure 5. A 120-kVp CT image from the Siemens CT (Forchheim, Germany) (a) shown together with Siemens DECT images reconstructed with a monoenergetic level of 110 keV (b) and with a DE-composition setting of −0.3 (c). A soft kernel and filtered backprojection are used for all images shown.

monoenergetic images of the GE CT, reconstructed without the MARS algorithm, were even scored as worse than the reference image for several image regions (up to 7/10 of the criteria). A few of the Siemens monoenergetic reconstruction series showed significantly overall improved image quality, but the total scores were relatively low. The highest total score for the Siemens monoenergetic images was 0.5, compared with maximum total scores of 1.2, 0.9 and 0.8 for the MAR algorithm images from the Philips CT, Toshiba CT and GE CT scanners, respectively.

Several previous studies of monoenergetic reconstructions have concluded effective reduction of artefacts caused by larger metallic implants,[19–22] whereas other authors have reported on less efficient reduction of metal artefacts when monoenergetic reconstruction is used solely, without any additional MAR algorithm.[5,12] Monoenergetic reconstruction is utilized to reduce beam hardening artefacts. However, since artefacts due to metallic implants are also caused by photon starvation among other effects, the exclusive use of monoenergetic reconstruction may not be sufficient for reducing artefacts caused by large orthopaedic implants, which the result of the current study also indicates.

Commercial MAR algorithms have been shown to reduce metal artefacts caused by large orthopaedic implants.[5–15] However, additional artefacts created by the MAR algorithms have been stated as a drawback.[5,7] Han et al[7] evaluated monoenergetic reconstructions of a GE CT scanner, with and without MAR software, and reported improved overall image quality in the pelvic cavity in patients with hip prostheses when MAR was used, but new artefacts were also seen when using the MAR algorithm. The creation of new artefacts when using MAR algorithms was clearly seen also in the current study, particularly when the SEMAR algorithm was used for the Toshiba CT scanner.

A previous study performed by Gondim Teixeira et al[8] showed that the SEMAR algorithm, in combination with IR, improved the image quality of periarticular soft-tissue structures in patients with hip prostheses. The depiction of structures adjacent to the prostheses, the iliopsoas tendon and the sciatic nerve, was improved when SEMAR was used but still of mediocre quality. Our evaluation of the SEMAR algorithm also showed that the image quality was improved in several image areas, but the image quality adjacent to the head and cup of the prostheses was, in general, considered worse or much worse compared with the reference image. This image degradation appeared as additional streaking artefacts close to the head of the prosthesis. The effect was especially distinct when the SEMAR algorithm was used in combination with IR.

In the current study, the radiologists generally scored the reproduction of the head and cup of the prosthesis (Criterion 7) in the GE MARS images as worse than or equal to the reference image. The reason for this might be that even though artefacts in some of the areas close to the head and cup were reduced, the cup itself was depicted as very thin and even partly disappeared. The disappearance of parts of the cup was confirmed by the image measurements performed by the radiologists (Figure 7). Disappearance of metal implants and underestimation of implant size in images reconstructed with MAR algorithms have previously been reported.[10–12] The findings of both disappearance of metal implants and the creation of new artefacts suggest that images reconstructed with and without the MAR algorithm should always be reviewed together to reduce the risk of misinterpretation.

An overall preferable choice of reconstruction kernel and reconstruction technique was not possible to decide based on this evaluation. The GE MARS images reconstructed with IR resulted in lower scores than the corresponding FBP images. For the Philips O-MAR images, on the other hand, the images reconstructed with IR resulted in further improved image quality, compared with the FBP images. However, reconstruction technique and reconstruction kernel used should be chosen according to the clinical question in the specific case.

Figure 6. The result of the visual grading study of the images from the four CT scanners; (a) Philips Healthcare (Best, Netherlands), (b) Toshiba Medical Systems (Otawara, Japan), (c) GE Healthcare (Milwaukee, WI) and (d) Siemen Healthcare (Forchheim, Germany). The scores are median values of the five radiologists' visual grading of the images as much worse, worse, equal, better or much better compared with the reference image. The scores for the different image quality criteria [Criteria (Cr.) 1–10] are shown by different colours. The total score is at maximum of 2 (corresponding to all of the 10 criteria considered much better) or at minimum of −2 (corresponding to all of the 10 criteria considered much worse). Where a certain criterion is not marked in the diagram, the median value was zero in that case. FBP, filtered backprojection; IR, iterative reconstruction.
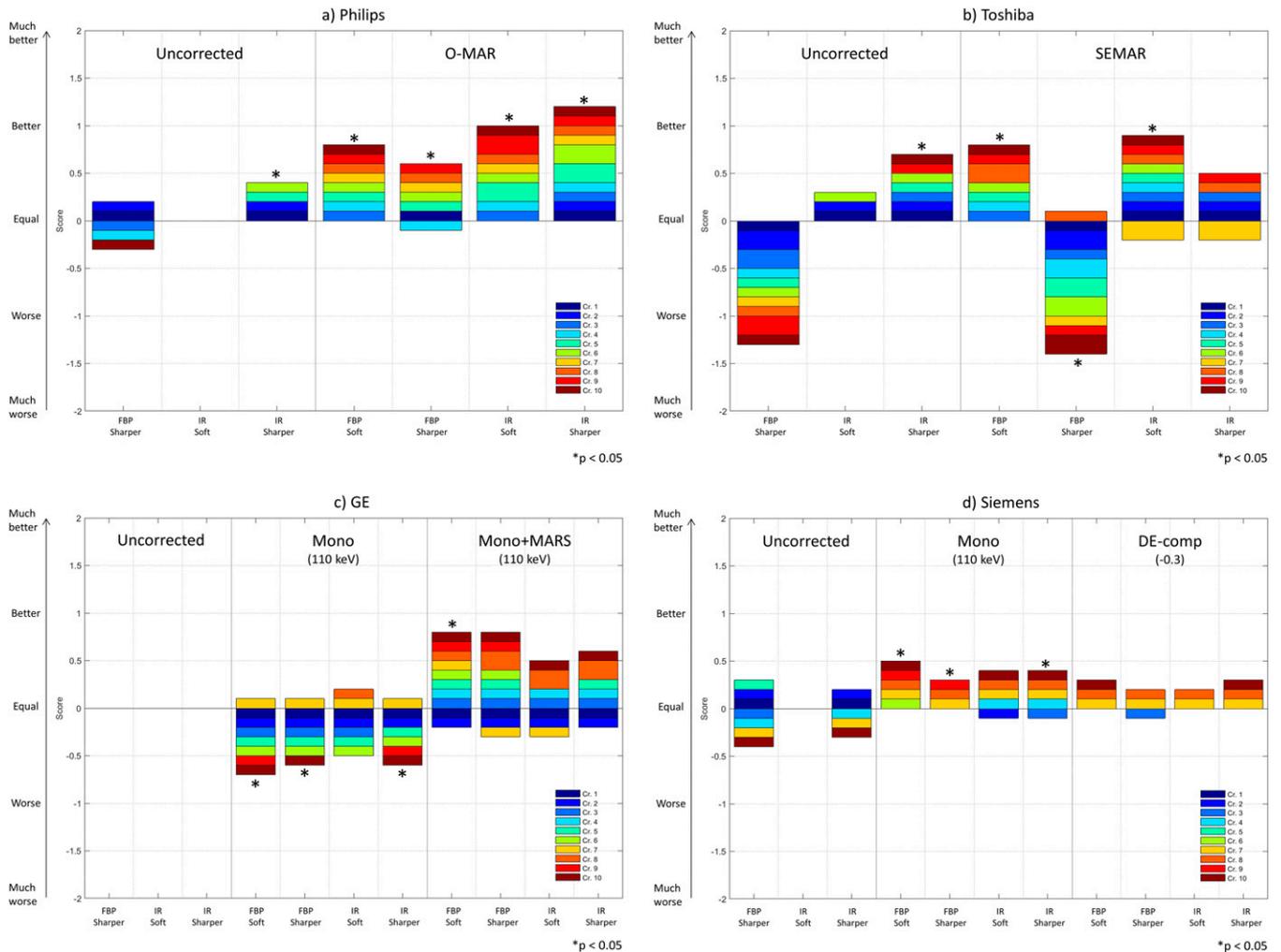
The use of a sharper kernel may not be appropriate for diagnosis of soft tissues and organs in the pelvic region, and a soft kernel in combination with a MAR algorithm may then be preferred.

The main objective of this study was to evaluate the visualization of bone close to hip prostheses; hence a phantom containing bone was used. The image quality of bone adjacent to the prosthesis is of interest when diagnosing prosthesis loosening or fractures. The result of this evaluation indicates for which clinical questions the different MAR techniques could be suitable. For example, if diagnosing prosthesis loosening is the purpose of the CT examination, improved image quality in areas close to the head, cup and stem of the prosthesis would especially be of interest (Criteria 7–8).

A study of a phantom containing soft tissue would also be of interest, to be able to analyse how the MAR techniques affect such structures. In this study, water was used as a substitute for soft tissue. Water has a lower attenuation coefficient than muscles which also implies that it may be of value to design a phantom where soft tissue substitutes some of the water volume, to obtain a more representative simulation of a human body. To evaluate the MAR techniques in the case of a unilateral hip prosthesis would also be of interest.

Another limitation of the design of the phantom used in this study may be that the dimensions of the calf bones were too large to represent human bones. It was noted that the new artefacts created adjacent to the head and cup of the prostheses

Table 2. The result of the statistical analysis of the visual grading evaluation where images from CT scanners from (a) Philips Healthcare (Best, Netherlands), (b) Toshiba Medical Systems (Otawara, Japan), (c) GE Healthcare (Milwaukee, WI) and (d) Siemens Healthcare (Forchheim, Germany) were compared with a reference series from the same CT. The estimated coefficient, shown together with the confidence level (CI), shows whether the tested image has a greater probability (positive value) or smaller probability (negative value) of receiving a higher score than reference image. The *p*-values (adjusted for multiple comparisons) which indicate a significant improvement in image quality are shown in bold

MAR	Reconstruction method	Kernel	Coefficient (95% CI)	*p*-value (adjusted)
Philips				
No	FBP	Sharper	−0.67 (−1.21, −0.12)	0.192
No	IR	Soft	0.23 (−0.52, 0.98)	1.000
No	IR	Sharper	1.25 (0.61, 1.90)	**0.003**
O-MAR	FBP	Soft	2.84 (1.99, 3.70)	**<0.001**
O-MAR	FBP	Sharper	0.83 (0.33, 1.32)	**0.021**
O-MAR	IR	Soft	18.90 (8.94, 28.85)[a]	**<0.001**
O-MAR	IR	Sharper	10.20 (8.78, 11.62)[a]	**<0.001**
Toshiba				
No	FBP	Sharper	−58.93 (−107.85, −10.01)[b]	0.209
No	IR	Soft	0.93 (0.29, 1.57)	0.083
No	IR	Sharper	2.32 (1.44, 3.21)	**<0.001**
SEMAR	FBP	Soft	1.70 (1.07, 2.32)	**<0.001**
SEMAR	FBP	Sharper	−2.72 (−3.45, −1.98)	<0.001
SEMAR	IR	Soft	2.46 (1.77, 3.15)	**<0.001**
SEMAR	IR	Sharper	0.72 (0.22, 1.22)	0.083
GE				
No	FBP	Sharper	–	1.000
No	IR	Soft	–	1.000
No	IR	Sharper	–	1.000
Mono	FBP	Soft	−1.69 (−2.38, −1.02)	<0.001
Mono	FBP	Sharper	−1.46 (−2.11, −0.81)	<0.001
Mono	IR	Soft	−0.81 (−1.37, −0.25)	0.074
Mono	IR	Sharper	−1.31 (−1.90, −0.71)	<0.001
MARS	FBP	Soft	1.13 (0.60, 1.66)	**<0.001**
MARS	FBP	Sharper	0.75 (0.24, 1.26)	0.074
MARS	IR	Soft	0.15 (−0.32, 0.62)	1.000
MARS	IR	Sharper	0.55 (0.04, 1.05)	0.330
Siemens				
No	FBP	Sharper	0.06 (−0.46, 0.57)	1.000
No	IR	Soft	–	1.000
No	IR	Sharper	−0.29 (−0.79, 0.21)	1.000
Mono	FBP	Soft	1.32 (0.63, 2.02)	**0.004**
Mono	FBP	Sharper	1.47 (0.73, 2.21)	**0.002**
Mono	IR	Soft	0.80 (0.21, 1.38)	0.105
Mono	IR	Sharper	1.79 (0.91, 2.67)	**0.002**
DE-composition	FBP	Soft	1.12 (0.38, 1.87)	0.060

(Continued)

Table 2. (Continued)

MAR	Reconstruction method	Kernel	Coefficient (95% CI)	*p*-value (adjusted)
DE-composition	FBP	Sharper	−0.14 (−0.64, 0.35)	1.000
DE-composition	IR	Soft	0.76 (0.16, 1.36)	0.182
DE-composition	IR	Sharper	0.91 (0.18, 1.64)	0.195

FBP, filtered backprojection; MAR, metal artefact reduction; IR, iterative reconstruction.
[a]Sandwich robust method.
[b]Jackknife method.

in the Toshiba SEMAR CT images were especially severe in the sections containing the thickest bone parts.

To further evaluate the MAR techniques, the scan protocols should be optimized both in the aspect of dose and of the level of IR. In this study, the same $CTDI_{vol32}$ and one intermediate level of IR was used to keep the scan protocol as similar as possible for the four CT scanners. In the case of monoenergetic reconstructions, different keV levels should also be tested further. The image quality of bone in sections without any metal present became worse when monoenergetic reconstructions were applied for the GE CT scanner. These kinds of effects should be carefully evaluated when varying the keV level of monoenergetic reconstructions.

CONCLUSION

This visual grading study of bilateral hip prosthesis phantom CT images showed that the MAR algorithms tested significantly improved the image quality. The image quality was in general improved based on the majority of the criteria. However, new artefacts and disappearance of parts of the metallic implants were noted. Hence, careful evaluation of a MAR algorithm, for the specific clinical situation considered, is always necessary. The use of the tested monoenergetic reconstructions alone only improved image quality in a few image regions or even worsened image quality based on several criteria. Monoenergetic reconstructions were therefore concluded to be insufficient for reducing metal artefacts caused by hip prostheses.

Figure 7. The result of the measurements of the cup thickness in the phantom images acquired with CT scanners from Philips Healthcare (Best, Netherlands), Toshiba Medical Systems (Otawara, Japan), GE Healthcare (Milwaukee, WI) and Siemens Healthcare (Forchheim, Germany). The cup thickness was physically measured to 7 mm. DE, dual energy; FBP, filtered backprojection; MAR, metal artefact reduction; IR, iterative reconstruction.
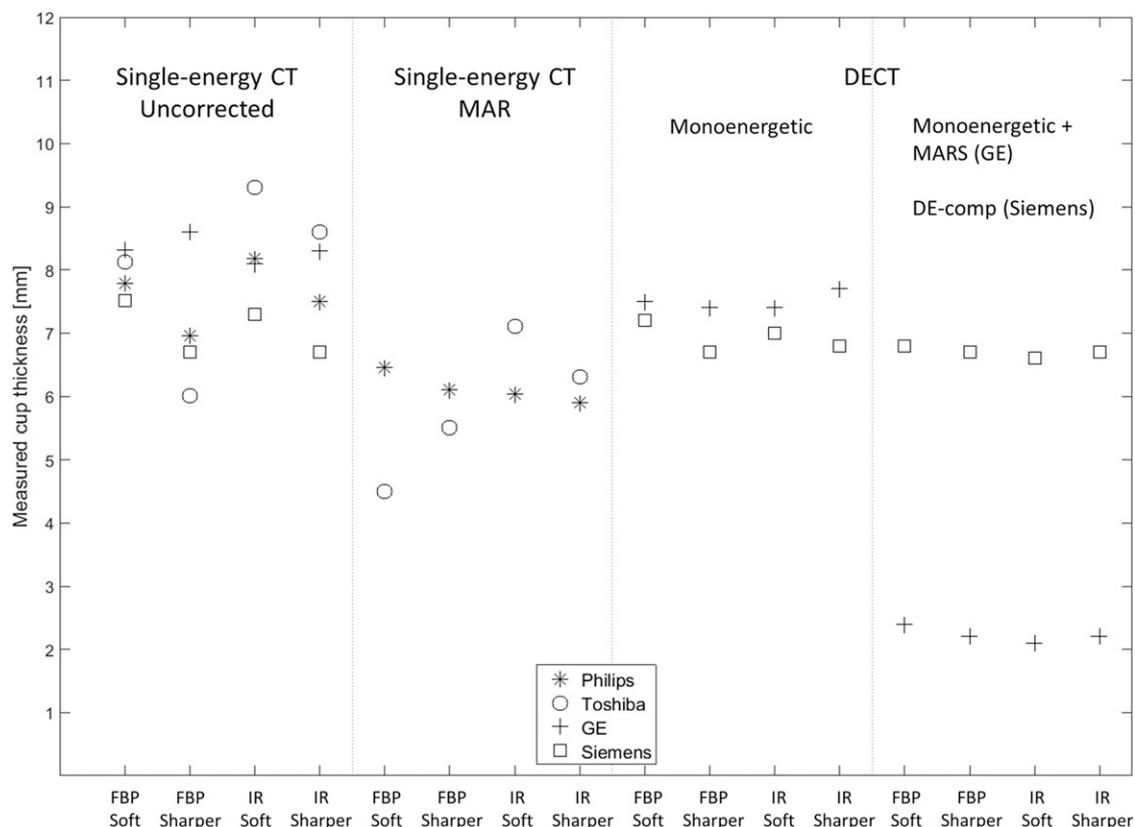
REFERENCES

1. Barrett JF, Keat N. Artefacts in CT: recognition and avoidance. *Radiographics* 2004; **24**: 1679–91. doi: http://dx.doi.org/10.1148/rg.246045065

2. Lee MJ, Kim S, Lee SA, Song HT, Huh YM, Kim DH, et al. Overcoming artifacts from metallic orthopedic implants at high-field-strength MR imaging and multi-detector CT. *Radiographics* 2007; **27**: 791–803. doi: http://dx.doi.org/10.1148/rg.273065087

3. Haramati N, Staron RB, Mazel-Sperling K, Freeman K, Nickoloff EL, Barax C, et al. CT scans through metal scanning technique *versus* hardware composition. *Comput Med Imaging Graph* 1994; **18**: 429–34. doi: http://dx.doi.org/10.1016/0895-6111(94)90080-9

4. *Metal artefact reduction for orthopedic implants (O-MAR)*. Cleveland, OH: Philips Healthcare; 2012. [Updated 8 January 2012; cited 9 May 2016]. Available from: http://clinical.netforum.healthcare.philips.com/us_en/Explore/White-Papers/CT/Metal-Artifact-Reduction-for-Orthopedic-Implants-(O-MAR)

5. Andersson KM, Nowik P, Persliden J, Thunberg P, Norrman E. Metal artefact reduction in CT imaging of hip prostheses—an evaluation of commercial techniques provided by four vendors. *Br J Radiol* 2015; **88**: 20140473. doi: http://dx.doi.org/10.1259/bjr.20140473

6. Andersson KM, Ahnesjö A, Vallhagen Dahlgren C. Evaluation of a metal artifact reduction algorithm in CT studies used for proton radiotherapy treatment planning. *J Appl Clin Med Phys* 2014; **15**: 4857. doi: http://dx.doi.org/10.1120/jacmp.v15i5.4857

7. Han SC, Chung YE, Lee YH, Park KK, Kim MJ, Kim KW. Metal artifact reduction software used with abdominopelvic dual-energy CT of patients with metal hip prostheses: assessment of image quality and clinical feasibility. *AJR Am J Roentgenol* 2014; **203**: 788–95. doi: http://dx.doi.org/10.2214/AJR.13.10980

8. Gondim Teixeira PA, Meyer JB, Baumann C, Raymond A, Sirveaux F, Coudane H, et al. Total hip prosthesis CT with single-energy projection-based metallic artifact reduction: impact on the visualization of specific periprosthetic soft tissue structures. *Skeletal Radiol* 2014; **43**: 1237–46. doi: http://dx.doi.org/10.1007/s00256-014-1923-5

9. Li H, Noel C, Chen H, Harold Li H, Low D, Moore K, et al. Clinical evaluation of a commercial orthopedic metal artifact reduction tool for CT simulations in radiation therapy. *Med Phys* 2012; **39**: 7507–17. doi: http://dx.doi.org/10.1118/1.4762814

10. Wang F, Xue H, Yang X, Han W, Qi B, Fan Y, et al. Reduction of metal artifacts from alloy hip prostheses in computer tomography. *J Comput Assist Tomogr* 2014; **38**: 828–33. doi: http://dx.doi.org/10.1097/RCT.0000000000000125

11. Lee YH, Park KK, Song HT, Kim S, Suh JS. Metal artefact reduction in gemstone spectral imaging dual-energy CT with and without metal artefact reduction software. *Eur Radiol* 2012; **22**: 1331–40. doi: http://dx.doi.org/10.1007/s00330-011-2370-5

12. Huang JY, Kerns JR, Nute JL, Liu X, Balter PA, Stingo FC, et al. An evaluation of three commercially available metal artifact reduction methods for CT imaging. *Phys Med Biol* 2015; **60**: 1047–67. doi: http://dx.doi.org/10.1088/0031-9155/60/3/1047

13. Hilgers G, Nuver T, Minken A. The CT number accuracy of a novel commercial metal artefact reduction algorithm for large orthopedic implants. *J Appl Clin Med Phys* 2014; **15**: 4597. doi: http://dx.doi.org/10.1120/jacmp.v15i1.4597

14. Sonoda A, Nitta N, Ushio N, Nagatani Y, Okumura N, Otani H, et al. Evaluation of the quality of CT images acquired with the single energy metal artifact reduction (SEMAR) algorithm in patients with hip and dental prostheses and aneurysm embolization coils. *Jpn J Radiol* 2015; **33**: 710–16. doi: http://dx.doi.org/10.1007/s11604-015-0478-2

15. Axente M, Paidi A, Von Eyben R, Zeng C, Bani-Hashemi A, Krauss A, et al. Clinical evaluation of the iterative metal artifact reduction algorithm for CT simulation in radiotherapy. *Med Phys* 2015; **42**: 1170–83. doi: http://dx.doi.org/10.1118/1.4906245

16. Kidoh M, Nakaura T, Nakamura S, Tokuyasu S, Osakabe H, Harada K, et al. Reduction of dental metallic artefacts in CT: value of a newly developed algorithm for metal artefact reduction (O-MAR). *Clin Radiol* 2014; **69**: e11–16. doi: http://dx.doi.org/10.1016/j.crad.2013.08.008

17. Funama Y, Taguchi K, Utsunomiya D, Oda S, Hirata K, Yuki H, et al. A newly-developed metal artefact reduction algorithm improves the visibility of oral cavity lesions on 320-MDCT volume scans. *Phys Med* 2015; **31**: 66–71. doi: http://dx.doi.org/10.1016/j.ejmp.2014.10.003

18. Pessis E, Campagna R, Sverzut JM, Bach F, Rodallec M, Guerini H, et al. Virtual monochromatic spectral imaging with fast kilovoltage switching: reduction of metal artefacts at CT. *Radiographics* 2013; **33**: 573–83. doi: http://dx.doi.org/10.1148/rg.332125124

19. Lewis M, Reid K, Toms AP. Reducing the effects of metal artefact using high keV monoenergetic reconstruction of dual energy CT (DECT) in hip replacements. *Skeletal Radiol* 2013; **42**: 275–82. doi: http://dx.doi.org/10.1007/s00256-012-1458-6

20. Meinel FG, Bischoff B, Zhang Q, Bamberg F, Reiser MF, Johnson TR. Metal artifact reduction by dual-energy computed tomography using energetic extrapolation: a systematically optimized protocol. *Invest Radiol* 2012; **47**: 406–14. doi: http://dx.doi.org/10.1097/RLI.0b013e31824c86a3

21. Bamberg F, Dierks A, Nikolaou K, Reiser MF, Becker CR, Johnson TR. Metal artefact reduction by dual energy computed tomography using monoenergetic extrapolation. *Eur Radiol* 2011; **21**: 1424–9. doi: http://dx.doi.org/10.1007/s00330-011-2062-1

22. Zhou C, Zhao Y, Luo S, Shi H, Li L, Zheng L, et al. Monoenergetic imaging of dual-energy ct reduces artifacts from implanted metal orthopedic devices in patients with factures. *Acad Radiol* 2011; **18**: 1252–7. doi: http://dx.doi.org/10.1016/j.acra.2011.05.009

23. Mangold S, Gatidis S, Luz O, König B, Schabel C, Bongers MN, et al. Single-source dual-energy computed tomography: use of monoenergetic extrapolation for a reduction of metal artefacts. *Invest Radiol* 2014; **49**: 788–93. doi: http://dx.doi.org/10.1097/RLI.0000000000000083

24. Bongartz G, Golding SJ, Jurik AG, Leonardi M, van Persijn van Meerten E, Rodríguez R, et al. European guidelines for multislice computed tomography. European Commission; 2014 [Cited 9 May 2016.] Available from: http://www.msct.eu/PDF_FILES/Part_two_final_document_quality_criteria_CT-TIP.pdf

25. Smedby Ö, Fredrikson F. Visual grading regression: analysing data from visual grading with regression models. *Br J Radiol* 2010; **83**: 767–75. doi: http://dx.doi.org/10.1259/bjr/35254923

26. Futing Liao T. *Interpreting probability models no. 101*. Los Angeles, CA: SAGE Publications Inc.; 1994.

27. Freedman DA. On the so-called "Huber sandwich estimator" and "Robust standard errors". *Am Statistician* 2006; **60**: 299–302. doi: http://dx.doi.org/10.1198/000313006X152207

28. Sahinler S, Topuz D. Bootstrap and jackknife resampling algorithms for estimation of regression parameters. *J Appl Quant Methods* 2007; **2**: 188–99.

29. Holm S. A simple sequential rejective multiple test procedure. *Scand J Stat* 1979; **6**: 65–70.