

Fully automated quantification method (FQM) of coronary calcium in an anthropomorphic phantom

Gijs D. van Praagh^{a)}

Department of Nuclear Medicine and Molecular Imaging, Medical Imaging Center, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands

Department of Radiology, Stanford University School of Medicine, Stanford, CA, USA

Niels R. van der Werf

Department of Radiology, University Medical Center Utrecht, Utrecht, The Netherlands

Department of Radiology and Nuclear Medicine, Erasmus University Medical Center, Rotterdam, The Netherlands

Jia Wang

Department of Environmental Health and Safety, Stanford University, Stanford, CA, USA

Fasco van Ommen

Department of Radiology, University Medical Center Utrecht, Utrecht, The Netherlands

Keris Poelhekkem

Department of Radiology, Medical Imaging Center, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands

Riemer H.J.A. Slart

Department of Nuclear Medicine and Molecular Imaging, Medical Imaging Center, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands

Department of Biomedical Photonic Imaging, Faculty of Science and Technology, University of Twente, Enschede, The Netherlands

Dominik Fleischmann

Department of Radiology, Stanford University School of Medicine, Stanford, CA, USA

Marcel J.W. Greuter

Department of Radiology, Medical Imaging Center, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands

Department of Robotics and Mechatronics, University of Twente, Enschede, The Netherlands

Tim Leiner

Department of Radiology, University Medical Center Utrecht, Utrecht, The Netherlands

Martin J. Willeminck

Department of Radiology, Stanford University School of Medicine, Stanford, CA, USA

(Received 12 September 2020; revised 19 February 2021; accepted for publication 15 April 2021; published 17 May 2021)

Objective: Coronary artery calcium (CAC) score is a strong predictor for future adverse cardiovascular events. Anthropomorphic phantoms are often used for CAC studies on computed tomography (CT) to allow for evaluation or variation of scanning or reconstruction parameters within or across scanners against a reference standard. This often results in large number of datasets. Manual assessment of these large datasets is time consuming and cumbersome. Therefore, this study aimed to develop and validate a fully automated, open-source quantification method (FQM) for coronary calcium in a standardized phantom.

Materials and Methods: A standard, commercially available anthropomorphic thorax phantom was used with an insert containing nine calcifications with different sizes and densities. To simulate two different patient sizes, an extension ring was used. Image data were acquired with four state-of-the-art CT systems using routine CAC scoring acquisition protocols. For interscan variability, each acquisition was repeated five times with small translations and/or rotations. Vendor-specific CAC scores (Agatston, volume, and mass) were calculated as reference scores using vendor-specific software. Both the international standard CAC quantification methods as well as vendor-specific adjustments were implemented in FQM. Reference and FQM scores were compared using Bland-Altman analysis, intraclass correlation coefficients, risk reclassifications, and Cohen's kappa. Also, robustness of FQM was assessed using varied acquisitions and reconstruction settings and validation on a dynamic phantom. Further, image quality metrics were implemented: noise power spectrum, task transfer function, and contrast- and signal-to-noise ratio among others. Results were validated using imQuest software.

Results: Three parameters in CAC scoring methods varied among the different vendor-specific software packages: the Hounsfield unit (HU) threshold, the minimum area used to designate a group of voxels as calcium, and the usage of isotropic voxels for the volume score. The FQM was in high agreement with vendor-specific scores and ICC's (median [95% CI] were excellent (1.000 [0.999-1.000] to 1.000 [1.000-1.000]). An excellent interplatform reliability of $\kappa = 0.969$ and $\kappa = 0.973$ was found. TTF results gave a maximum deviation of 3.8% and NPS results were comparable to imQuest. **Conclusions:** We developed a fully automated, open-source, robust method to quantify CAC on CT scans in a commercially available phantom. Also, the automated algorithm contains image quality assessment for fast comparison of differences in acquisition and reconstruction parameters. © 2021 The Authors. *Medical Physics* published by Wiley Periodicals LLC on behalf of American Association of Physicists in Medicine [https://doi.org/10.1002/mp.14912]

Key words: agatston scores, automated scoring, computed tomography, coronary calcium scores

1. INTRODUCTION

Coronary artery calcium (CAC) score is a strong predictive value for future adverse cardiovascular events, including myocardial infarction and sudden cardiac death, and a powerful tool in primary prevention.¹⁻³ In 1990, Agatston and colleagues developed a specific quantification method for CAC using electron beam tomography (EBT).⁴ This so-called Agatston score—currently quantified using cardiac computed tomography (CT)—is clinically used for further risk classification of asymptomatic individuals at intermediate risk.⁵⁻⁷ In addition to the Agatston score, two other metrics were introduced to quantify CAC, namely the volume and mass score.^{8,9}

It is well known that CAC scores vary between different CT scanners. Not only do CAC scores differ between scanners of different vendors but also between different systems from the same vendor, and between the same systems from the same vendor if a slightly different starting position is applied.¹⁰⁻¹² Moreover, CAC scores can vary greatly due to motion of the coronary arteries during the scan phase of a CAC scoring CT acquisition.¹³ In order to study these differences, their possible impact on clinical outcome, and to optimize acquisition protocols, dedicated coronary calcium phantoms are frequently used. In the well-established international standard developed for CAC quantification by McCollough and colleagues, a commonly evaluated commercially available anthropomorphic phantom was used (thorax and CCI phantom, QRM, Möhrendorf, Germany).⁸ With this phantom not only the Agatston score but also the volume and mass score of the calcifications in the phantom can be studied among different scanners and different vendors for influences of acquisition and reconstruction parameters. This phantom also contains calibration rods, allowing for adequate mass score assessment.

However, manual assessment of the CAC scores is time consuming and cumbersome, especially when several scan and/or reconstruction parameters have been systematically varied resulting in a large number of scans. Therefore, the aim of our study was to develop and validate a fully automated quantification method (FQM) for coronary calcium in a standardized phantom. In order to be useful for a variety of

CT scanners of different vendors, we sought to develop an automated scoring method that replicates CAC scores.

2. MATERIALS AND METHODS

2.A. Phantom

We used a standard, commercially available anthropomorphic thorax phantom (QRM-Thorax, QRM, Möhrendorf, Germany) (Fig. 1a). The static phantom comprises artificial lungs, a spine, and a shell of soft tissue-equivalent material. X-ray attenuation of the phantom's materials is similar to human tissues when data are acquired at a peak tube potential of 120 kVp. To simulate two different patient sizes, an extension ring (QRM-Extension ring, QRM, Möhrendorf, Germany) of fat equivalent material (−100 Hounsfield Units (HU)) was used. With this extension ring, outer dimensions of the phantom increased from 300 × 200 mm to 400 × 300 mm, similar to a small and large patient size, respectively.¹⁴ Within the thorax, a commercially available calcium-containing insert (Cardiac Calcification Insert (CCI), QRM, Möhrendorf, Germany) was placed, which is commonly used in coronary calcium studies (Fig. 1b).^{8,15-20} The insert consisted of nine hydroxyapatite (HA)-containing calcifications and two large calibration rods. These calibration rods consisted of water-equivalent material and 200 mgHAcm⁻³. The calcifications had diameters and lengths of 1.0, 3.0, and 5.0 mm, defined as small, medium, and large, respectively. For each calcification size, three densities were present in the phantom: 200, 400, and 800 mgHAcm⁻³, defined as low, medium, and high density, respectively.

To assess the performance of our automatic scoring method on dynamic data, a robotic arm (QRM Sim2D, QRM, Möhrendorf, Germany) moved an artificial coronary artery in a water-filled compartment, which was positioned in the center of the anthropomorphic thorax phantom. Two artificial arteries were used, where each artery consisted of two calcifications. These calcifications were equal in dimensions (5.0 ± 0.1 mm in diameter, with a length of 10.0 ± 0.1 mm) but different in density. Densities were 196 ± 3, 380 ± 2, 408 ± 2, and 800 ± 2 mgHAcm⁻³. The arteries were moved at four constant velocities (0-30 mm/s,

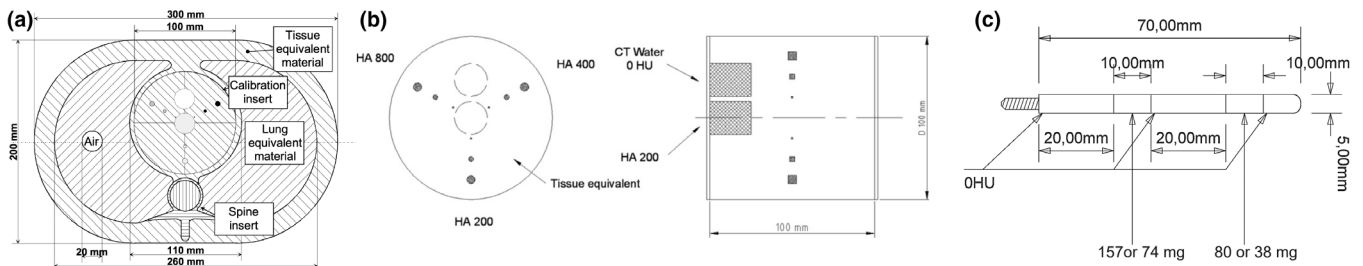


FIG. 1. (a) Axial sketch of the thoracic phantom including the cardiac calcification insert. (b) Axial and lateral sketch of the cardiac calcification insert containing the nine calcifications and the two calibration rods. (c) Sketch of the cylindrical artificial coronary artery containing two calcified inserts with a diameter of 5.0 ± 0.1 mm and a length of 10 ± 0.1 mm.

increment of 10 mm/s) along the x-axis, comparable to heart rates of 0, <60, 60-75, and >75 bpm.²¹ Electrocardiography trigger output was used to ensure that acquisition was done during linear motion of the calcifications.¹³

2.B. Acquisition and reconstruction

Static phantom image data were acquired with four state-of-the-art CT systems, one from each of the main CT manufacturers: CT-1: Aquilion One Vision (Canon Medical Systems, Otawara, Japan); CT-2: Brilliance iCT (Philips Healthcare, Best, The Netherlands); CT-3: Revolution CT (GE Healthcare, Waukesha, Wisconsin, USA); and CT-4: SOMATOM Force (Siemens Healthineers, Erlangen, Germany), respectively. Routine CAC scoring acquisition protocols for small and large patients were used (Table I). To simulate interscan variability each acquisition of the thorax phantom with and without extension ring was done five times with small translations and/or rotations of approximately

2 mm and 2 degrees, respectively. Raw data were reconstructed with filtered back projection (FBP) (Table I).

2.C. Vendor-specific CAC scores

For all acquisitions, vendor-specific CAC scores were derived using each vendor’s commercial software implementation (Table I). These CAC scores included Agatston, volume, and mass scores. For each vendor, CAC scores derived with their respective software were used as reference CAC scores for the analysis. The CT specific mass calibration factor was determined for each CT system according to standard methodology.⁸

2.D. CAC score standard: Automated algorithms

The international standard for quantification of CAC scores was implemented in a fully automated algorithm (FQM) for CAC scoring of the CCI phantom. This was done

TABLE I. Acquisition and reconstruction parameters for all CT systems used in this study.

Parameter	CT1	CT2	CT3	CT4	Dynamic
Manufacturer	Canon	Philips	GE	Siemens	Siemens
CT system	Aquilion One Vision	Brilliance iCT	Revolution	SOMATOM Force	SOMATOM Flash
Acquisition mode	Axial	Axial	Axial	Axial	Axial
Tube voltage [kVp]	120	120	120	120	120
Tube current time product [mAs]	Small: 15 Large: 84	Small: 50 Large: 50	Small: 30 Large: 161	Small: 44 Large: 194	80
Automatic exposure correction	SD=55	Off	Off	Off	Off
CTDI _{vol} [mGy]	Small: 2.3 Large: 12.8	Small: 4.7 Large: 4.4	Small: 1.49 Large: 7.2	Small: 1.5 Large: 6.7	Large: 2.8
Collimation [mm]	280x0.5	128x0.625	224x0.625	160x0.6	128x0.6
Field of View [mm]	250	250	250	250	250
Rotation time [s]	0.35	0.27	0.28	0.25	0.28
Slice thickness [mm]	3.0	3.0	2.5	3.0	3.0
Increment [mm]	3.0	3.0	2.5	3.0	3.0
Reconstruction kernel	FC12	XCA	Standard	Qr36d	B35f*
Matrix size [pixels]	512 × 512	512 × 512	512 × 512	512 × 512	512 × 512
Reconstruction	FBP	FBP	FBP	FBP	FBP
Calcium scoring software	Vitrea FX 6.5.0 (S1)	Heartbeat-CS (S2)	SmartScore 4.0 (S3)	Syngo Calcium Scoring (S4)	Syngo Calcium Scoring (S4)

*Based on vendor-recommended protocol of earlier software version than used for the static phantom.

in two popular programming languages to allow for wide usage: MATLAB® R2020a (Mathworks, Natick, Massachusetts, USA) and Python (Python 3.7.3). Both algorithms were made publicly available via Github (https://github.com/nwerf/FQM_Analysis) to assist in any research where the CCI insert is used.

After importing a DICOM series into FQM (module 1), the center of the insert (module 2) and two main locations in the CCI were found: the largest calcifications (module 3) and the 200 mg HA calibration rod (module 4; Figure 2). These calcified areas were found using a connected component analysis (4-connected) with the standard CAC scoring threshold of 130 HU.⁴ Next, a mask based on the locations of the nine calcifications was

determined. First, the largest calcifications were defined based on the area of the connected components. For each density, the locations of the other calcifications were determined using the known distances between the calcifications of different sizes, on the connecting lines between the center of the insert and the center of the large calcification. The mean HU value of each of the large calcifications was used to determine the density of the calcifications, with the highest mean HU value corresponding to the highest density, etc. By using this methodology, the exact position of the phantom within the CT system, and any rotation of the CCI insert within the thorax phantom, was made irrelevant, consequently adding to the robustness of FQM.

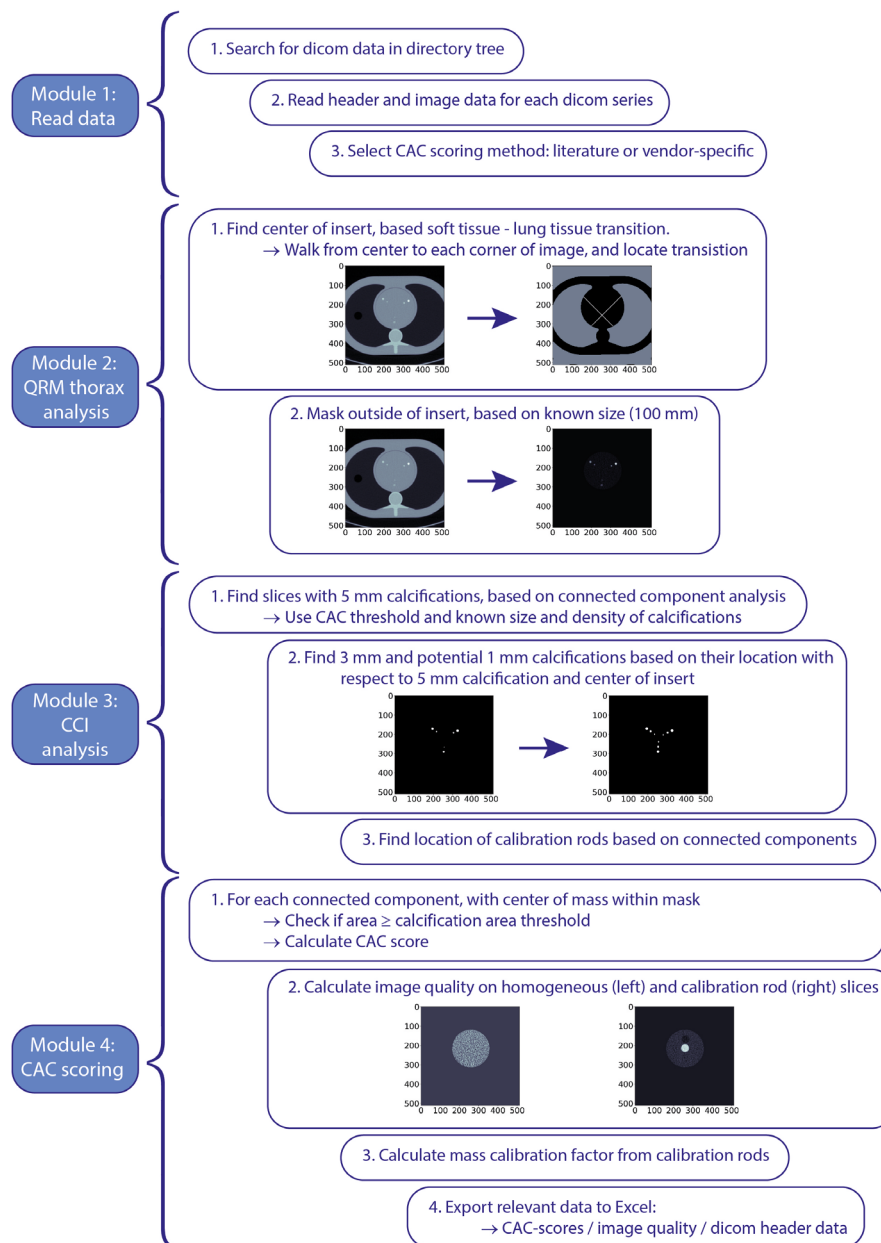


Fig. 2. Flowchart of FQM. [Color figure can be viewed at wileyonlinelibrary.com]

The international standard implementation for all three CAC scoring methods (Agatston, volume, and mass scores) were in accordance with their respective definitions from literature.^{4,8,9} All methods used a minimum in-plane area of 1 mm² for pixels >130 HU to identify calcium-containing lesions. The Agatston scores were derived for each calcified area per slice from a multiplication of that area with an associated weighting factor depending on the maximum HU within the area: 130 to 200 HU = 1; 200 to 300 HU = 2; 300 to 400 HU = 3; and ≥ 400 HU = 4. The Agatston score per calcification was defined as the summation of all Agatston scores per slice.

The volume score was determined according to Callister et al. based on a linear interpolation to create isotropic voxels.⁹ To achieve this, the slice thickness was decreased to match in-plane pixel spacing by means of a linear grid interpolation in 3D. To limit computation time, this was only performed for the slices containing the calcifications. For each slice, the volume score was calculated by multiplication of the number of voxels per lesion with the interpolated voxel volume.

Lastly, mass scores were determined according to McColough et al. using scan-specific mass calibration factors.⁸ Mean CT numbers (HU) for the calibration factor calculation were measured in the center slice of the large cylinder-shaped calibration rods with a region of interest of 1.5 cm². The calibration rods were automatically located based on the known specifications of the phantom. Then, mean CT numbers (HU) for both calibration rods were used to calculate the scan-specific calibration factor. Finally, mass scores of the calcifications were calculated by multiplication of the calibration factor with the calcified volume (without interpolation) and the mean CT number of the lesion.

To assess robustness of FQM, additional acquisitions with varying acquisition settings were made on CT-4 and scored with its vendor-specific software. In these acquisitions, parameters that have a well-known influence on CAC scores were changed: tube potential was changed from 120 to 100 and 80 kVp, tube current time product was changed from 44 to 34 and 22 mAs, convolution kernel was changed from Qr36 to Qr32 and Qr44, iterative reconstruction was applied at levels 2 and 4, and lastly, field-of-view was changed from 250 to 200 and 320 mm. Finally, robustness was assessed for a dynamic phantom on another CT system: SOMATOM Definition Flash (Siemens Healthineers, Erlangen, Germany). A routinely used clinical CT CAC protocol was used for acquisition and reconstruction (Table I; Fig. 3).

2.E. Vendor-specific CAC scores: Automated algorithms

In addition, FQM was adapted in such a way that the calculation of the calcium scores matched the methodology used in the vendor-specific software packages. These adjustments were based on scoring mechanism descriptions in manuals, and information provided by the vendors. The following

parameters were adapted: HU threshold used to designate a pixel as calcium, the threshold used to indicate the minimum area necessary for calcium scoring, and the use of interpolation for specific CAC scores (Table II). Vendor-specific parameters were automatically extracted by FQM from the DICOM header information, which also identified the vendor-specific CT system that was used to acquire the data. In addition, the algorithms allowed for manual selection of vendor-specific scoring parameters. With this, images from any of the four vendors can be evaluated with scoring parameters from any other vendor.

2.F. Image quality assessment

For the automated analysis of the CCI phantom, several image quality metrics were included to assess image quality differences for changing acquisition or reconstruction parameters. These image quality metrics both concerned image noise and contrast measurement. For the image noise, first the standard deviation (SD) of the mean CT value (HU) of a square region of interest (ROI) of 55 × 55 mm in a noncalcium-containing slice of the CCI insert was calculated. Second, image noise was characterized with a noise power spectrum (NPS) analysis. This analysis was implemented according to the methodology of the International Commission on Radiation Units and Measurements (ICRU), as previously implemented by Van Ommen et al.^{22,23} For this, 18 radially dispersed ROIs of 15 × 15 mm were used. Both 2D and 1D NPS results were extracted.

For the contrast-related image quality metrics, first the mean HU and SD of the three large calcifications and two calibration rods were calculated. For each calcification, the mean HU was calculated over the entire volume of each calcification. The mean HU and SD of a circular ROI of 1.5 cm² in the calibration rod were calculated within the center slice of these rods. Second, the signal-to-noise ratio (SNR) and contrast-to-noise ratio (CNR) were calculated for the calcifications.

Lastly, the task-transfer function (TTF) was computed. The TTF is a type of modulation-transfer function, which is also valid for nonlinear systems and incorporates contrast and noise.²⁴ For this, the ICRU implementation for modulation transfer function calculation was used.²² For robustness, the TTF was calculated by radially averaging the edge spread function (ESF) of the calibration rod, as described previously by Van Ommen et al.²³ Due to the proximity of the water-equivalent calibration rod, the ESF in the direction of this rod was excluded from the analysis. In addition, image data were linearly interpolated by a factor 4, to reduce pixel-size effects. For quick evaluation purposes, 50% and 10% TTF were also calculated.

NPS and TTF results were validated by comparison with the CT image analysis tool (imQuest (Duke University, Durham, 2018)) described in Task Group 233 of the American Association of Physicists in Medicine (AAPM) for two datasets, reconstructed with different reconstruction kernels (Qr44, Qr32). For the NPS calculation, only one ROI was

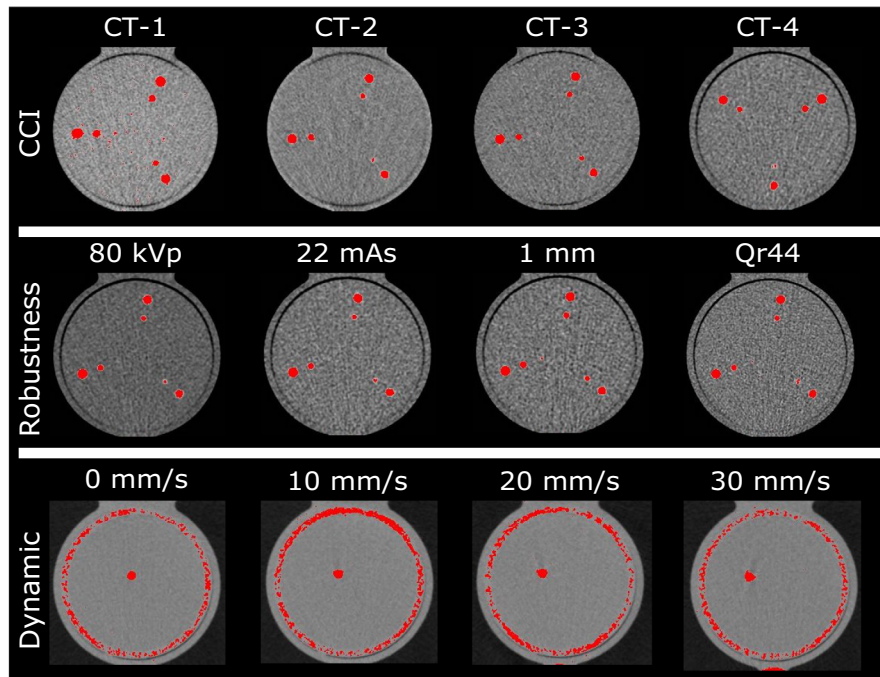


FIG. 3. Axial views of the cardiac calcification insert from the four CT systems used in static experiments (top row), a few examples of the robustness scans where acquisition or reconstruction settings were changed (middle row; from left to right: tube voltage, tube current, slice thickness, and kernel), and the dynamic phantom with four different speed settings (bottom row). Red overlay is used to highlight the pixels above the 130 HU threshold. Screenshots are made with ImageJ (U.S. National Institutes of Health, Bethesda, Maryland, USA). [Color figure can be viewed at wileyonlinelibrary.com]

TABLE II. International standard and vendor-specific CAC scoring parameters for all vendors and commercial vendor neutral software. Light-grey entries indicate equal parameter values with respect to literature. Darker-grey entries are vendor-specific parameters.

Parameters	CAC score	International standard	S1	S2	S3	S4
Connectivity	All	4	4	4	4	4
HU threshold	Agatston	130	130	130	130	130
	Volume	130	130	130 or 100/c ^a	Patented	130
	Mass	130	130	100/c	Patented	130
Calcification area threshold	All	1 mm ²	3 pixels	0.5 mm ²	1 mm ²	0
	Interpolation ^b	Volume	Yes	No	No	Patented

c = calibration factor.

^aDepending on availability of CT system-specific calibration factor within the scoring software

^bLinear interpolation algorithm used to calculate isotropic voxels

placed at the center of the insert for both tools for the current comparison due to potential measurement errors resulting from manual placement of 18 ROIs for imQuest.

2.G. Statistical analysis

To assess the accuracy of our FQM, automatically quantified CAC scores were compared with reference scores obtained with vendor-specific software. Agreement between

FQM and reference CAC scores was assessed using Bland-Altman analyses. Reliability between the methods was determined by calculating intraclass correlation coefficients (ICCs) and root mean square error (RMSE). Reference and FQM scores were classified per calcification according to the Agatston risk stratification: 0 – absent; >0 and <10 – minimal; ≥10 and <100 – mild; ≥100 and <400 – moderate; and ≥400 – severe. Calcifications classified differently by FQM from the reference classifications were defined as reclassifications. Subsequently, reliability of reclassification between FQM and reference scores was determined by calculating Cohen’s kappa (κ). All statistical analyses were performed with SPSS for Windows, version 26.0. A *P*-value <0.05 was used to determine significant differences.

3. RESULTS

3.A. Vendor-specific CAC scores: automated algorithms

Vendor-specific adjustments to our generic CAC scoring methods were necessary to match vendor-specific scores. An overview of all parameters, including vendor-specific parameters, is shown in Table II. Three parameters varied among the different vendor-specific software packages. First, the HU threshold, used to indicate whether a pixel contains CAC, varied. In general, a threshold of 130 HU was used for all vendors, for all CAC scores. However, for one vendor, the threshold was 100 mg HA, when a CT system-specific calibration factor was available in the software. When this

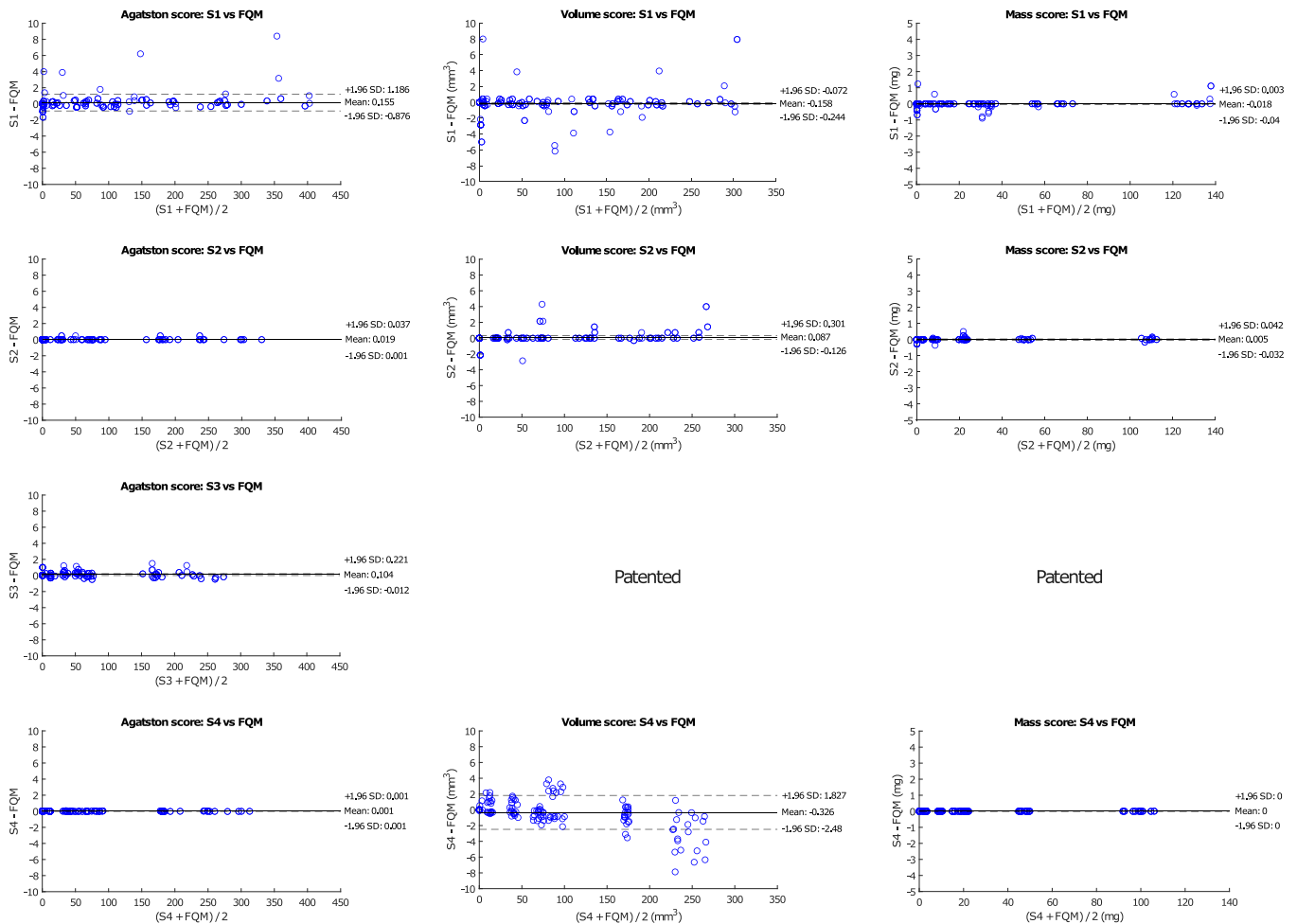


FIG. 4. Bland-Altman plots of all CAC scoring software compared to the FQM. From left to right the Agatston, volume, and mass scores are shown, respectively. From top to bottom, S1 to S4 are shown. Volume and mass scoring method of S3 were patented (the manufacturer was not able to provide any information) and could, therefore, not be implemented into the FQM. [Color figure can be viewed at wileyonlinelibrary.com]

calibration factor was not available, the normal threshold of 130 HU was used. Second, the minimum area used to designate a group of pixels as calcium varied. For a group of pixels with HU above the CAC scoring threshold, this minimum area varied between >0 pixels and 1 mm^2 . Last, some vendors used an interpolation algorithm to create isotropic voxels for the volume score, and some vendors did not. Parameters for the volume and mass score of CT-3 were kept confidential by the vendor and could, therefore, not be determined.

With these vendor-specific CAC scoring parameters implemented, FQM scores were in high agreement with the vendor-specific software scores for all CAC scoring methods (Fig. 4). Smallest confidence interval (CI) (95%) range of absolute differences between the FQM and vendor-specific scores was 0.000 to 0.000 mg for the mass score when FQM was compared to S4. Largest CI range was -2.480 to 1.827 mm^3 for the volume score when FQM was compared to S4. ICCs were excellent for all comparisons between FQM and the vendor-specific software. The ICC of the volume score of S4 and FQM was 1.000 (0.999-1.000); all other

comparisons gave an ICC of 1.000 (1.000-1.000). RMSE for Agatston, volume, and mass score ranged between 0.02 and 1.01, 0.80 and 1.64 mm^3 , and 0.00 and 0.22 mg, respectively. Reclassification of the calcifications occurred 7 times of 90 calcifications (7.8%) at CT-1 and 3 times of 90 calcifications (3.3%) at CT-3. All reclassifications were from zero to minimal or vice versa. No reclassifications occurred with CT-2 and CT-4. This gave an interplatform reliability of $\kappa = 0.969$ ($P < 0.0001$), 95% CI [0.947, 0.991], between $\text{FQM}_{\text{MATLAB}}$ and the vendor-specific software and $\kappa = 0.973$ ($P < 0.0001$), 95% CI [0.953, 0.993], between $\text{FQM}_{\text{Python}}$ and the vendor-specific software.

3.B. Algorithm robustness

FQM scores were also in high agreement with the vendor-specific software packages after varying the acquisition settings for all CAC scores. When FQM scores were compared with the vendor-specific scores, mean (95% CI) differences for Agatston, volume, and mass scores were -0.001 (-0.033 to 0.031), -0.2 (-0.365 to -0.035) mm^3 , and -0.071

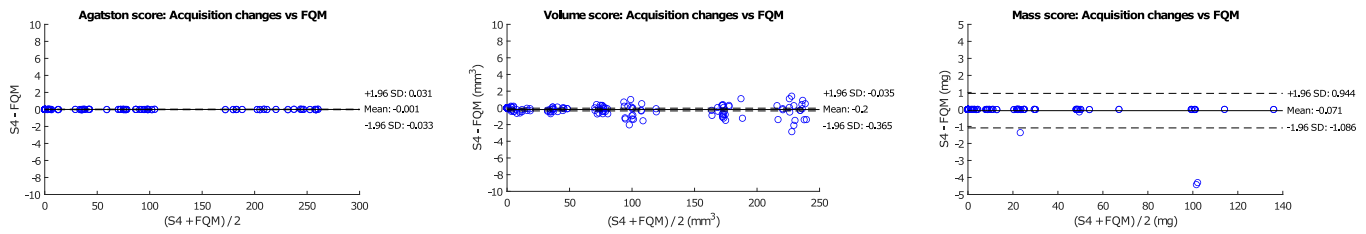


FIG. 5. Bland-Altman plots of S4 compared to the FQM of all CAC scores. Acquisition parameters were changed for assessment of algorithm robustness. [Color figure can be viewed at wileyonlinelibrary.com]

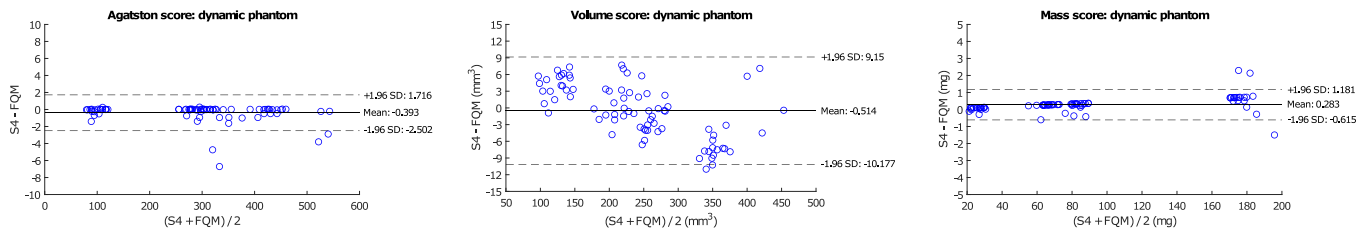


FIG. 6. Bland-Altman plots of S4 compared to the FQM of all CAC scores. Scans were acquired with a dynamic phantom and scored with FQM for assessment of algorithm robustness. [Color figure can be viewed at wileyonlinelibrary.com]

(−1.086 to 0.944) mg HA, respectively (Fig. 5). ICCs (mean [95% CI]) were excellent (1.000 [1.000–1.000] for all CAC scores). No reclassifications occurred. RMSE were between 0.012 and 0.020 for Agatston scores, 0.220 and 0.835 mm³ for volume scores, and 0.003 and 1.063 mg for mass scores. Remarkably, all RMSEs of mass scores were below 0.034 mg except for field-of-view changes, where RMSE scores were 1.042 and 1.063 mg for FOV 320 and 200, respectively.

For the dynamic phantom, FQM scores were in high agreement with vendor-specific software too. When FQM scores were compared with the vendor-specific scores, mean (95% CI) differences for Agatston, volume, and mass scores were −0.393 (−2.502 to 1.716), −0.514 (−10.177 to 9.15) mm³, and −0.283 (−0.651 to 1.181) mg HA, respectively (Fig. 6). ICCs (mean [95% CI]) were excellent (1.000 [1.000–1.000] for Agatston and mass scores and 0.999 [0.999–1.000] for volume scores). RMSE was 1.139, 4.926 mm³, and 0.536 mg for Agatston, volume, and mass scores, respectively.

On a regular desktop computer (Windows 7, i5-6500 CPU 3.2 GHz, 8 GB RAM), evaluating a single scan with FQM took on average 3 or 6 seconds without or with interpolation for the volume score, respectively. In contrast, manual analysis of the phantom (without advanced image quality assessment) is in the order of minutes.

3.C. Image quality

For two datasets which were reconstructed with different reconstruction kernels (Qr44, Qr32), 10% and 50% TTF results were calculated (Fig. 7). For the NPS analysis, images, ROI placement, and resulting 1D NPS curve results are shown in Figure 8. For both reconstruction kernels, NPS results were comparable between FQM and imQuest.

4. DISCUSSION

In this study, we successfully developed an open-source, fully automated, vendor-independent, robust method to quantify CAC in two commonly used commercially available phantoms. In addition, we implemented vendor-specific scoring methods from four major calcium scoring software vendors with excellent agreement. Two scoring methods could not be implemented in our method due to non-disclosures. Also, image quality metrics, useful for comparison of CT scans with varying imaging parameters, were automatically extracted from the image data. These advanced image quality metrics can aid in assessing the influence of nonlinear (post) processing steps on CAC scores.

Our algorithm is focused on a fully automated analysis of a standard anthropomorphic cardiac phantom. The main reasons for this focus are the substantial reduction of evaluation time and the lack of inter- and intraobserver variability, manual notation errors, and software problems because of acquisition settings. An example of the latter is that some software programs are not able to process calcium scoring scans with a slice thickness different from the usual 3 mm, which can be rather inconvenient for research purposes. This in contrast to FQM, which is written in both MATLAB and Python, making it widely usable, depending on programming language preference. This phantom is often used for careful evaluation of novel technical advances in CT, eg, acquisition techniques, such as novel photon-counting detector elements, or reconstruction techniques, such as kernels which allow for tube voltage-independent CAC acquisitions, before clinical usage.^{16,25} FQM can aid in these experiments, as larger number of scans can easily be analyzed in a fully automatic manner.

Although the predictive role in risk stratification of low nonzero calcium scores caused by microcalcifications is still unknown, zero CAC scores are proven to be a strong negative

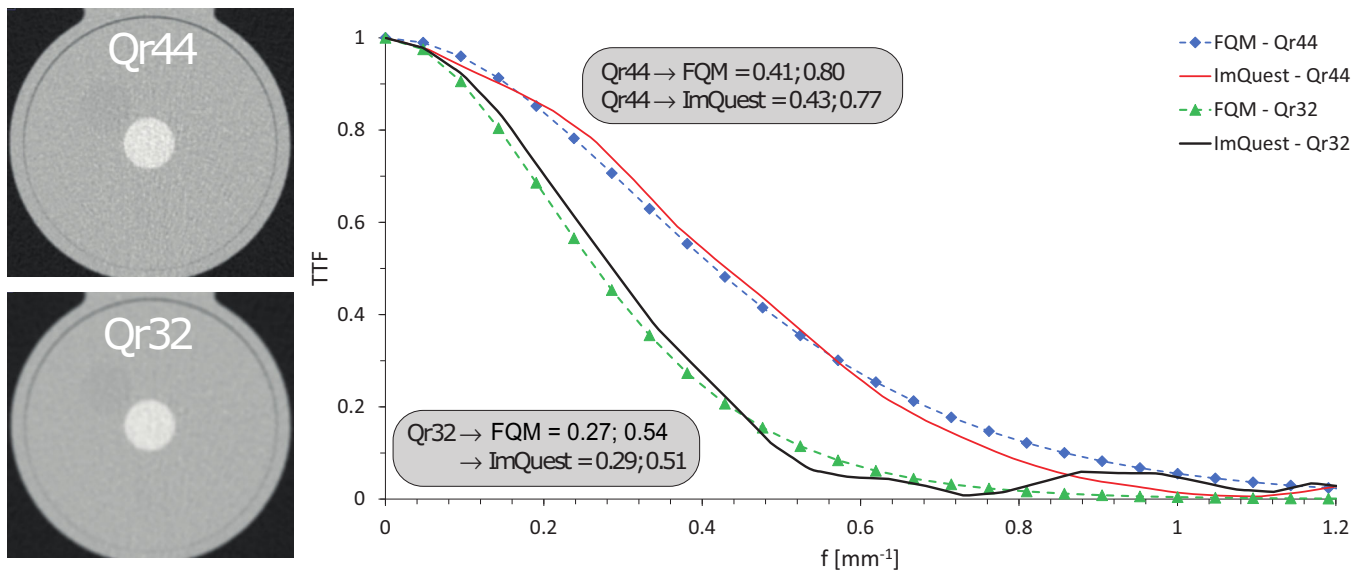


FIG. 7. TTF results for both FQM and ImQuest for two datasets, reconstructed with different reconstruction kernels. In addition, deviations at 50% and 10% TTF between results from both analyses are shown. [Color figure can be viewed at wileyonlinelibrary.com]

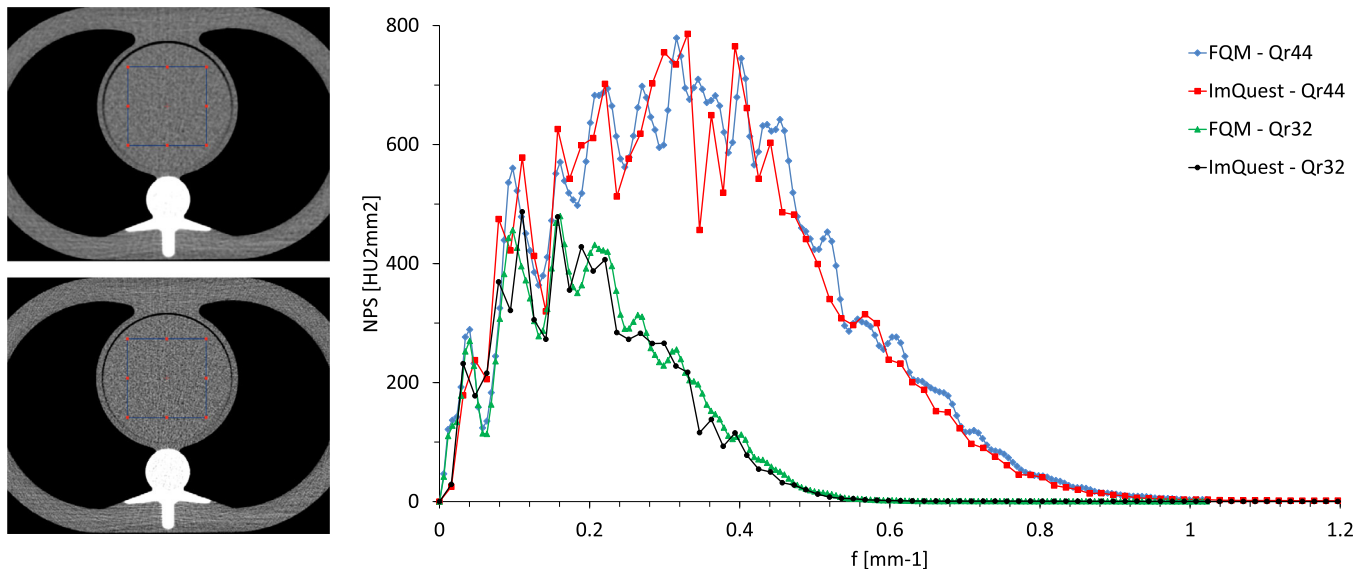


FIG. 8. NPS results for both FQM and ImQuest. Left, images for the Qr32 (upper) and Qr44 (lower) reconstruction kernel are shown, together with the placed ROI. Right, resulting 1D NPS results are shown. Small differences between both results are expected to be due to small differences in ROI placement. [Color figure can be viewed at wileyonlinelibrary.com]

predictor of CAD.^{26–28} Also, Criqui and colleagues found an inversely proportional association of density on future cardiovascular events.²⁹ Therefore, the detection of small and low-density calcifications is of utmost importance. In our study, we found three main software parameters, which influence CAC detection and, therefore, quantification. First, the threshold used to discriminate calcium-containing voxels from noncalcium-containing voxels. Second, the minimum calcification area threshold used to discriminate between noise and calcium-containing voxels. And third, the use of isotropic interpolation for volume scores. All factors have an important impact on the detection of microcalcifications, especially for high noise acquisitions. For these acquisitions,

lower thresholds and use of interpolation will increase CAC area, and smaller minimum calcification areas will increase the number of false positives due to noise effects. It is, thus, important to investigate the exact influence of these parameters on CAC scores and the impact of scoring method standardization on differences in CAC scores between scanners. Besides that, the need of an improved CAC scoring method is high.^{28,30,31} Both Agatston and volume scores show high variability in scores within and between CT systems.^{11,12} The mass score is a more reliable score in terms of variability, although small differences still exist.³² FQM is, thus, a helpful tool in the development of new CT acquisition/reconstruction protocols and new scoring methods.

A few studies developed an automatic CAC scoring algorithm for patient CT angiography scans.^{33–36} However, to the best of our knowledge, this is the first study that developed a fully automated, vendor-neutral method for quantification of CAC scores in a phantom. Also, no other study examined and reproduced the exact scoring methods of the four major calcium scoring software vendors. Only a few studies compared software platforms in CAC scores. However, these were either with platforms that are nowadays no longer widely used, or they compared scores, but did not go into detail about the parameters.^{37–39} Weininger and colleagues used three different workstations, Syngo Calcium Scoring (Siemens), Aquarius (TeraRecon), and Vitrea (Vital Images), to acquire CAC scores of 59 patients.³⁹ Total Agatston and volume scores were compared between these systems. Although all results were numerically different, they found excellent correlations between the three workstations for both scoring methods.³⁹

Our study has limitations. First, we were not able to implement the volume and mass quantification method of GE Healthcare. The vendor explained that they make use of a patented algorithm which adapts the threshold to help correct for beam hardening and overestimation. This adaptive threshold is used for both volume and mass scores. Another limitation of this study is that, currently, FQM can only be used in the described phantoms and not in patients or other phantoms as it makes use of the physical properties of these phantoms. However, these are commonly used phantoms for coronary calcium studies and FQM provides simple and fast analyses. Also, the main body of FQM can be rewritten to include other phantoms as we have shown in our flowchart and by validating both a static and a dynamic phantom. This increases the usability of FQM. Finally, only in-plane resolution measurements were added to the current version of FQM. Longitudinal measurements, based on the edge of the calibration rod, could be added in a future release.

5. CONCLUSIONS

In conclusion, we developed a fully automated, open-source, robust method in MATLAB and Python to quantify CAC in a commercially available and widely used phantom. The algorithm contains the international standard quantification methods described in literature, as well as almost all scoring methods of four major calcium scoring software vendors with an excellent agreement. The need for manual calcium scoring was completely eliminated with our fully automated method. Also, the automated algorithm contains image quality assessment for fast comparison of differences in acquisition and reconstruction parameters.

CONFLICT OF INTEREST

Gijs D van Praagh: This work was supported in part by an unconditional grant from PUSH: a collaboration between Siemens Healthineers and the University Medical Center Groningen. The sponsor had no role in the conceptualization,

interpretation of findings, writing or publication of the article. Niels R van der Werf: the author has no relevant conflicts of interest to disclose. Jia Wang: the author has no relevant conflicts of interest to disclose. Fasco van Ommen: the author has no relevant conflicts of interest to disclose. Keris Poelheken: the author has no relevant conflicts of interest to disclose. Riemer HJA Slart: the author has no relevant conflicts of interest to disclose. Dominik Fleischmann: the author has received research support from Siemens; the author has ownership interest in IschemaView Inc., and in Segmed Inc., none of which is related to cardiac CT or this project. Marcel JW Greuter: the author has no relevant conflicts of interest to disclose. Tim Leiner: the author has no relevant conflicts of interest to disclose. Martin J Willeminck: Activities related to the present article: Disclosed no relevant relationships. Activities not related to the present article: Received a research grant from Philips Healthcare. Co-founder, advisor, and stockholder of Segmed, Inc. Other relationships: Disclosed no relevant relationships.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Gijs D van Praagh and Niels R van der Werf should be considered joint first author

^{a)} Author to whom correspondence should be addressed. Electronic mail: g.d.-van.praagh@umcg.nl; Telephone: +31 50 361 70 47.

REFERENCES

1. Malguria N, Zimmerman S, Fishman EK. Coronary artery calcium scoring: Current status and review of literature. *J Comput Assist Tomogr.* 2018;42:887–897.
2. Van Der Bijl N, De Bruin PW, Geleijns J, et al. Assessment of coronary artery calcium by using volumetric 320-row multi-detector computed tomography: Comparison of 0.5 mm with 3.0 mm slice reconstructions. *Int J Cardiovasc Imaging.* 2010;26:473–482.
3. Keelan PC, Bielak LF, Ashai K, et al. Long-term prognostic value of coronary calcification detected by electron-beam computed tomography in patients undergoing coronary angiography. *Circulation.* 2001;104:412–417.
4. Agatston AS, Janowitz WR, Hildner FJ, Zusmer NR, Viamonte MJ, Detrano R. Quantification of coronary artery calcium using ultrafast computed tomography. *J Am Coll Cardiol.* 1990;15:827–832.
5. Divakaran S, Cheezum MK, Hulten EA, et al. Use of cardiac CT and calcium scoring for detecting coronary plaque: Implications on prognosis and patient management. *Br J Radiol.* 2015;88:20140594.
6. Hecht H, Blaha MJ, Berman DS, et al. Clinical indications for coronary artery calcium scoring in asymptomatic patients: Expert consensus statement from the Society of Cardiovascular Computed Tomography. *J Cardiovasc Comput Tomogr.* 2017;11:157–168.
7. Greenland P, Blaha MJ, Budoff MJ, Erbel R, Watson KE. Coronary calcium score and cardiovascular risk. *J Am Coll Cardiol.* 2018;72:434–447.
8. McCollough CH, Ulzheimer S, Halliburton SS, Shanneik K, White RD, Kalender WA. Coronary artery calcium: A multi-institutional, multimanufacturer international standard for quantification at cardiac CT. *Radiology.* 2007;243:527–538.
9. Callister TQ, Cooil B, Raya SP, Lippolis NJ, Russo DJ, Raggi P. Coronary artery disease: Improved reproducibility of calcium scoring with an electron-beam CT volumetric method. *Radiology.* 1998;208:807–814.

10. van der Werf NR, Willeminck MJ, Willems TP, Greuter MJW, Leiner T. Influence of dose reduction and iterative reconstruction on CT calcium scores: A multi-manufacturer dynamic phantom study. *Int J Cardiovasc Imaging*. 2017;33:899–914.
11. Willeminck MJ, Vliegenthart R, Takx RAP, et al. Coronary artery calcification scoring with state-of-the-art ct scanners from different vendors has substantial effect on risk classification. *Radiology*. 2014;273:695–702.
12. Rutten A, Isgum I, Prokop M. Coronary Calcification: Effect of Small Variation of Scan Starting Position on Agatston, Volume, and Mass Scores. *Radiology*. 2008;246(1):90–98. <https://doi.org/10.1148/radiol.2461070006>.
13. van der Werf NR, Willeminck MJ, Willems TP, Vliegenthart R, Greuter MJW, Leiner T. Influence of heart rate on coronary calcium scores: A multi-manufacturer phantom study. *Int J Cardiovasc Imaging*. 2018;34:959–966.
14. Willeminck MJ, Abramiuc B, den Harder AM, et al. Coronary calcium scores are systematically underestimated at a large chest size: A multivendor phantom study. *J Cardiovasc Comput Tomogr*. 2015;9:415–421.
15. McCollough CH, Primak AN, Saba O, et al. Dose performance of a 64-channel dual-source CT scanner. *Radiology*. 2007;243:775–784.
16. Booij R, van der Werf NR, Budde RPJ, Bos D, van Straten M. Dose reduction for CT coronary calcium scoring with a calcium-aware image reconstruction technique: A phantom study. *Eur Radiol*. 2020;30:3346–3355.
17. Tang YC, Liu YC, Hsu MY, Tsai HY, Chen CM. Adaptive iterative dose reduction 3D integrated with automatic tube current modulation for CT coronary artery calcium quantification: Comparison to traditional filtered back projection in an anthropomorphic phantom and patients. *Acad Radiol*. 2018;25:1010–1017.
18. Vonder M, Pelgrim GJ, Huijssse SEM, et al. Coronary artery calcium quantification on first, second and third generation dual source CT: A comparison study. *J Cardiovasc Comput Tomogr*. 2017;11:444–448.
19. Blobel J, Mews J, Goatman KA, Schuijff JD, Overlaet W. Calibration of coronary calcium scores determined using iterative image reconstruction (AIDR 3D) at 120, 100, and 80 kVp. *Med Phys*. 2016;43:1921–1932.
20. Schindler A, Vliegenthart R, Schoepf UJ, et al. Iterative image reconstruction techniques for CT coronary artery calcium quantification: Comparison with traditional filtered back projection in vitro and in vivo. *Radiology*. 2014;270:387–393.
21. Husmann L, Leschka S, Desbiolles L, et al. Coronary artery motion and cardiac phases: Dependency on heart rate - Implications for CT image reconstruction. *Radiology*. 2007;245:567–576.
22. The International Commission on Radiation Units and Measurements. ICRU Report no.87 - Radiation dose and image-quality assessment in computed tomography. *J ICRU*. 2012;12:1–149.
23. van Ommen F, Bennink E, Vlassenbroek A, et al. Image quality of conventional images of dual-layer SPECTRAL CT: A phantom study. *Med Phys*. 2018;45:3031–3042.
24. Robins M, Solomon J, Richards T, Samei E. 3D task-transfer function representation of the signal transfer properties of low-contrast lesions in FBP- and iterative-reconstructed CT. *Med Phys*. 2018;45:4977–4985.
25. Sandfort V, Persson M, Pourmorteza A, Noël PB, Fleischmann D, Willeminck MJ. Spectral photon-counting CT in cardiovascular imaging. *J Cardiovasc Comput Tomogr*. 2020; (In Press). <https://doi.org/10.1016/j.jcct.2020.12.005>
26. Blaha M, Budoff MJ, Shaw LJ, et al. Absence of coronary artery calcification and all-cause mortality. *JACC Cardiovasc Imaging*. 2009;2:692–700.
27. Sarwar A, Shaw LJ, Shapiro MD, et al. Diagnostic and prognostic value of absence of coronary artery calcification. *JACC Cardiovasc Imaging*. 2009;2:675–688.
28. Blaha MJ, Mortensen MB, Kianoush S, Tota-Maharaj R, Cainzos-Achirica M. Coronary artery calcium scoring: Is it time for a change in methodology? *JACC Cardiovasc Imaging*. 2017;10:923–937.
29. Criqui MH, Denenberg JO, Ix JH, et al. Calcium density of coronary artery plaque and risk of incident cardiovascular events. *JAMA*. 2014;311:271.
30. Willeminck MJ, van der Werf NR, Nieman K, Greuter MJW, Koweek LM, Fleischmann D. Coronary artery calcium: A technical argument for a new scoring method. *J Cardiovasc Comput Tomogr*. 2019;13:347–352.
31. Arnold BA, Budoff MJ, Child J, Xiang P, Mao SS. Coronary calcium test phantom containing true CaHA microspheres for evaluation of advanced CT calcium scoring methods. *J Cardiovasc Comput Tomogr*. 2010;4:322–329.
32. Dijkstra H, Greuter MJW, Groen JM, et al. Coronary calcium mass scores measured by identical 64-slice MDCT scanners are comparable: A cardiac phantom study. *Int J Cardiovasc Imaging*. 2010;26:89–98.
33. Yang G, Chen Y, Ning X, Sun Q, Shu H, Coatrieux JL. Automatic coronary calcium scoring using noncontrast and contrast CT images. *Med Phys*. 2016;43:2174–2186.
34. Lessmann N, Van Ginneken B, Zreik M, et al. Automatic calcium scoring in low-dose chest CT using deep neural networks with dilated convolutions. *IEEE Trans Med Imaging*. 2018;37:615–625.
35. de Vos BD, Wolterink JM, Leiner T, de Jong PA, Lessmann N, Isgum I. Direct automatic coronary calcium scoring in cardiac and chest CT. *IEEE Trans Med Imaging*. 2019;38:2127–2138.
36. Wolterink JM, Leiner T, de Vos BD, van Hamersvelt RW, Viergever MA, Isgum I. Automatic coronary artery calcium scoring in cardiac CT angiography using paired convolutional neural networks. *Med Image Anal*. 2016;34:123–136.
37. Yamamoto H, Budoff MJ, Lu B, Takasu J, Oudiz RJ, Mao S. Reproducibility of three different scoring systems for measurement of coronary calcium. *Int J Cardiovasc Imaging*. 2002;18:391–397.
38. Adamzik M, Schmermund A, Reed JE, Adamzik S, Behrenbeck T, Sheedy PF. Comparison of two different software systems for electron-beam CT- derived quantification of coronary calcification. *Invest Radiol*. 1999;34:767–773.
39. Weininger M, Ritz KS, Schoepf UJ, et al. Interplatform reproducibility of CT coronary calcium scoring software. *Radiology*. 2012;265:70–77.