European Radiology
https://doi.org/10.1007/s00330-018-5348-8

CARDIAC

# Automated estimation of image quality for coronary computed tomographic angiography using machine learning

Rine Nakanishi[1] · Sethuraman Sankaran[2] · Leo Grady[2] · Jenifer Malpeso[1] · Razik Yousfi[2] · Kazuhiro Osawa[1] · Indre Ceponiene[1] · Negin Nazarat[1] · Sina Rahmani[1] · Kendall Kissel[1] · Eranthi Jayawardena[1] · Christopher Dailing[1] · Christopher Zarins[2] · Bon-Kwon Koo[3] · James K. Min[4] · Charles A. Taylor[2] · Matthew J. Budoff[1]

Received: 5 October 2017 / Revised: 15 January 2018 / Accepted: 23 January 2018


## Abstract

**Objectives** Our goal was to evaluate the efficacy of a fully automated method for assessing the image quality (IQ) of coronary computed tomography angiography (CCTA).

**Methods** The machine learning method was trained using 75 CCTA studies by mapping features (noise, contrast, misregistration scores, and un-interpretability index) to an IQ score based on manual ground truth data. The automated method was validated on a set of 50 CCTA studies and subsequently tested on a new set of 172 CCTA studies against visual IQ scores on a 5-point Likert scale.

**Results** The area under the curve in the validation set was 0.96. In the 172 CCTA studies, our method yielded a Cohen's kappa statistic for the agreement between automated and visual IQ assessment of 0.67 ($p < 0.01$). In the group where good to excellent ($n = 163$), fair ($n = 6$), and poor visual IQ scores ($n = 3$) were graded, 155, 5, and 2 of the patients received an automated IQ score > 50 %, respectively.

**Conclusion** Fully automated assessment of the IQ of CCTA data sets by machine learning was reproducible and provided similar results compared with visual analysis within the limits of inter-operator variability.

**Key points**

- *The proposed method enables automated and reproducible image quality assessment.*
- *Machine learning and visual assessments yielded comparable estimates of image quality.*
- *Automated assessment potentially allows for more standardised image quality.*
- *Image quality assessment enables standardization of clinical trial results across different datasets.*

**Keyword** Computed tomography angiography · Coronary vessels · Cardiac imaging techniques · Machine learning · Image enhancement

**Abbreviations**

| | |
|---|---|
| AUC | Area under the curve |
| CAD | Coronary artery disease |
| CCTA | Coronary computed tomographic angiography |
| CNR | Contrast-to-noise ratio |
| FFR | Fractional flow reserve |
| ICA | Invasive coronary angiography |
| IQ | Image quality |
| ML | Machine learning |

Rine **Nakanishi** and Sethuraman **Sankaran** these two authors contributed equally to this work.

✉ Matthew J. Budoff
mbudoff@labiomed.org

[1] Los Angeles Biomedical Research Institute at Harbor UCLA Medical Center, Torrance, CA, USA

[2] HeartFlow Inc., Redwood City, CA, USA

[3] Department of Medicine, Seoul National University Hospital, Seoul, South Korea

[4] Department of Radiology, Weill Cornell Medical College and the New York Presbyterian Hospital, New York, NY, USA

## Introduction

Advances in coronary computed tomographic angiography (CCTA) have improved our ability to assess plaque coronary characteristics, anatomical stenosis severity, and subsequently functional significance [1–3], which helps in the detection of coronary artery disease (CAD) [4, 5]. In general, CCTA images are obtained from prospective padding or retrospective gating protocols from multiple phases and can provide rich

Published online: 23 March 2018

information to evaluate coronary arteries. Quantification of image quality (IQ) provides useful information regarding the ability to extract information from the CCTA and is usually evaluated visually by readers [6]. Therefore, the selection of the optimal phase with the best IQ to be used for assessment of CAD is a time-consuming process. The IQ may also vary within and between observers.

Studies on automated assessment of IQ have generally focused on either the global characteristics of poor IQ, such as low contrast, high noise, and a low contrast-to-noise ratio (CNR), or local IQ metrics such as motion or misregistration [7–9]. While an automated assessment of such IQ metrics may be beneficial, validation against core-lab assessment by expert readers is necessary for such metrics to be clinically useful.

In this article, we describe a new method for automated assessment of IQ of CCTA data sets by comparing it with a visually estimated IQ score.

## Materials and methods

### Study population

Briefly, the DeFACTO study and the DISCOVER-FLOW study were multinational prospective clinical trials for evaluating the diagnostic accuracy of fractional flow reserve (FFR) derived from CT ($FFR_{CT}$) compared with invasive FFR [3, 10]. All patients were clinically referred to invasive coronary angiography (ICA) for evaluation of CAD. All patients underwent a CCTA and an ICA with FFR. Patients with bypass grafts, stents, and chronic total occlusions were excluded. Each participating institution obtained Institutional Review Board approval, and all patients signed informed consent forms. Among 388 patients enrolled in the DeFACTO study ($n$ = 285) or the DISCOVER-FLOW study ($n$ = 103) [3, 10], we identified 297 patients for the current study after excluding 91 because of the absence of contrast ($n$ = 3) or insufficient automated centreline extraction ($n$ = 88). Though cases with insufficient automated centrelines also had many regions of un-interpretability, we excluded them because the ML algorithm relies on centreline location to calculate features. Overall, 125 patients for the training set ($n$ = 75) and validation set ($n$ = 50) were randomly selected for the ML algorithm, and the remaining 172 patients were used for comparing IQ between the automated and manual assessments (Fig. 1).

### CT image acquisition protocol

Details regarding the CT image protocol and analysis were documented in prior publications [10, 11]. In brief, all patients underwent a ≥ 64-slice CCTA scan (Lightspeed VCT, GE Healthcare, Milwaukee, WI; Somatom Sensation and Definition CT, Siemens, Forchheim, Germany; Brilliance 256 and 64, Philips, Surrey, UK; Aquilion One and 64, Toshiba, Otawara, Japan). In accordance with the Society of Cardiovascular Computed Tomography guidelines [6], patients received oral and/or intravenous beta-blocker medication to achieve a target heart rate of 60 beats or less per minute. They also received sublingual nitroglycerin for coronary artery dilation. In study 1, 72 % of patients received beta-blockers, but this information was not available for study 2. Scan parameters were obtained as follows: tube voltage 100 or 120 kVp, ≤ 0.75 mm slice thickness, and 512 × 512 matrix size. Helical or axial scans were obtained with prospective or retrospective electrocardiogram triggering. Scan parameters of non-contrast CT were obtained as follows: tube voltage 120 kVp, ≤ 3 mm slice thickness, and 512 × 512 matrix size.

### IQ assessment at an automated IQ core laboratory

An ML approach was used to map features of the CCTA images and automatically compute an IQ score in a blinded manner (Fig. 2). A manual IQ assessment was used as ground truth to train the system. This assessment was performed by trained readers who labelled cases deemed readable or unreadable as 1 or 0, respectively. We trained on this binary score instead of the Likert score to ensure a more reproducible ground truth. Since we train multiple random trees, an averaging of the binary outputs of individual trees can be scaled to 4 and interpreted as a Likert score. Some CCTA studies with different artefacts as well as two cases annotated as excellent and poor IQ are shown in Fig. 3. Readers were allowed to reject a scan for reasons such as vessels being uninterpretable because of high noise or CNR, heavy cardiac motion, blooming, calcification, and/or a large number of misregistrations.

Automated ostium detection, centreline extraction, and lumen segmentation were performed on these CCTA scans prior to the readers' assessment. The centrelines and lumen segmentation information were necessary because a key feature in the algorithm was the use of local IQ metrics, which were defined based on centrelines and lumen segmentation. In-house algorithms were used for the extraction of the centreline tree and the lumen segmentation. For purposes of this study, cases that fail the automated centreline detection or lumen segmentation were removed from the ground truth data set.

Global features: Figure 2 illustrates the schema of the ML method. First, misregistration, which occurs because of patient motion and is visible as a shift across image slices, was calculated directly from the image based on intensity correlation across neighbouring slices. Gradients in image intensity were calculated for each slice as $G_i = \sqrt{G_{x_i}^2 + G_{y_i}^2}$, where $G_i$ is the magnitude of intensity gradients for slice i, and $G_x$ and $G_y$ are gradients in the x and y directions respectively. The

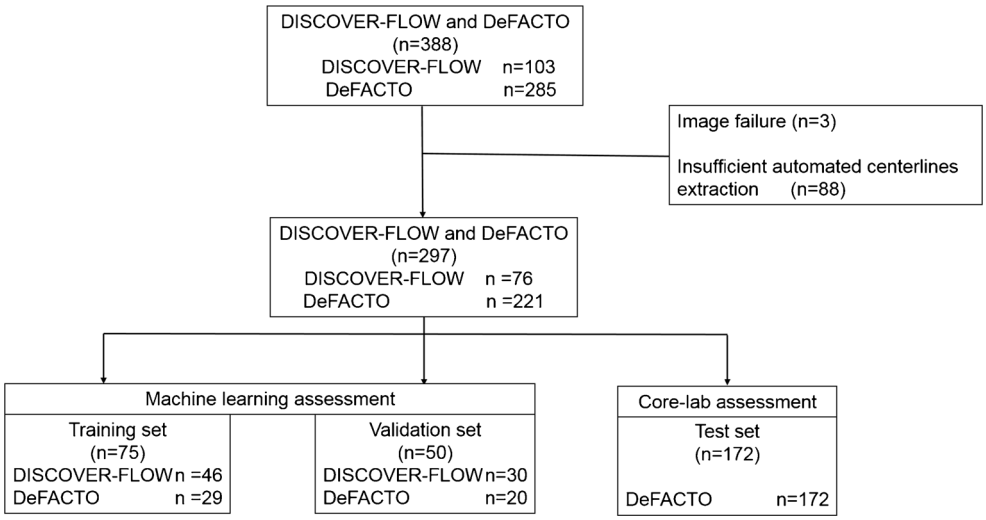


**Fig. 1** Schematic of the study design. The figure also shows the exclusion criterion and the split of the data into training, validation and test sets

cross-correlations between image intensity and gradients of neighbouring slices were used to arrive at a misregistration score. The mean magnitude, maximum magnitude, and number of misregistrations were calculated and included as features. Following this, voxels within and surrounding the aorta were identified to calculate noise, contrast, and background intensity. The average intensity of the voxels representing the aorta was calculated, and the average background intensity (voxels outside the aorta) was subtracted; the result (intensity – background) was assigned as contrast. The standard deviation of voxels within the aorta was also calculated and assigned as noise.

Local features: Image intensities on a 3 × 3 pixel grid around each centreline point were extracted. Statistics of these metrics (standard deviation, maximum and minimum) were calculated. In addition, the entropy value of lumen intensities (H) was calculated as $H = \sum_i p_i \log(p_i)$ where $p_i$ are probabilities of different intensity levels. Finally, an image sharpness estimate was calculated using a wavelet transform.

An additional feature, the standard deviation in the lumen area, was derived from the local features. The ground truth for this feature was based on a variability sub-study, wherein 15 readers performed manual lumen segmentation on the same set of patients and the coefficient of variation (COV) in the lumen area was assessed. A total of around 3000 vessel sections from five patients was used for assessment of COV. A linear regressor was used to map the local features to the COV in the lumen area. This value reflects the difficulty in assessment of lumen boundaries. A subset of the training data was used for this task.

Un-interpretability index: A predictive classifier is built that maps the local features and the standard deviation in the lumen area into a local binary un-interpretability index trained

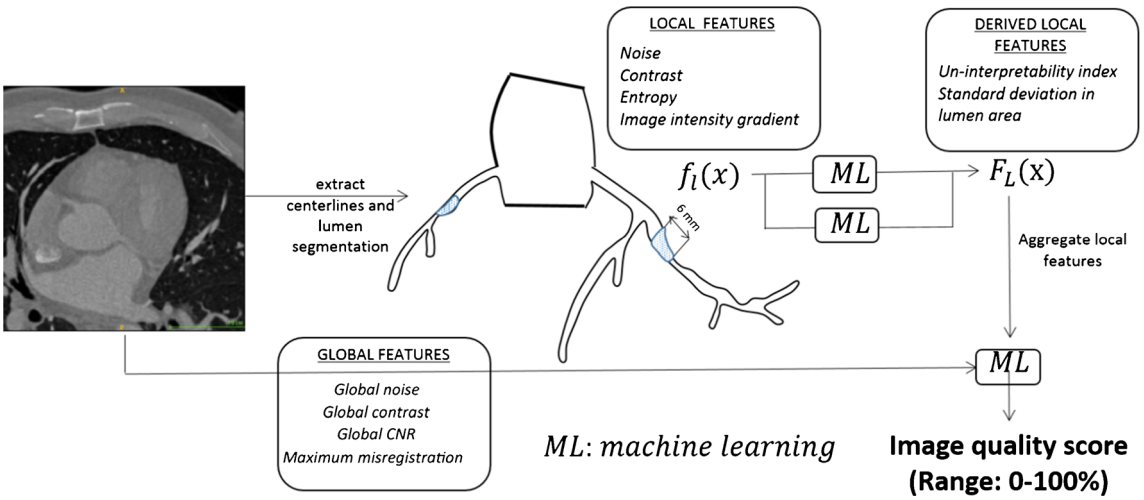


**Fig. 2** A schematic of the algorithm used to calculate the overall IQ score from a CCTA. First, global features that can be extracted directly from the image are computed. Then, an automated centreline extraction and lumen segmentation step are performed. A set of local features, which vary with space, is then calculated. These features are fed into two machine-learning algorithms for predicting lumen uninterpretability and the coefficient of variation in the lumen area respectively. The local features are statistically aggregated and combined with the global features to yield an overall IQ score. IQ, image quality; CCTA, coronary computed tomographic angiography, ML, machine learning, CNR, contrast-to-noise ratio

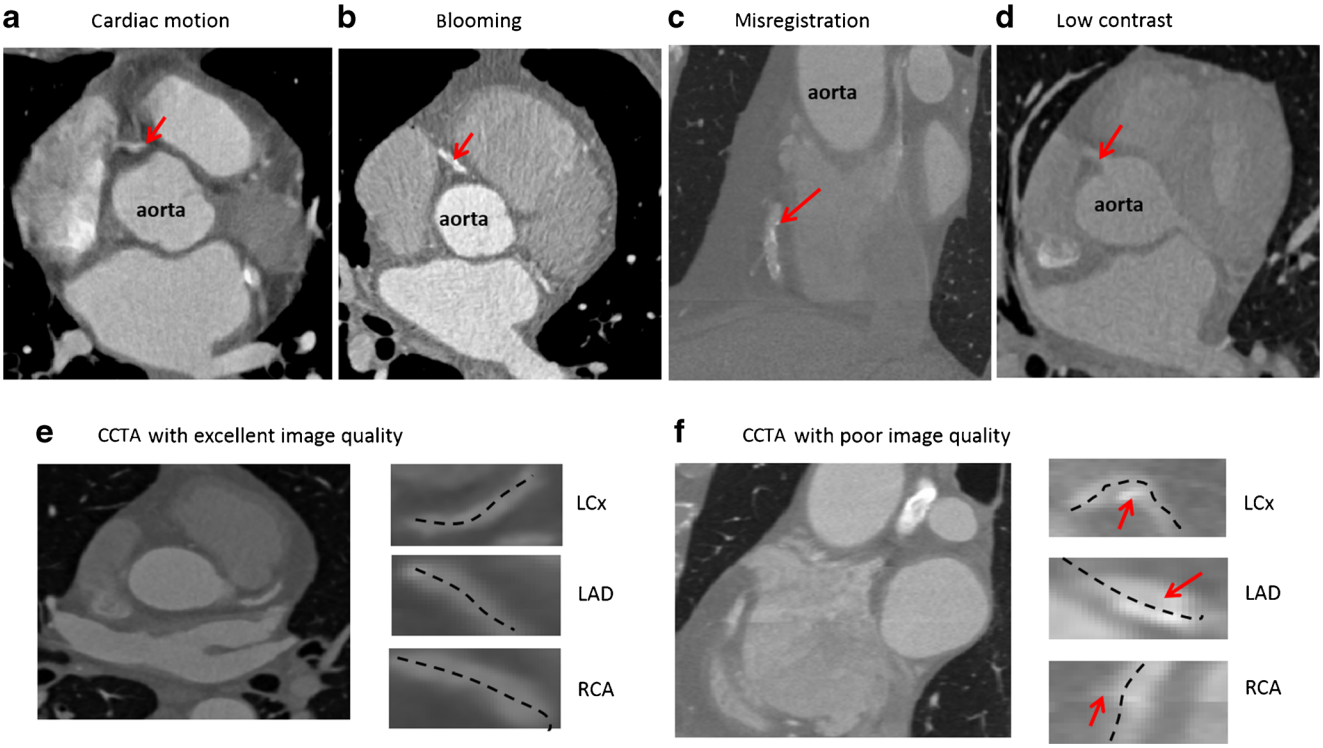


**Fig. 3** Examples of some cases that were manually rejected because of (**a**) cardiac motion, (**b**) noise, (**c**) blooming (dotted)/misregistration, and (**d**) low contrast. The red arrows point to locations where the impact of the artefact is visible. It is not uncommon for CCTA scans to have multiple artefacts simultaneously. Examples of two CCTA studies, one with excellent and one with poor IQ, are shown in (**e**) and (**f**) respectively. LAD, left anterior descending artery; LCx, left circumflex; RCA, right coronary artery

on the un-interpretable regions identified by human readers. A random forest regressor was used to map the local image features to the un-interpretability index. The random forest regressor is an aggregate of random trees, where each random tree is a decision tree built on a random subset of features. For instance, one decision tree might be built on the CNR and motion, whereas another may be built on the COV in the lumen area and misalignment. The power of random forests lies in combining many trees, where the trees are built with only a subset of the features. The decision-making criterion in each tree is decided such that the random forest regressor provides optimal output on the training data.

IQ score: The global features listed in the previous subsection and the derived local features (the mean and maximum un-interpretability index) are used to calculate an overall IQ score. A bootstrap aggregated random forest is used for classification and an IQ score is calculated as the mean value of the individual decision trees, converted to percentage.

The score ranges between 0 and 100 % for each CCTA. A ≥ 50 % is defined as good to excellent IQ and < 50 % is poor to unevaluable IQ. The calculation of automated IQ score per CCTA required less than a minute.

The ML algorithm was validated on two sets of data—initially, the binary uninterpretability index was validated on a binary test set of 50 CCTA studies. Subsequently, the actual score was tested on manual core-lab scores on 172 CCTA studies. The rationale for not training on the Likert score was the larger variability and subjectivity in the Likert scores compared with a binary uninterpretability index. Cross-validation on the training set was used to find the optimal parameters for the decision forest. Further, we used a Cohen's kappa statistic to quantify the agreement between manual and automated IQ.

## IQ assessment at a CT core laboratory

The overall IQ was reviewed on axial, sagittal, and coronal reformats in segments with > 1.5-mm luminal diameter at the CT core laboratory. IQ per segment was evaluated on multiplanar reformats using an 18-segment American Heart Association coronary model [6] by a CT core laboratory (Los Angeles Biomedical Research Institute, Torrance, CA, USA) in a blinded manner. Two expert readers with Society of Cardiovascular Computed Tomography level 3 reviewed the images and manually provided IQ scores for each individual segment. The score was given based on a 5-point Likert scale as follows: 4, excellent: clear delineation of vessel lumen boundaries without artefacts; 3, good: reserved ability to evaluate vessel lumen boundaries in the presence of artefacts; 2, fair: sufficient information to evaluate vessel lumen boundaries having a reduced IQ due to artefacts; 1, poor: impaired ability to evaluate vessel lumen boundaries because of

artefacts; 0, unevaluable: unable to evaluate vessel lumen boundaries because of artefacts [12–14] (Fig. 4). The visual IQ metric was normalised to a range of 0 and 4 and multiplied by 25, replicating a metric similar to the automated IQ score between 0 and 100. All the scores were independently and blindly assessed from automated IQ reads.

### Statistical analysis

Continuous variables for age, height, weight, body mass index, systolic pressure, diastolic blood pressure and heart rate were expressed as the mean ± standard deviation (SD) in the descriptive statistics. Median was chosen for the Agatston score because it was not normally distributed. The Wilcoxon signed-rank test was used to compare IQ between automated and core-lab assessments. A $p < 0.05$ was used to indicate statistical significance. A Cohen's kappa statistic was used to quantify the agreement between manual and automated IQ. Since a binary classification was used to assess manual reproducibility, we also measured the Cohen's kappa statistic in a similar manner. Though we stratified images into five buckets based on the Likert scale, we used a binary classification (the automated and manual image qualities either fall into the same bucket and are concordant, classified as 1, or discordant, classified as 0) for the calculation of kappa values. Weighted kappa on the 5-point Likert scale was not appropriate here because we compared the kappa against reproducibility between two independent readers, which were reported in the literature using a binary classification. Confidence intervals in Cohen's kappa statistic were also measured.

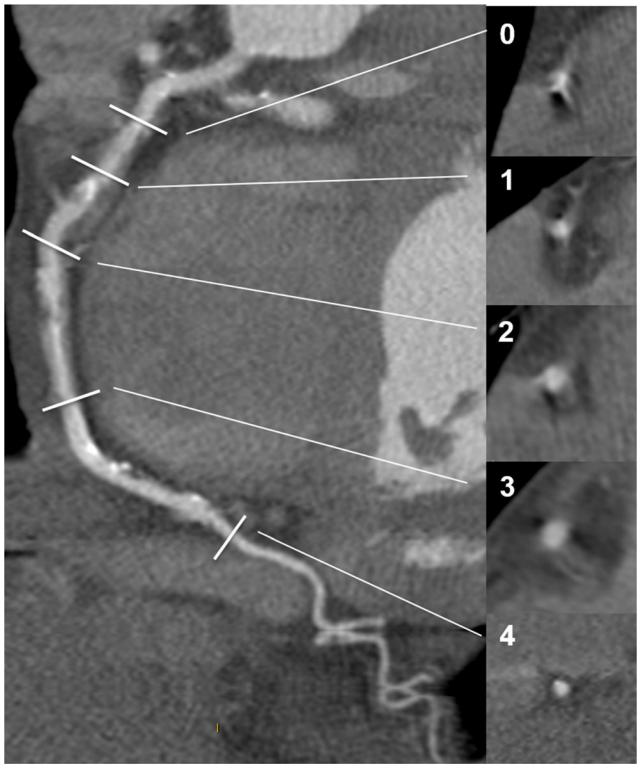


**Fig. 4** The figures demonstrate an example of the 5-point Likert scale. A score of 0 is given when vessel lumen boundaries are unevaluable because of artefacts (unevaluable). A score of 1 is given when delineating vessel lumen boundaries are impaired because of artefacts (poor). A score of 2 is given when vessel lumen boundaries can be sufficiently evaluated in the presence of artefacts (fair). A score of 3 is given when vessel lumen boundaries are fully evaluable in the presence of mild artefacts (good). Lastly, a score of 4 is given when clear delineations of vessel lumen boundaries are possible in the absence artefacts (excellent). IQ, image quality

## Results

Baseline patient characteristics for the study population of 172 patients are listed in Table 1. The fraction of male patients was 71.5 %. Two-thirds of patients had hypertension or hyperlipidaemia and 22 % of those had a history of CAD. Mean heart rate was 60.2 ± 0.7 beats/minute. The median coronary artery calcium Agatston score was 358. Overall, there was no significant difference between the study population in the training/validation and the test sets ($p > 0.05$ for all).

Based on a five-fold cross-validation on the training set, six features per tree on 101 decision trees was found to be optimal. The sensitivity, specificity, and accuracy of correctly classifying the un-interpretability index during cross-validation were 89 %, 93 %, and 92 % respectively. The performance on the validation (test) set showed a high accuracy of 94%, with a corresponding sensitivity, specificity, and positive and negative predictive value of 96 %, 91 %, 90 %, and 96 % respectively. The area under the receiver-operating characteristics curve for the un-interpretability index was 0.96 (Fig. 5).

Figure 6a shows the comparison of the COV in the lumen area between a machine learning method and the ground truth using a linear regressor. A correlation coefficient of 0.78 was achieved with the corresponding mean absolute error and root mean square error being 0.051 and 0.065 respectively. Figure 6b illustrates a histogram of the difference between automated IQ and core-lab assessments. The 95 % confidence intervals were within a Likert score of 2. The automated IQ score had a kappa of 0.67 (0.59-0.75), $p < 0.01$ against the visual IQ score, corresponding to a true-positive (TP), true-negative (TN), false-positive (FP), and false-negative (FN) count of 17, 141, 7, and 7 respectively. In the group where a good-to-excellent ($n = 163$), fair ($n = 6$), and poor visual IQ score ($n = 3$) were graded, 155, 5 and 2 of patients received an automated IQ score > 50 %, respectively.

Overall, 2392 segments among 172 patients were assessed for the visual IQ score. During segmental analysis, 43 (1.8 %), 30 (1.3 %), 192 (8.0 %), and 2127 segments (89.1 %) were visually scored as having an IQ of 0 or 1, 2, 3, and 4,

**Table 1** A summary of study demographics for the training/validation and test set

| | Training/validation set ($n$ = 125) | Test set ($n$ = 172) | $p$ value |
|---|---|---|---|
| Age (years) | 63.4 ± 8.5 | 62.8 ± 8.8 | 0.47 |
| Male gender | 69.5 % | 71.5 % | 0.49 |
| Height (cm) | 169.5 ± 9.3 | 169.7 ± 9.1 | 0.50 |
| Weight (kg) | 79.6 ± 15.4 | 76.6 ± 14.2 | 0.61 |
| Body mass index (kg/m$^2$) | 27.5 ± 3.9 | 26.5 ± 3.7 | 0.38 |
| Hypertension | 69.9 % | 71.8 % | 0.45 |
| Diabetes | 20.5 % | 21.2 % | 0.57 |
| Hyperlipidaemia | 86.7 % | 76.2 % | 0.22 |
| Never smoker | 50 % | 38.4 % | 0.27 |
| Former smoker | 39 % | 41.3 % | 0.63 |
| Current smoker | 11 % | 20.3 % | 0.31 |
| History of coronary artery disease | 14.5 % | 22.2 % | 0.29 |
| Coronary artery calcium score (median) | 320 | 358 | 0.19 |
| Systolic blood pressure (mmHg) | 134.9 ± 19.9 | 134.8 ± 17.9 | 0.50 |
| Diastolic blood pressure (mmHg) | 76.8 ± 10.4 | 77.9 ± 11.2 | 0.55 |
| Heart rate (bpm) | 63.5 ± 10.4 | 60.2 ± 9.5 | 0.43 |

respectively. Overall, in 265 segments with visual IQ scores < 4, there were 277 reasons for reducing the IQ. The most common reason for artefact was misalignment (35.4 %), followed by motion (30.0 %), image noise (22.0 %), coronary calcification (10.1 %), poor contrast (1.8 %), and others including beam hardening or segment after long total occlusion (0.7 %). Table 2 lists the distribution of reasons causing visual artefact by IQ scores. Motion artefacts were commonly associated with very poor/fair IQ scores, followed by misalignment, image noise, and coronary calcification.

Finally, we evaluated the reproducibility of the image quality method by running the algorithm thrice on the same image inputs. The results were identical, thereby demonstrating that the automated algorithm is 100 % reproducible.

## Discussion

The goal of our study was to demonstrate the performance and efficiency of a fully automated ML method to assess IQ. To our knowledge, this is the first study investigating automated assessment of CCTA IQ with a direct comparison with a manual core-lab read. We demonstrated that the agreement between the automated and visual IQ scores was similar to inter-observer variability in IQ reported by Sun et al. [16] with a kappa of 0.68. In addition, the proposed method enables fast patient-specific estimation of IQ on CCTA studies. We also demonstrated the reproducibility of the algorithm, which is not surprising since, given a training data set, there are no stochastic components in the algorithm.

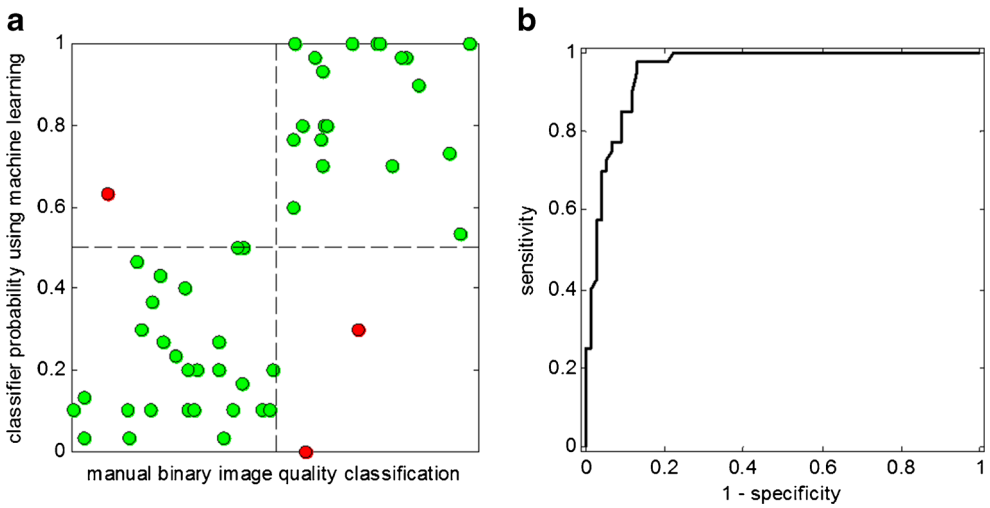


**Fig. 5** (**a**) Performance of the machine-learning algorithm on an initial validation set of 50 patients. Green dots represent correctly classified cases, with the x-axis denoting ground truth and y-axis being the automated IQ score. The x-axis is randomised to show the scatter of the points, since the ground truth values are binary and (**b**) the corresponding ROC curve having an area under the curve of 0.96. ROC, receiver-operator curve; IQ, image quality

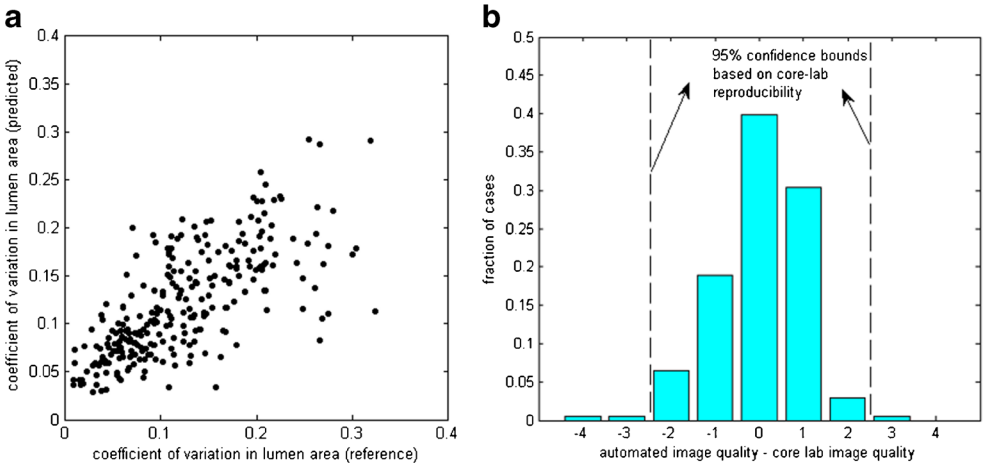


**Fig. 6** The figure shows (**a**) comparison of COV in the lumen area estimated using a machine-learning method compared with ground truth data. This sub-analysis was performed by having the same image assessed by 15 readers to estimate the coefficient of variation in the lumen area and (**b**) comparison of the difference between the manual and automated IQ algorithm normalised to a scale between 0 and 4. The 95 % confidence bounds of core-lab reproducibility based on inter-observer variability as quantified by Sun et al. [15] are also shown, demonstrating that the difference between the automated and manual IQ score in most of the CCTA studies lies within these bounds

The most recent guideline has suggested that standardised and optimised interpretations of the CCTA results help guide optimal patient care [17]. Despite advanced CCTA technologies, the prevalence of artefacts remains a limitation of this technique. In a previous meta-analysis of 27 studies analysing 22,798 segments, 4.2 % of segments were excluded from analysis because of unassessable IQ [18]. A previous multicentre study showed lower un-assessable segments with 2.9 % including motion artefact (75.6 %), coronary calcium (15.3 %), and poor contrast (8.4 %), whereas 15 % of patients still possessed at least one unevaluable segment [19]. Our results are consistent with these studies, demonstrating 3.1 % of segments were visually assessed as poor/unevaluable IQ. Since these cases in general take the most time for manual assessment, the automated algorithm would result in significant savings in analyst time. Moreover, motion artefact was the most common reason causing poor IQ and was seen in two-thirds of un-interpreted segments in the current study (1.8 % of total). This observation is also concordant with the study, showing 2.2 % of total segments were presented with motion artefacts [19].

IQ assessment by the proposed ML approach has the potential to reduce CT image evaluation time, thereby enabling automated selection of the optimal CT phase. In fact, the calculation time of automated IQ score per phase was less than a minute in the current study. Since we did not record how long the visual evaluation took, we could not directly compare the required time between the two methods. This automated method for measuring IQ may also be used to stratify data sets in clinical trials so that the performance may be assessed across different levels of image quality. Automated IQ measurement per se would not improve study quality, but would help a physician to select the best phase with best IQ from multiple phases, which may avoid diagnostic misinterpretations caused by the selection of a phase with artefacts.

Our findings demonstrated good concordance of IQ assessed by machine learning and tested against manual assessment by expert readers, which potentially allows for a more standardised IQ. Further, such standardisation by the automated algorithm would negate any effects of analyst competence and experience, which are critical in manual assessment of image quality.

**Table 2** Segmental scores and their source corresponding to calcification, motion, noise, contrast, and misalignment

| | Total ($n$ = 277 reason in 237 segments) | Visual IQ score 0 ($n$ = 45 reasons in 43 segments) | Visual IQ score 1 ($n$ = 2 reasons in 2 segments) | Visual IQ score 2 ($n$ = 35 reason in 30 segments) | Visual IQ score 3 ($n$ = 195 reason in 192 segments) |
|---|---|---|---|---|---|
| Calcification ($n$, %) | 28 (10.1 %) | 1 (2.2 %) | 0 (0 %) | 5 (14.3 %) | 22 (11.3 %) |
| Motion ($n$, %) | 83 (30.0 %) | 29 (64.5 %) | 2 (100 %) | 14 (40.0 %) | 38 (19.5 %) |
| Image noise ($n$, %) | 61 (22.0 %) | 5 (11.1 %) | 0 (0 %) | 12 (34.3 %) | 44 (22.6 %) |
| Poor contrast ($n$, %) | 5 (1.8 %) | 0 (0 %) | 0 (0 %) | 2 (5.7 %) | 3 (1.5 %) |
| Misalignment ($n$, %) | 98 (35.4 %) | 9 (20.0 %) | 0 (0 %) | 2 (5.7 %) | 87 (44.6 %) |
| Others ($n$, %) | 2 (0.7 %) | 1 (2.2 %) | 0 (0 %) | 0 (0 %) | 1 (0.5 %) |

There are numerous clinical applications of the proposed methodology. First, the method allows a real-time assessment of IQ. Second, the IQ is possibly an indicator of confidence in diagnostic accuracy though clinical validation studies are needed to confirm this.

There are certain limitations in the current study. Our study was a sub-study of the prospective multicentre study evaluating the diagnostic accuracy of $FFR_{CT}$ compared with invasive FFR. Most of the cases met the inclusion criteria of the study with good/excellent IQ for assessing $FFR_{CT}$. However, the dependency of the machine-learning algorithm on the ability to extract centrelines resulted in some of the patients being excluded from the study, though a majority of such cases also had an uninterpretable CCTA. One potential way to mitigate this is to train a second machine learning classifier that operates on a reduced feature set that does not take the centreline locations as inputs. In this regard, the feasibility of an automated method for evaluating IQ score in the real world remains uncertain. However, the prevalence of segments with un-assessable IQ in the current study was similar to that in a prior meta-analysis [18]. In addition, in this work, the confusion matrix (TP, TN, FP, and FN) does not contain a similar number of good and poor IQ data. In the final test set, there are more cases with good IQ than poor IQ. While our operating point sensitivity and positive predictive value on the validation set were 96 % and 90 % respectively, we did not have a sufficient number of rejected samples in the test set to power this observation statistically. Further, we were missing data on prior medication, which could have potentially impacted image quality. Finally, the relationship between image quality and diagnostic performance needs to be assessed.

## Conclusion

We developed an automated and reproducible method for assessment of IQ that compares well with the limits of inter-operator variability.

**Funding** The authors state that this work has not received any funding.

## Compliance with ethical standards

**Guarantor** The scientific guarantor of this publication is Dr. Matthew J. Budoff.

**Conflict of interest** The authors of this manuscript declare relationships with the following companies: Dr. Matthew Budoff receives grant support from GE Healthcare. Dr. Sankaran, Dr. Grady, Mr. Yousfi, Dr. Zarins, and Dr. Taylor are employees of HeartFlow. Dr. Min received modest speakers' bureau medical advisory board compensation and significant research support from GE Healthcare. All other authors of this manuscript declare no relationships with any companies, whose products or services may be related to the subject matter of the article.

**Statistics and biometry** No complex statistical methods were necessary in the current study.

**Informed consent** Written informed consent was obtained from all subjects (patients) in this study.

**Ethical approval** Institutional Review Board approval was obtained.

**Study subjects or cohorts overlap** Some study subjects have been previously reported in the JAMA and JACC.

**Methodology**

• This is a retrospective observational study using two previous multicentre studies.

## References

1. Budoff MJ, Dowe D, Jollis JG et al (2008) Diagnostic performance of 64-multidetector row coronary computed tomographic angiography for evaluation of coronary artery stenosis in individuals without known coronary artery disease: results from the prospective multicenter ACCURACY (Assessment by Coronary Computed Tomographic Angiography of Individuals Undergoing Invasive Coronary Angiography) trial. J Am Coll Cardiol 52(21):1724–1732
2. Achenbach S, Moselewski F, Ropers D et al (2004) Detection of calcified and noncalcified coronary atherosclerotic plaque by contrast-enhanced, submillimeter multidetector spiral computed tomography: a segment-based comparison with intravascular ultrasound. Circulation. 109(1):14–17
3. Min JK, Leipsic J, Pencina MJ et al (2012a) Diagnostic accuracy of fractional flow reserve from anatomic CT angiography. JAMA. 308(12):1237–1245
4. Douglas PS, Hoffmann U, Patel MR et al (2015) Outcomes of anatomical versus functional testing for coronary artery disease. N Engl J Med. 372(14):1291–1300
5. SCOT-HEART investigators (2015) CT coronary angiography in patients with suspected angina due to coronary heart disease (SCOT-HEART): an open-label, parallel-group, multicentre trial. Lancet 385(9985):2383–91.
6. Leipsic J, Abbara S, Achenbach S et al (2014) SCCT guidelines for the interpretation and reporting of coronary CT angiography: a report of the Society of Cardiovascular Computed Tomography Guidelines Committee. J Cardiovasc Comput Tomogr. 8(5):342–358
7. Naeemi MDRJ, Hollcroft N, Alessio AM, Roychowdhury S (2016) Application of big data analytics for automated estimation of CT image quality. Big Data (Big Data). IEEE International Conference on. IEEE 2016:3422–3431
8. Fronthaler HKK, Bigun J, Fierrez J, Alonso-Fernandez F, Ortega-Garcia J, Gonzalez-Rodriguez J (2008) Fingerprint image-quality estimation and its application to multialgorithm verification. IEEE Transactions on Information Forensics and Security. 3(2):331–338
9. Marin TKM, Hendrik PP, Wernick MN, Yang Y, Brankov JG (2011) Numerical observer for cardiac motion assessment using machine learning. In Proc. of SPIE. 7966:79660G–796601
10. Koo BK, Erglis A, Doh JH et al (2011) Diagnosis of ischemia-causing coronary stenoses by noninvasive fractional flow reserve computed from coronary computed tomographic angiograms. Results from the prospective multicenter DISCOVER-FLOW (Diagnosis of Ischemia-Causing Stenoses Obtained Via

Noninvasive Fractional Flow Reserve) study. J Am Coll Cardiol. 58(19):1989–1997
11. Min JK, Berman DS, Budoff MJ et al (2011) Rationale and design of the DeFACTO (Determination of Fractional Flow Reserve by Anatomic Computed Tomographic AngiOgraphy) study. J Cardiovasc Comput Tomogr. 5(5):301–309
12. Leipsic J, Labounty TM, Heilbron B et al (2010) Adaptive statistical iterative reconstruction: assessment of image noise and image quality in coronary CT angiography. AJR Am J Roentgenol. 195(3):649–654
13. Min JK, Koo BK, Erglis A et al (2012b) Effect of image quality on diagnostic accuracy of noninvasive fractional flow reserve: results from the prospective multicenter international DISCOVER-FLOW study. J Cardiovasc Comput Tomogr. 6(3):191–199
14. Abbara S, Blanke P, Maroules CD et al (2016) SCCT guidelines for the performance and acquisition of coronary computed tomographic angiography: A report of the society of Cardiovascular Computed Tomography Guidelines Committee: Endorsed by the North American Society for Cardiovascular Imaging (NASCI). J Cardiovasc Comput Tomogr. 10(6):435–449
15. Sun K, Li K, Han R et al (2015) Evaluation of high-pitch dual-source CT angiography for evaluation of coronary and carotid-cerebrovascular arteries. Eur J Radiol. 84(3):398–406
16. Zir LM, Miller SW, Dinsmore RE, Gilbert JP, Harthorne JW (1976) Interobserver variability in coronary angiography. Circulation. 53(4):627–632
17. Cury RC, Abbara S, Achenbach S et al (2016) CAD-RADS(TM) Coronary Artery Disease - Reporting and Data System. An expert consensus document of the Society of Cardiovascular Computed Tomography (SCCT), the American College of Radiology (ACR) and the North American Society for Cardiovascular Imaging (NASCI). Endorsed by the American College of Cardiology. J Cardiovasc Comput Tomogr. 10(4):269–281
18. Hamon M, Biondi-Zoccai GG, Malagutti P et al (2006) Diagnostic performance of multislice spiral computed tomography of coronary arteries as compared with conventional invasive coronary angiography: a meta-analysis. J Am Coll Cardiol. 48(9):1896–1910
19. Puchner SB, Liu T, Mayrhofer T et al (2014) High-risk plaque detected on coronary CT angiography predicts acute coronary syndromes independent of significant stenosis in acute chest pain: results from the ROMICAT-II trial. J Am Coll Cardiol. 64(7):684–692