Rahil Shahzad, MSc, Theo van Walsum, PhD, Michiel Schaap, PhD, Alexia Rossi, MD, Stefan Klein, PhD, Annick C. Weustink, MD, PhD, Pim J. de Feyter, MD, PhD, Lucas J. van Vliet, PhD, Wiro J. Niessen, PhD

Rationale and Objectives: The aim of this study was to automatically detect and quantify calcium lesions for the whole heart as well as per coronary artery on non-contrast-enhanced cardiac computed tomographic images.

Materials and Methods: Imaging data from 366 patients were randomly selected from patients who underwent computed tomographic calcium scoring assessments between July 2004 and May 2009 at Erasmus MC, Rotterdam. These data included data sets with 1.5-mm and 3.0-mm slice spacing reconstructions and were acquired using four different scanners. The scores of manual observers, who annotated the data using commercially available software, served as ground truth. An automatic method for detecting and quantifying calcifications for each of the four main coronary arteries and the whole heart was trained on 209 data sets and tested on 157 data sets. Statistical testing included determining Pearson’s correlation coefficients and Bland-Altman analysis to compare performance between the system and ground truth. Wilcoxon’s signed-rank test was used to compare the interobserver variability to the system’s performance.

Results: Automatic detection of calcified objects was achieved with sensitivity of 81.2% per calcified object in the 1.5-mm data set and sensitivity of 86.6% per calcified object in the 3.0-mm data set. The system made an average of 2.5 errors per patient in the 1.5-mm data set and 2.2 errors in the 3.0-mm data set. Pearson’s correlation coefficients of 0.97 (P < .001) for both 1.5-mm and 3.0-mm scans with respect to the calcium volume score of the whole heart were found. The average R values over Agatston, mass, and volume scores for each of the arteries (left circumflex coronary artery, right coronary artery, and left main and left anterior descending coronary arteries) were 0.93, 0.96, and 0.99, respectively, for the 1.5-mm scans. Similarly, for 3.0-mm scans, R values were 0.94, 0.94, and 0.99, respectively. Risk category assignment was correct in 95% and 89% of the data sets in the 1.5-mm and 3-mm scans.

Conclusions: An automatic vessel-specific coronary artery calcium scoring system was developed, and its feasibility for calcium scoring in individual vessels and risk category classification has been demonstrated.

Key Words: Calcium scoring; coronary arteries; machine learning.

Coronary artery disease is one of the leading causes of mortality worldwide (1). Many clinical studies have shown that the amount of calcium in coronary artery plaques correlates with the risk for future cardiovascular events (2–7).

Calcium scoring is routinely performed on low-dose, non-contrast-enhanced computed tomographic (CT) scans by manually annotating all calcium objects present in the main vessels of the coronary artery tree: the left main (LM), left circumflex, left anterior descending (LAD), and right coronary arteries. Subsequently, on the basis of all selected objects per patient, the Agatston (8), mass (9), or volume (10) score is determined. Recently, it has been suggested that vessel-specific calcium scoring or rather risk assessment on the basis of individual vessels is more informative compared to whole-heart calcium scoring (11). Similar findings have been reported in other large population studies. Williams et al (12) observed that the mortality rate of the patients they followed increased proportionally with the rise in the number of calcified lesions, and they also observed that all patients who had Agatston scores \( \geq 1000 \) in the LM died during follow-up. Mohlenkamp et al (13) observed that LM disease was an independent predictor of hard events. Vessel-specific scores also facilitate in better understanding calcium progression in longitudinal studies. Budoff and Raggi (14) found that calcium increases by approximately 20% to 30% each year. The Multi-Ethnic Study of Atherosclerosis (15) investigated the relationship between calcium scores at baseline and stenosis in individual vessels. It reported a positive correlation between calcium scores and vessel-specific scores in the individual artery beds.
Manual scoring is a time-consuming task because it consists of drawing contours to obtain the region of interest or clicking inside all calcium objects. Isgum et al (16) demonstrated the feasibility of automating this task, but the feasibility of automatic per vessel calcium scoring has not been demonstrated.

The purpose of our work was to develop and evaluate an automatic calcium scoring system for electrocardiographically gated, non-contrast-enhanced cardiac CT scans that yields scores for the whole heart as well as for the individual coronary arteries. Our system uses an atlas-based estimate of the coronary artery locations, which permits the system to assign calcium lesions to the correct coronary arteries (17). The system uses a machine-learning approach to discriminate true calcium objects from all detected candidate objects.

MATERIALS AND METHODS

Data Description

We retrospectively selected a random subset of patients who underwent cardiac non-contrast-enhanced CT scans for calcium score evaluation between July 2004 and May 2009 at the Erasmus Medical Center for whom calcium scoring reports were electronically available along with the scans in the picture archiving and communication system. In total, 366 scans (280 in men, 86 in women) were retrieved. The scans were acquired using four different generations of Siemens scanners (Definition Flash, Definition AS+, Sensation 64, and Definition; Siemens Medical Solutions, Forchheim, Germany). A detailed description of the data characteristics is provided in Table 1. A tube voltage of 120 kV was used for all scans. Two different slice spacings were used in tomographic reconstruction: 234 scans had 1.5-mm slice spacing, and 132 scans had 3.0-mm slice spacing, henceforth referred to as the 1.5-mm and 3.0-mm scans, respectively. Both types of scans had a slice thickness of 3.0 mm and a field of view of approximately 180 mm. The in-plane resolution was 0.35 × 0.35 mm² on average. Both the 1.5-mm and 3.0-mm scans were randomly divided into different sets for designing, training, and testing of the calcium scoring system, as shown in Figure 1. The age distribution of male and female patients over the data sets was similar.

System Overview

The complete calcium scoring system consists of the following stages. First, candidate calcium objects are determined from the CT scan. Subsequently, a classifier that uses local image features is applied to determine which of the detected candidate objects are coronary artery calcifications. Finally, a coronary artery location estimate is used to assign the calcifications to one of the main coronary arteries. These calcifications are used to compute the calcium scores. The following paragraphs describe the candidate detection, the design of the classifier including how it was trained and which features are used, and how the final calcium scores are computed and presented. The work flow of our system, including the calcified object labeling, is shown in Figure 2.

Candidate Detection

Candidate calcium objects were obtained from the CT scans by thresholding at 130 Hounsfield units and discarding all objects >1500 mm³ and also those <1.5 mm³, which are assumed to correspond to bone and noise, respectively (9). The candidate objects consist of true calcified objects, including arterial calcifications, aortic calcifications, and calcifications in the valves and false objects due to noise and imaging artifacts.

Classifier

We developed a classifier for the candidate objects that differentiates between arterial calcium and the rest on the basis of local image information (features). The classifier was trained using manually annotated CT data (the training set). We used the design set to experimentally determine the optimal classifier and feature set. The image features that were investigated, classifier selection, and feature selection are presented in the following sections. We built and trained two classifiers, one for the 1.5-mm data and one for the 3.0-mm data. The 1.5-mm training data had a total of 366,876 candidate calcium objects, of which 1155 were true calcium objects. The 3.0-mm training data had a total of 112,302 candidate calcium objects, of which 439 were true calcium objects. The system automatically determines the slice spacing of the data set to be analyzed and applies the appropriate classifier.

Features

A feature-based classification (18) approach was adopted for classifying candidate objects. In total, 62 features were considered, which are listed in Table 2 and explained below.

Object-based features. Five different types of object-based features were computed: the volume of the candidate object, the

---

TABLE 1. Details of the Data Sets

<table>
<thead>
<tr>
<th>Variable</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data sets</td>
<td>366</td>
</tr>
<tr>
<td>Total</td>
<td>Women</td>
</tr>
<tr>
<td>Men</td>
<td>280 (76.5%)</td>
</tr>
<tr>
<td>Age, women (y)</td>
<td>57 (28–83)</td>
</tr>
<tr>
<td>Age, men (y)</td>
<td>57 (21–84)</td>
</tr>
<tr>
<td>Scanner type*</td>
<td>Definition (3.0-mm scans)</td>
</tr>
<tr>
<td>Sensation 64</td>
<td>205 (56%)</td>
</tr>
<tr>
<td>Definition Flash</td>
<td>16 (4.4%)</td>
</tr>
<tr>
<td>Definition AS+</td>
<td>13 (3.5%)</td>
</tr>
</tbody>
</table>

Data are expressed as number (percentage) or as mean (range). *All from Siemens Medical Solutions (Forchheim, Germany).
maximum and average intensity of the object, and two shape features, blob likeness and plate likeness (19).

**Multi-scale image derivatives.** The intensity of the object at the maximum intensity point after Gaussian image derivatives was computed at five different scales (Gaussian standard deviation in-slice between 0.3 and 4.8 mm with one sample per octave and between slices from 1.5 to 24 mm with one sample per octave) up to the second order.

**Coronary artery location estimate.** An atlas-based method (20) was used to determine a location estimate for the coronary artery locations. Coronary artery locations from 85 different CT angiographic scans were used to build 10 atlases encoding the distribution of coronary arteries across the population. These atlases were registered to the CT scan to obtain the patient-specific coronary artery location estimate for each individual artery. The selection criteria of the 10 atlas scans and the exact procedure to compute the artery location estimates are described in detail in the recent work of Shahzad et al (17). The artery location feature is used both for the classification of candidate objects and for the calculation of calcium score per artery.

**Position-based features.** Positions (x, y, and z) in actual image space and in a standardized coordinate space were used. We introduced the standardized space to account for the varying position of the heart on CT scans. This standardized space is constructed by pairwise registration of 10 atlas images (the same 10 atlases that were used for creating the coronary artery location estimates). Registration was performed in two stages: an initial affine registration followed by a nonrigid registration using ElastiX, a publicly available registration package (21). The standardized space is then obtained by averaging the resulting deformations. The midpoint between the right and left coronary ostia is defined as the origin in the...
standardized space. A CT scan of a new patient is mapped into this standardized space, such that relative positions in this standardized coordinate system can be used.

**Feature Selection and Classifier Selection**

On the design set, feature selection was performed to determine the set of features that gives optimal performance in detecting calcified objects in the coronary arteries. In our case, because the training set was limited (on average, there were only seven calcium objects per data set), it was also important to reduce the number of features to avoid the curse of dimensionality (ie, having too few training samples in a high-dimensional feature space) (18). Twenty-one features were selected on the basis of a forward feature selection step. We also applied a backward feature selection algorithm and observed that the selected feature set was identical.

We investigated the performance of the \( k \)-nearest-neighbor classifier for different numbers of neighbors, \( k \), ranging from one to 15 (odd values), with forward feature selection. We found that the best classifier for this problem is a nine-nearest-neighbor classifier, which yielded the smallest classification error at the object level.

The set of 21 best features retained were volume; maximum and average intensity; \( z \) position in the image and mean space; coronary artery position estimate; Gaussian filter with scale 1, 2, 4 and 8; first-order Gaussian derivative with scale 1, 8, and 16 in \( z \) direction; second-order Gaussian derivative with scale 1, 2, and 4 in the \( x \) direction; and the same scales in the \( y \) direction and \( z \) direction with scale 1, 2, and 4 in the \( z \) direction. The unit of the scale was 0.3 mm in the \( x \) and \( y \) directions and 1.5 mm in the \( z \) direction (corresponding to size of a voxel in the 1.5-mm data set).

**Calcium Scores**

The system calculates Agatston, mass, and volume scores for the detected calcium objects (22). The system presents the scores for the whole heart as well as the individual arteries. The assignment of the calcium object to one of the arteries is achieved by using the coronary artery location estimate feature. Because we use population-based information from the atlases to compute the location estimates of the coronary arteries and because of the large anatomic variation in the length of the LM coronary artery, we decided to label the LM and LAD as a single vessel. The other two vessels were labeled as the left circumflex and right coronary arteries.

**Risk Categorization**

Patients were assigned to different risk categories on the basis of the whole-heart Agatston scores (23,24). Hoff et al (25) studied the distribution of calcium lesions in 35,246 patients with respect to age and gender and showed that the calcium score distribution depended on the age and sex of the patients. They proposed to present calcium scores as 10th, 25th, 50th, 75th, and 90th percentile rank groups. Our system assigns patients to the appropriate risk category using this method, hence accounting for age and sex.
Reference Standard

The reference standard calcium scores were obtained from the calcium scoring reports that were stored in the picture archiving and communication system along with the patient scans. Calcium scoring for generating the reports was performed manually using the syngo Calcium Scoring tool (Siemens Medical Solutions). We reproduced the set of calcium objects in the scans using the reports.

Statistical Analysis

We report the system's sensitivity and specificity with respect to object detection. To evaluate the scores per patient, Pearson's correlation coefficient ($R$) was calculated and Bland-Altman plots (26) were created for the entire heart as well as for individual arteries. The analysis was performed using MATLAB version 7.9.0. (The MathWorks, Natick, MA). A confusion matrix was used to report errors in automatic risk categorization from the whole-heart calcium scores. The accuracy of the automatic method for calcium scoring was compared to interobserver variability on a subset of 50 patients; differences in total calcium scores (for Agatston, mass, and volume scores) were analyzed using Wilcoxon's signed-rank test, for which we report the $Z$ statistic to indicate significance.

RESULTS

Overall Performance of the System for Calcium Object Detection

1.5-mm scans. The system was tested on 101 data sets comprising 281,138 candidate objects, of which 787 were true calcium objects. The system yielded per object sensitivity of 81.2% and specificity of 99.6%.

3.0-mm scans. The system was tested on 56 data sets comprising 64,555 candidate objects, of which 300 were true calcium objects.
objects. The system had per object sensitivity of 86.7% and specificity of 97.4%.

**Performance of the System on the Patient Calcium Scores**

**1.5-mm scans.** On average, the system made one false-positive error and 1.5 false-negative errors per patient. We obtained Pearson’s correlation coefficients of 0.97, 0.95, and 0.97 ($P < .001$) between automatic and manual scoring on the whole heart with respect to the Agatston, mass, and volume scores. The corresponding correlation coefficients for each of the arteries are shown in Table 3. Of the 101 patients, five were assigned to different risk categories; two were off by two categories, and the others were off by one category. Note that cases close to a boundary can easily move to the neighboring category. The confusion matrix is shown in Table 4; 95% of the scans were assigned to the correct risk percentile.

**3.0-mm scans.** On average, the system made 1.5 false-positive errors and 0.7 false-negative errors per patient. The correlation coefficients between automatic and manual scoring with respect to the Agatston, mass, and volume scores were all 0.96 ($P < .001$). The corresponding correlation coefficients for each of the arteries are shown in Table 3. Of the 56 patients, six were assigned to the wrong risk categories; only one scan was off by two categories. The confusion matrix is shown in Table 4; 89.3% of the scans were assigned to the correct risk percentile.

The whole-heart volume score correlations between the system and the manual observers are presented in Figure 3.
along with explanations of the few classification errors. The vessel-specific correlation curves on the 1.5-mm data sets are shown in Figure 4, and the corresponding correlation values for the 3.0-mm scans are presented in Table 3. A few examples of misclassified objects are shown in Figure 5.

**Method Performance in Left-dominant and Balanced Subjects**

We performed an additional experiment to estimate the mislabeling errors made by our method when quantifying calcium in left-dominant and balanced subjects. We randomly selected 100 subjects from our data set and selected all calcium objects that were in the region that could possibly be supplied by the three different side branches resulting from the different dominant systems. We could accurately regionalize this area by using information from our standardized coordinate space. We found 26 calcium lesions belonging to 20 subjects in this region; the total volume of the lesions was 687.7 mm$^3$. Using the knowledge that on average, 13% of a population is not right dominant (27), and assuming a similar distribution over our subjects, we would make a total of 3.3 mislabeling errors over the 100 subjects. The average mislabeling error made was 0.26 objects, and the volumetric error per left-dominant or balanced subject would thus be 6.8 mm$^3$, which is negligible compared to the average volumetric calcium score present in the main branch.

**Interobserver Variability**

The correlation coefficients between the observers were 0.98 for all three scoring methods, and for the automatic method with respect to each of the observers, the mean R values for the three scores were 0.97 and 0.95 ($P < .001$). Table 5 lists the median values, $Z$ statistics (on the basis of Wilcoxon’s signed-rank test), and Bland-Altman limits of agreement. From the $Z$ statistics, it can be observed that the system did not make statistically significant errors compared to each of the observers.

**DISCUSSION**

Our automatic system obtained good sensitivity and high specificity in detecting calcified objects. We also found that the agreement between the automatic and manual scores was very close to the interobserver agreement. Isgum et al (16) reported sensitivity of 73.8% with respect to object detection, obtained by automatic whole-heart calcium scoring on a cohort of only female patients, sampled to obtain a slice spacing of 3.0 mm. This sensitivity is less than the sensitivity obtained by our system, which was 83.9% on the entire data set. When we compared the results of only female patients from our test set (42 in total), we obtained sensitivity of 82.2%.

The main advantage of our method is the automatic artery-specific calcium scoring, permitting large-scale epidemiology
or long-term prognostic evaluation studies (15,28) to better investigate the value of individual artery calcium scoring as a risk predictor (11).

The atlases used for estimating in which coronary artery calcifications are located were derived from a random population of 85 subjects. We calculated the overall spatial distribution of the individual coronary arteries and their branches over this population. Thus, we did not differentiate between left-dominant and right-dominant subjects. Because the majority of the subjects in a given population are right dominant (27), our atlases will label the posterior descending artery as belonging to the right coronary artery. This mislabeling of the vessels for the left-dominant and balanced subjects does have a negligible effect on the correlation graphs between the automatic method and the ground truth, because of the small size of the calcium lesions found in the smaller vessels. Hence we conclude that mislabeling errors in the left and balanced subjects by our method will not have a huge impact on vessel-specific calcium scoring.

The automatic artery-specific scores generated by our system correlate very well with those obtained manually. The bias, obtained from the Bland-Altman analysis, suggests that the system slightly underestimates scores on 1.5-mm scans, while scores are slightly overestimated on 3.0-mm scans. Most errors in vessel labeling were caused by incorrect object classification. The exceptions were some errors in the LM and the proximal part of the LAD. Variability in LM and LAD anatomy hampers the use of global information for separating the LM from the LAD, which is why we present the scores for LM and LAD combined. The correlation coefficients for the LM and LAD, respectively, were 0.83 and 0.98 on 1.5-mm scans and 0.57 and 0.97 on 3.0-mm scans. The errors made while distinguishing the LM from the LAD can be resolved if we would use image information from a corresponding contrast-enhanced CT scan of the same patient, in which the arteries are clearly distinguishable. However, in our study, we assumed that only non-contrast-enhanced CT scans were available. Even though our system made two errors per patient, this did not have adverse effects in categorizing the patients into risk percentiles; only three patients were assigned to risk percentiles that was off by two categories. Also, there were 27 patients in total who had calcium scores of zero. Our system correctly assigned scores of zero to 18 of these patients, and the average error made in the remaining scans was a score of 2.8 (Agatston). None of these 27 patients was assigned to a different risk percentile.

The system is completely automatic. It can automatically provide calcium scores when viewing an image, provided the reconstruction phase of the scan was completed. The user may want to glance through the scan and correct for the misclassified objects, which on average are limited to two objects per scan. The process of correcting false-positives, which generally occur around the aorta and the mitral valve, can be made easier by assigning a separate color to the suspicious objects.

<table>
<thead>
<tr>
<th>TABLE 5. Coronary Artery Calcium Scores of the Automated System (A) and Observers O1 and O2 (Median), Pearson’s Correlation Coefficient (R), Z Statistics Using Wilcoxon’s Signed-rank test, and Bland-Altman Limits of Agreement for Interobserver and the System with Respect to the Observers</th>
<th>Score</th>
<th>Median (Range)</th>
<th>R*</th>
<th>Z Statistics</th>
<th>Bland-Altman</th>
</tr>
</thead>
<tbody>
<tr>
<td>Volume</td>
<td>A</td>
<td>204 (0 to 1700)</td>
<td>214 (0 to 1769)</td>
<td>241 (0 to 1789)</td>
<td>215 (0 to 2177)</td>
</tr>
<tr>
<td>Mass</td>
<td>A</td>
<td>28 (0 to 378)</td>
<td>39 (0 to 389)</td>
<td>41 (0 to 366)</td>
<td>41 (0 to 366)</td>
</tr>
<tr>
<td>Agatston</td>
<td>A</td>
<td>216 (0 to 2117)</td>
<td>215 (0 to 2177)</td>
<td>217 (0 to 2177)</td>
<td>217 (0 to 2177)</td>
</tr>
</tbody>
</table>

*P < .001.
A limitation of our study is that the patient scans were acquired on equipment from only one vendor, Siemens. We did not investigate the performance of our system with data acquired using scanners from other vendors. However, because our method learns from example data sets, we presume that it can be adapted to data from other vendors or protocols, provided that a set of training data are available for the learning step.

CONCLUSIONS

We developed an automatic vessel-specific coronary artery calcium scoring system and have demonstrated the feasibility of calcium scoring and risk category classification using this system.

REFERENCES