# Standardized evaluation framework for evaluating coronary artery stenosis detection, stenosis quantification and lumen segmentation algorithms in Computed Tomography Angiography

H.A. Kirişli[a,b], M. Schaap[a], C.T. Metz[a], A.S. Dharampal[c,d], W.B. Meijboom[d], S.L. Papadopoulou[c], A. Dedic[d], K. Nieman[c,d], M.A. de Graaf[f,g], M.F.L. Meijs[e], M.J. Cramer[e], A. Broersen[b], S. Cetin[l], A. Eslami[u,v], L. Flórez-Valencia[p], K.L. Lor[m], B. Matuszewski[y], I. Melki[n,o], B. Mohr[k], I. Öksüz[r], R. Shahzad[a,i], C. Wang[t], P.H. Kitslaar[b], G. Unal[l], A. Katouzian[u,w], M. Orkisz[q], C.M. Chen[m], F. Precioso[x], L. Najman[n], S. Masood[k], D. Ünay[s], L. van Vliet[i], R. Moreno[t], R. Goldenberg[h], E. Vuçini[j], G.P. Krestin[c], W.J. Niessen[a,i], T. van Walsum[a,*]

[a] *Biomedical Imaging Group Rotterdam, Dept. of Radiology and Med. Informatics, Erasmus MC, Rotterdam, the Netherlands*
[b] *Div. of Image Processing, Dept. of Radiology, Leiden UMC, Leiden, the Netherlands*
[c] *Dept. of Radiology, Erasmus MC, Rotterdam, the Netherlands*
[d] *Dept. of Cardiology, Erasmus MC, Rotterdam, the Netherlands*
[e] *Dept. of Cardiology, UMC Utrecht, Utrecht, the Netherlands*
[f] *Dept. of Cardiology, Leiden UMC, Leiden, the Netherlands*
[g] *The Interuniversity Cardiology Institute of the Netherlands, Utrecht, the Netherlands*
[h] *Rcadia Medical Imaging, Haïfa, Israël*
[i] *Quantitative Imaging Group, Imaging Science and Technology, Faculty of Applied Sciences, Delft Univ. of Technology, Delft, the Netherlands*
[j] *VRVis Research Center for Virtual Reality and Visualization, Vienna, Austria*
[k] *Toshiba Medical Visualization Systems, Edinburgh, UK*
[l] *Faculty of Engineering and Natural Sciences, SabancıUniversity, Turkey*
[m] *Institute of Biomedical Engineering, National Taiwan University, Taipei, Taiwan*
[n] *Université Paris-Est, Laboratoire d'Informatique Gaspard-Monge, Equipe A3SI, Noisy-le-Grand, France*
[o] *GE Healthcare, Buc, France*
[p] *Grupo Takina, Departamento de Ingeniería de Sistemas, Pontificia Universidad Javeriana, Bogotá, Colombia*
[q] *Université de Lyon, CREATIS; CNRS UMR 5220; INSERM U 1044; INSA-Lyon, Lyon, France*
[r] *Electrical and Electronics Engineering, Bahçeşehir University, Istanbul, Turkey*
[s] *Biomedical Engineering, Bahçeşehir University, Istanbul, Turkey*
[t] *Center for Medical Imaging Science and Visualization, Department of Medical and Health Sciences, Linköping University, Linköping, Sweden*
[u] *Computer Aided Medical Procedures, Technical University of Munich, Munich, Germany*
[v] *Institute for Biomathematics and Biometry, Helmholtz Zentrum Munich, Germany*
[w] *Biomedical Engineering Department, Columbia University, New York, USA*
[x] *University Nice-Sophia Antipolis, Laboratory of Informatics, Signal and Systems (I3S), Nice Sophia Antipolis, France*
[y] *School of Computing Engineering and Physical Sciences, University of Central Lancashire, Preston, UK*

**Abstract**

Though conventional coronary angiography (CCA) has been the standard of reference for diagnosing coronary artery disease in the past decades, computed tomography angiography (CTA) has rapidly emerged, and is nowadays widely used in clinical practice. Here, we introduce a standardized evaluation framework to reliably evaluate and compare the performance of the algorithms devised to detect and quantify the coronary artery stenoses, and to segment the coronary artery lumen in CTA data. The objective of this evaluation framework is to demonstrate the feasibility of dedicated algorithms to: 1) (semi-)automatically detect and quantify stenosis on CTA, in comparison with quantitative coronary angiography (QCA) and CTA consensus reading, and 2) (semi-)automatically segment the coronary lumen on CTA, in comparison with expert's manual annotation. A database consisting of 48 multicenter multivendor cardiac CTA datasets with corresponding reference standards are described and made available. The algorithms from 11 research groups were quantitatively evaluated and compared. The results show that 1) some of the current stenosis detection/quantification algorithms may be used for triage or as a second-reader in clinical practice, and that 2) automatic lumen segmentation is possible with a precision similar to that obtained by experts. The framework is open for new submissions through the website, at *http://coronary.bigr.nl/stenoses/*.

*Keywords:* standardized evaluation framework, coronary arteries, Computed Tomography Angiography (CTA), Quantitative Coronary Angiography (QCA), stenoses, detection, quantification, lumen segmentation, multicenter, multivendor

*Corresponding author. P.O. Box 2040, 3000 CA Rotterdam, the Netherlands.
*Email address:* coronarystenoses@bigr.nl (T. van Walsum)
*URL:* www.bigr.nl (T. van Walsum)

## 1. Introduction

Coronary artery disease (CAD) is a major cause of death worldwide (Roger et al., 2012). Oxygen and nutrients, which are required for normal heart function, are supplied to the myocardium (the muscular tissue responsible for the contraction of the heart) by the blood traveling through the coronary arteries. If a coronary artery becomes narrowed or occluded owing to the build-up of plaque (e.g. calcium, fat and cholesterol), the amount of blood flowing to the myocardium is reduced and, thus, less oxygen and nutrients are delivered to these myocardial regions. The restriction in blood and oxygen is called *ischemia*; *atherosclerosis* is the condition in which plaques build-up in the coronary artery, and the narrowing of a vessel is referred to as *stenosis*. Atherosclerotic plaques can either be stable or unstable (also called vulnerable); the latter are prone to rupture (Virmani et al., 2006; Achenbach, 2008). Stable atherosclerotic plaques may cause temporary changes to ischemic myocardial regions, while unstable/vulnerable atherosclerotic plaques may induce irreversible defects to the myocardium, and result in *myocardial infarction* (heart attack). Though identification of stenoses prone to cause ischemic events through rupture is difficult (Achenbach, 2008), it is crucial to detect coronary artery plaques in an early stage.

Various cardiovascular imaging techniques are used to assess and quantify the presence and state of coronary artery stenoses. The choice of which cardiovascular imaging techniques to perform is determined by the patient's history and current symptoms. In current clinical practice, conventional coronary angiography (CCA) is the gold standard imaging technique to diagnose CAD. With CCA, the location, number and severity of the stenoses can be assessed. Computed tomography coronary angiography (CTA) is gaining popularity (Weustink and de Feyter, 2011). From 2006 to 2008, the number of coronary CTA scans (with and without quantitative evaluation of coronary calcium) performed in the U.S. has doubled, growing from 35,578 to 71,122 utilizations (Medicare, Shaw et al. (2010)). CTA is less invasive than CCA, provides high-resolution three-dimensional (3D) images of the cardiac and coronary artery anatomy, and allows the interpreter to assess the presence, extent and type (calcified or non-calcified) of coronary plaques. CTA has evolved as a reliable gatekeeper of CCA in patients with low to intermediate pre-test probability of CAD[1](Achenbach et al., 2012).

CTA images are currently interpreted using several visualization techniques (Raff et al., 2009). Transaxial image stacks are the basic visualization mode, and consist of a series of 2-dimensional (2D) axial images stacked in the longitudinal (i.e. cranio-caudal) direction. Such a visualization is characterized by minimal distortion and maximum resolution; however, 3-dimensional (3D) anatomical information, such as the coronary artery lumen morphology, is to be "mentally" reconstructed by the interpreter. As a complement, (curved) multi-planar reformatted (MPR/cMPR) images permit to visualize the coronary artery in orthogonal and oblique planes; such visualizations are especially recommended to delineate the morphology of the lumen (Figure 1). Maximum intensity projection (MIP) images may also be used to visualize vessels that run out of a given plane; MIP images are obtained by projecting the voxels with maximum intensity within a slab volume onto a plane. Currently, eyeballing (visual inspection and quantification) of CTA is the standard procedure in clinical practice to assess the coronary arteries.

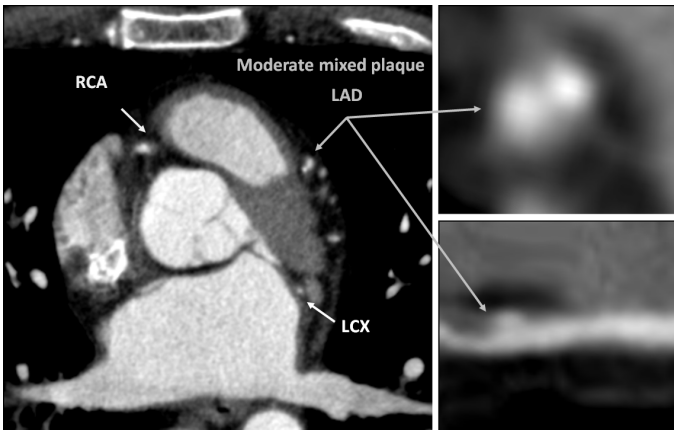



Figure 1: Visualization techniques used to interpret CTA images: transaxial (left) and multi-planar images (cross-sectional view at the upper-right, longitudinal view at the bottom right). Left anterior descending artery (LAD), right coronary artery (RCA) and left circumflex artery (LCX). This patient (training dataset #05) presents a moderate mixed plaque in segment #8 of the LAD.

The CTA interpretation is then summarized into a report(Raff et al., 2009), which contains, beside patient's clinical data, technical procedure information (i.e. image acquisition, image quality), clinical scan findings and interpretation. For each coronary artery lesion present in one of the modified 17-AHA-segments (American Heart Association, Fig. 2), the interpreter reports: 1) the stenosis location (origin, proximal, mid, distal, end) 2) the stenosis severity (mild, moderate, severe, occluded), 3) the stenosis plaque type (non-calcified, mixed, calcified), 4) the overall image quality / artifacts, and 5) the confidence in the interpretation. The final clinical decision making is based on these reported coronary findings; it is thus crucial to accurately detect or rule out significant CAD on CTA. Various studies investigated the diagnostic accuracy of CTA as compared to CCA (Meijboom et al., 2008; den Dekker et al., 2012). It has been demonstrated that CTA is 1) highly sensitive for detecting and ruling out significant CAD, and 2) moderately specific, even with severe coronary calcification (64-slices scanners and above).

The purpose of our work is to investigate to what extent automated approaches can be used to interpret cardiac CTA data for the presence of CAD. This paper has two main contributions: first, we introduce a framework to evaluate (semi-)automatic methods for coronary artery stenosis detection and quantification, and lumen segmentation, and second, we report on the results of this evaluation framework, comparing several state-of-the art coronary artery stenosis detection, quantification and segmentation algorithms.

[1]Pre-test probability of obstructive CAD are estimated using the Duke risk score (Pryor et al., 1993)

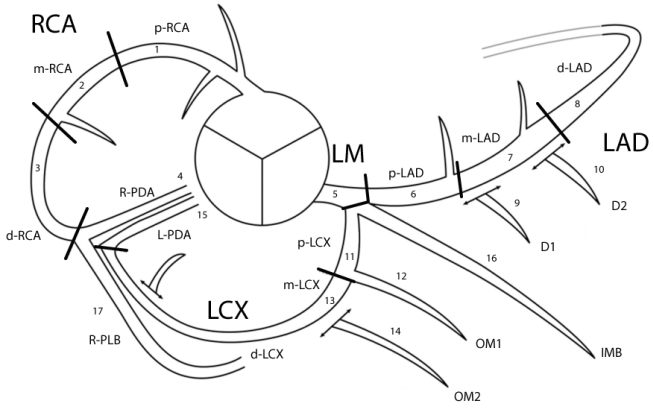



Figure 2: Coronary segmentation diagram - Axial coronary anatomy definitions derived, adopted, and adjusted from Austen et al. (1975)

Table 1: Quantitative stenosis grading and stenosis types

| Grade | | Description |
|---|---|---|
| 0 | Normal | Absence of plaque and no luminal stenosis |
| 1 | Mild | Plaque with 20% - 49% stenosis |
| 2 | Moderate | Plaque with 50% - 69% stenosis |
| 3 | Severe | Plaque with 70% - 99% stenosis |
| 4 | Occluded | Complete occlusion of the lumen |

| Type | Description |
|---|---|
| Non-calcified | Plaque without calcium |
| Calcified | Plaque with $\geq$ 50% calcium |
| Mixed | Plaque with $\leq$ 50% calcium |

In Section 2, we discuss previous work on detection and quantification of stenoses in CTA images. The evaluation framework is presented in Section 3. It includes a publicly available multicenter multivendor database of CTA data (Section 3.1), two reference standards derived from CCA and CTA (Section 3.4), a set of well-defined evaluation measures (Section 3.5), and an on-line tool to compare methods' performances (3.9). Section 4 gives a description of the first use of the framework during a MICCAI workshop, and includes a short description of the methods that were tested. The results of these methods as produced by the framework are presented in Section 5 and discussed in Section 6. Concluding remarks are made in Section 7.

## 2. Previous work

Here, we give an overview of the previously published stenosis detection, quantification and grading methods, and report how they were evaluated; we refer readers to Lesage et al. (2009) for an extensive review on vessel lumen segmentation methods.

Recently, the number of publications presenting and/or evaluating coronary artery stenosis detection and quantification techniques in cardiac CTA datasets is growing, thus increasing the need for a standardized evaluation framework.

Table 2 gives an overview of the previously published methods, evaluated against CTA and/or QCA. As presented in Figure 3, these methods can be categorized into two groups: 1) the ones that use accurate lumen segmentation together with either an intensity threshold or an estimation of the healthy vessel diameter to detect stenoses (Wesarg et al., 2006; Khan et al., 2006; Saur et al., 2008; Zhou et al., 2010; Kelm et al., 2011; Arnoldi et al., 2010; Halpern and Halpern, 2011; Xu et al., 2012) , and 2) the ones that use feature extraction computed along a centerline to directly detect plaque (Teßmann et al., 2009; Mittal et al., 2010; Zuluaga et al., 2011). Note that the latter methods focus on *plaque* detection rather than *stenosis* detection. In their evaluation stage, binary (healthy or diseased) labels were assigned to each cross-section by the observers, based on the presence of *plaque* rather than based on the presence/severity of lumen narrowing.

Most algorithms were quantitatively evaluated mainly on their detection rate (i.e. how accurately can a significant stenosis be detected by the algorithm); two articles (Halpern and Halpern, 2011; Xu et al., 2012) introduced more granularity (more grades) in the stenosis quantification. Moreover, solely three algorithms (Khan et al., 2006; Boogers et al., 2010; Halpern and Halpern, 2011) were compared to QCA.

To the best of our knowledge, the only commercially available system that automatically detects significant coronary artery stenosis in CTA is the COR Analyzer (Rcadia Medical Imaging Ltd., Haifa, Israel). The summary of 14 clinical trials evaluating the system is available in Goldenberg and Peled (2011). The system reports location and type (calcified, soft, mixed) of significant coronary lesions ($\geq$ 50% stenosis). It is positioned as a computer-aided simple triage (CAST) system (Goldenberg et al., 2012) to rule out significant coronary artery disease. It may also serve as a second opinion diagnostic aid and as a prioritization tool for high volume practices. The QAngio CT RE system (Medis Specials, Medis Medical Imaging bv, Leiden, the Netherlands; *www.medisspecials.com*) is commercially available, but is currently used for research purposes. This system has been evaluated in Boogers et al. (2010, 2012); it addresses the three tasks (detection, quantification, lumen segmentation) in a fully automatic fashion, but is intended to be used with minimal user interaction.

Since a few years, the number of initiatives that set up a publicly available evaluation framework in the medical image analysis community is growing (*http://www.grand-challenge.org/*). For instance, in the cardiovascular domain, Schaap et al. (2009a) and Hameeteman et al. (2011) successfully compared algorithms for coronary artery centerline extraction (*http://coronary.bigr.nl/centerlines*) and for carotid artery lumen segmentation and stenosis grading (*http://cls2009.bigr.nl/*) in CTA datasets. Up to now, no standardized evaluation methodology has been published to reliably evaluate and compare the performance of existing or newly developed stenosis detection/quantification and lumen segmentation algorithms. The proposed evaluation framework will provide such a large-scale standardized evaluation methodology and reference database.

Table 2: Overview of the previously published stenoses detection, quantification and grading methods. The analyses were performed in at least the 4 main arteries (left main, LAD, LCX, RCA), and possibly in the first-order arterial branches (diagonal, ramus, obtuse marginal, or posterior descending artery). The reported evaluation measures were computed *lesion-based*. TP, FP, FN, TN are the true positive, the false positive, the false negative and the true negative detections; PPV and NPV are the true positive value and false positive value; sens. and spec. refer to sensitivity and specificity and acc. to the accuracy.

| **Article** | **Patients/ Observers** | **Reference** | **Quantification?** | **Type** | **Used evaluation measures** |
|---|---|---|---|---|---|
| Wesarg et al. (2006) | 10/1 | CTA | - | Calcified | TP, FP, FN |
| Khan et al. (2006) | 50/1 | CTA/QCA | ≥ 50% | All | sens., spec. |
| Saur et al. (2008) | 127/1 | CT/CTA | - | Calcified & mixed | TP, FP, PPV |
| Teßmann et al. (2009) | 45/1 | CTA | - | All | TP, FP, FN, PPV |
| Mittal et al. (2010) | 165/1 | CTA | - | Calcified | PPV |
| Arnoldi et al. (2010) | 59/2 | QCA | ≥ 50% | All | sens., spec., acc., PPV, NPV |
| Zhou et al. (2010) | 20/2 | CTA | ≥ 50% | All | FP, FN |
| Halpern and Halpern (2011) | 207/1 | CTA | 3 grades | All | TP, FP |
| Kelm et al. (2011) | 229/3 | CTA | ≥ 50% | Non-calcified | sens., FP |
| Boogers et al. (2010) | 100/1 | CTA/QCA | All | All | Bland-Altman (% stenosis) |
| Zuluaga et al. (2011) | 9/2 | CTA | - | All | sens., spec., acc., Kappa |
| Xu et al. (2012) | 13/3 | CTA | 4 grades | All | Kappa |

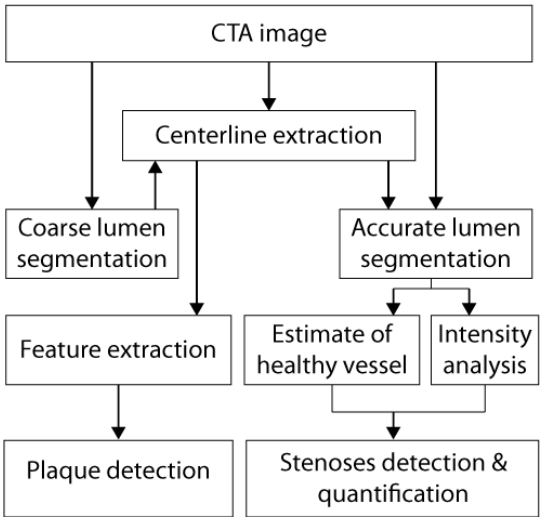



Figure 3: Overview of the building blocks and workflow of the previously published algorithms for coronary artery plaque detection and stenosis detection & quantification in CTA images.

## 3. Evaluation framework

In this section, we describe the datasets, the reference standards, the evaluation measures, as well as the ranking, used in our evaluation framework.

### *3.1. Cardiac CTA data*

*Study design.* The study was designed to include image data of symptomatic patients who presented either stable or unstable anginal syndromes, and who underwent both CTA and CCA examinations. Datasets were retrospectively acquired in three university hospitals, and evaluated anonymously. Thus, no IRB approval was required, according to the ethics committee guidelines of the involved medical centers: the Erasmus University Medical Center (ErasmusMC, Rotterdam, the Netherlands), the University Medical Center Utrecht (UMCU, Utrecht, the Netherlands) and the Leiden University Medical Center (LUMC, Leiden, the Netherlands).

*Patient selection.* Patients were selected such that they are representative of the population undergoing CTA examination for the assessment of obstructive CAD. According to the AHA guidelines (Budoff et al., 2006) and to the alternative diagnostic algorithm of Weustink and de Feyter (2011), patients with a low to intermediate pre-test probability of disease and an Agatston coronary calcium score (CCS) between 0 and 400 are, in current clinical practice, likely to undergo a CTA test. Therefore, patients were selected based on their CCS, and distributed over five CCS risk categories (Table 3); the number of patients included in each category was derived from the work of Nieman et al. (2009).

Our study population consists of 48 symptomatic patients, aged between 41 and 80 years old (58.76 ± 8.71 y.o.), enrolled in three university hospitals between June 2005 and June 2011; patients' characteristics are listed in Table 4. Patients with a previous history of percutaneous coronary stent placement, coronary artery bypass surgery, pacemaker, an impaired renal

Table 3: Distribution of patients (percentage of males) per coronary calcium score (CCS) category and per vendor. CCS refers to the Agatston score. The distribution of patients over the CCS categories was deduced from the work of Nieman et al. (2009), who reported on incidence of the different groups.

| Center | Vendor | Scanner | CCS | | | | | Total N (% males) |
|---|---|---|---|---|---|---|---|---|
| | | | 0 *Low* | 0.1-10 *Minimal* | 11-100 *Mild* | 101-400 *Moderate* | +400 *High* | |
| **EMC** | **SIEMENS** | Somatom Definition | 6 (100%) | 1 (100%) | 3 (80%) | 4 (50%) | 2 (50%) | 16 (75%) |
| **UMCU** | **PHILIPS** | Brillance 64 | 3 (33%) | 3 (66%) | 5 (80%) | 3 (33%) | 2 (50%) | 16 (56%) |
| **LUMC** | **TOSHIBA** | Aquilion ONE 320 | 2 (50%) | 2 (0%) | 6 (80%) | 4 (75%) | 2 (50%) | 16 (68%) |
| | | **All** | 11 (72%) | 6 (50%) | 14 (78%) | 11 (55%) | 6 (50%) | 48 (67%) |

function (serum creatinine ≥120 $\mu$mol/l), persistent arrhythmias, an inability to perform a breath hold of 15 s, a known allergy to iodinated contrast material, or a CTA of non-diagnostic image quality (motion artifacts) were excluded from our study.

*Scan protocol.* The CTA data was acquired on : 1) a dual-source CT scanner (Somatom Definition, Siemens, Forchheim, Germany) at the ErasmusMC, 2) a 64-slice CT scanner (Brillance 64, Philips Medical Systems, Best, the Netherlands) at the UMCU, and 3) a 320-slice CT scanner (Aquilion ONE 320, Toshiba Medical Systems, Tokyo, Japan) at the LUMC. The effective radiation dose was 11.3 ± 4.3 mSv for Siemens data, 18.4 ± 3.2 mSv for Philips data, and 3.8 ± 1.8 mSv for Toshiba data. A non-enhanced CT scan was performed before the CTA; the total calcium scores of all patients were calculated using dedicated software in each center. A bolus-tracking technique was used to synchronize the start of image acquisition with the arrival of contrast agent in the coronary arteries.

*Image reconstruction.* A single image per patient was used, reconstructed at the mid-to-end diastolic phase (350 ms before the next R-wave or at 65% to 70% of the R-R interval), with either retrospective (Siemens and Philips data) or prospective (Toshiba data) electrocardiographic gating.

### *3.2. Training and testing datasets*

Eighteen of the 48 CTA images, together with the CTA and CCA reference standards, were made available for training; the remaining thirty datasets were used for testing the algorithms; for those, only the CTA images were made available. The training and testing datasets were selected with respect to the different vendors, the CCS categories, and the disease prevalence, i.e. distribution of stenoses over the different degrees and coronary arteries; the distribution is shown in Figure 4 and Table 5. The 26% and 32% of the lesions are significant (≥ 50% luminal narrowing) for training and test datasets respectively.

### *3.3. Sub-challenges*

In our framework, three sub-challenges are defined: 1) coronary artery stenosis detection, 2) coronary artery stenosis detection & quantification, and 3) coronary artery stenosis detection & quantification and coronary artery lumen segmentation. Coronary artery stenosis detection is a mandatory task, as it is the focus of the evaluation framework. As some of the methods can also output, next to the stenosis detection, the stenosis grade and/or the lumen segmentation, we additionally provide the possibility to evaluate those two outputs.

Table 5: Distribution of the coronary artery lesions (≥ 20%) for the training and testing datasets. A lesion is considered as being significant if the luminal narrowing is ≥ 50%.

| | Artery | | | | |
|---|---|---|---|---|---|
| | RCA | LAD | LCX | IMB | **All** |
| **Training CTA** | | | | | |
| ≥ 20% | 36 | 51 | 12 | 4 | 103 |
| ≥ 50% | 12 | 10 | 5 | 0 | 27 |
| **Testing CTA** | | | | | |
| ≥ 20% | 50 | 73 | 18 | 2 | 143 |
| ≥ 50% | 18 | 22 | 7 | 0 | 47 |

Generally, semi-automatic algorithms may be used as aids for visual inspection of studies by clinicians; therefore, mainly accurate stenosis quantification is important. Fully automatic systems, on the other hand, may be used for triage, and therefore, should be able to identify patients without CAD with high specificity (usually above 60%, to not overwhelm the expert with a considerable amount of false positive detections and speed-up the diagnostic process), while maintaining very high sensitivity (usually above 90%). Every miss would then result, in the best case, in a delayed treatment for the patient.

### *3.4. Reference standard*

#### *3.4.1. Reference standard from CTA*

The multicenter multivendor CTA scans were analyzed at the Erasmus MC, University Medical Center Rotterdam (Rotterdam, the Netherlands).

*Stenoses detection/quantification.* Three independent experienced observers (A.S.D., W.B.M., S.L.P.), unaware of the results of the CCA, graded the CTA datasets; a unique reference standard was then derived from the three observers' grades following the protocol outlined in Figure 5.

A dedicated tool implemented in MeVisLab was used (*http://www.mevislab.de*) by the observers for the annotations. The axial source images, as well as MPR and cMPR views, were used to evaluate the CTA datasets for the presence of

Table 4: Patient's information

| | All | EMC | UMCU | LUMC |
|---|---|---|---|---|
| **Scan date** | | | | |
| min | 06/2005 | 08/2006 | 06/2005 | 06/2008 |
| max | 06/2011 | 12/2008 | 06/2006 | 06/2011 |
| **Age** | | | | |
| mean ± std | 58.76 ± 8.71 | 58.81 ± 11.05 | 57.31 ± 7.25 | 60.17 ± 7.05 |
| [min, max] | [41, 80] | [43, 80] | [41, 69] | [52, 74] |
| **Gender** | | | | |
| males (%) | 32(67%) | 12(75%) | 11(69%) | 10(63%) |
| **CV risk factors** | | | | |
| Obesity | 6 (13%) | 0 (0%) | 3 (19%) | 3 (19%) |
| Smoking | 20 (42%) | 3 (18%) | 8 (50%) | 9 (56%) |
| Hypertension | 21 (44%) | 7 (44%) | 6 (38%) | 8 (50%) |
| Diabetes | 5 (10%) | 0 (0%) | 1 (6%) | 4 (25%) |
| Fam. Hist. | 23 (48%) | 9 (56%) | 6 (38%) | 8 (50%) |

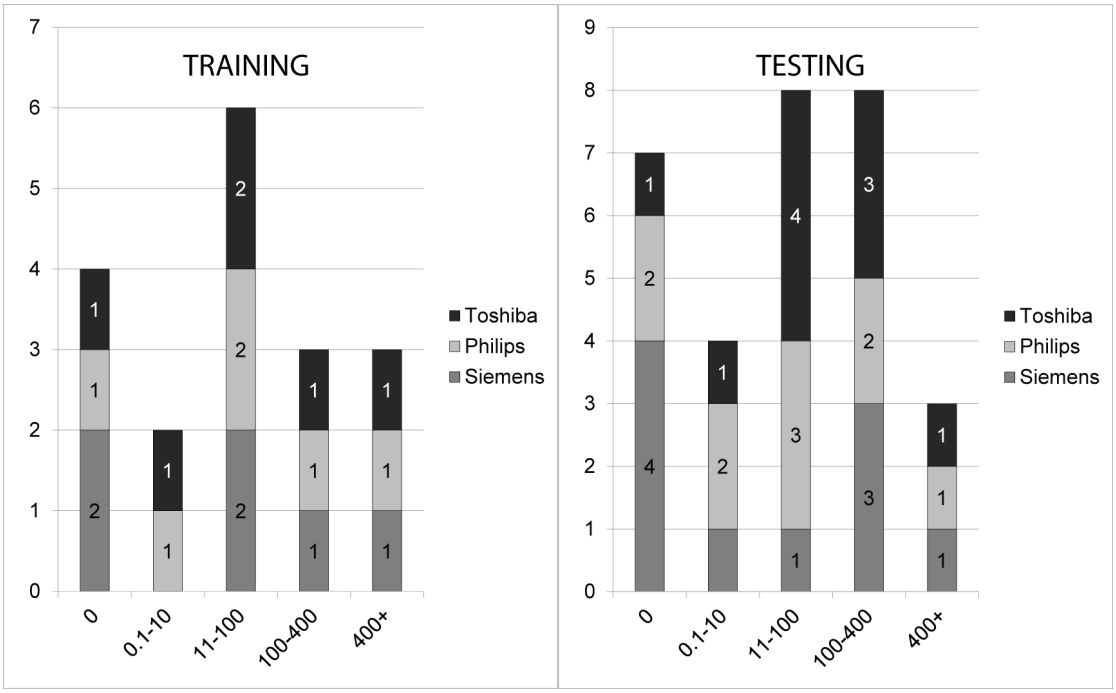



Figure 4: Distribution of 18 training and 30 testing datasets with respect to the different CCS categories and vendors.

coronary obstructions (i.e. lesions with ≥ 20% luminal narrowing). For each lesion with visually ≥ 20% luminal narrowing, the observers had to report the location, the plaque type (non-calcified, calcified, mixed), and the degree, according to the categories of Table 1. All segments from the 17 modified AHA segment model (Figure 2), which are present and have a diameter greater than 1.5 mm, were included in the analysis. Finally, each segment was scored as having significant CAD if at least one stenosis with ≥ 50% luminal narrowing was reported during the visual assessment.

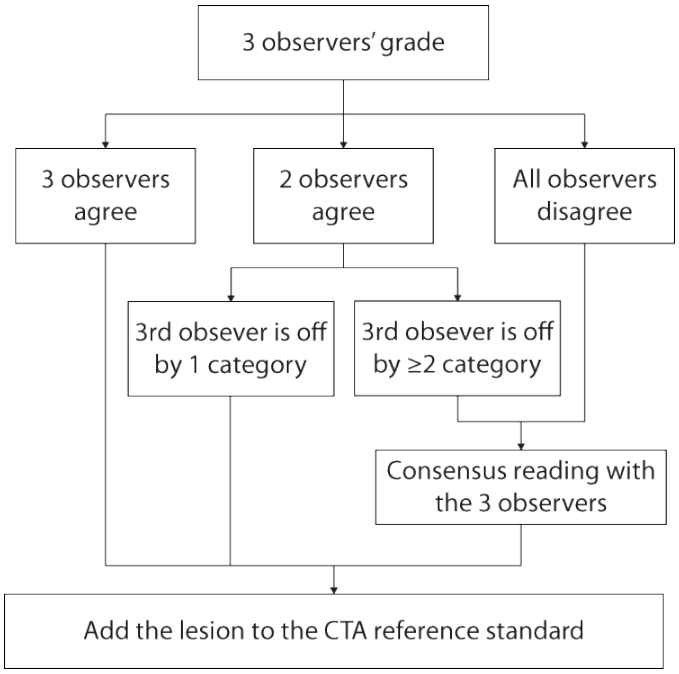



Figure 5: CTA stenoses detection/quantification reference standard protocol. Consensus reading with the 3 observers was necessary in 3% of the cases.

*Lumen segmentation.* Three independent experienced observers (A.S.D., W.B.M., A.D.), unaware of the results of the CCA, segmented a set of selected vessel segments in CTA images. All segments presenting a significant stenosis (≥ 50% luminal narrowing), as well as three additional segments (one in each of the main vessels) were randomly selected to be segmented; segments with a complete occlusion in the CTA stenoses detection/quantification reference standard were excluded. Figure 6 gives the details per vendor and per vessel of the number of segments being considered in the lumen segmentation evaluation. For the annotation process, we followed the same procedure as was used in the cls2009 challenge presented in Hameeteman et al. (2011). First, one observer annotated the centerline of each of the 17-segments of the modified AHA model (Figure 2) by clicking points in axial, sagittal and coronal views, followed by a centerline refinement step in cross-sectional views and cMPR images. Subsequently, using this centerline, three observers independently drew lumen contours in six cMPRs. These longitudinal contours were then used to construct cross-sectional contours on cross-sectional images sampled along the centerline. As a final refinement step, these cross-sectional contours could be manually edited. This procedure resulted in a set of cross-sectional contours along the vessel centerline, for each vessel segment selected and for each observer. These contours determine the reference standard for the evaluation of the lumen segmentations. An example of the CTA reference standard is presented in Figure 7.

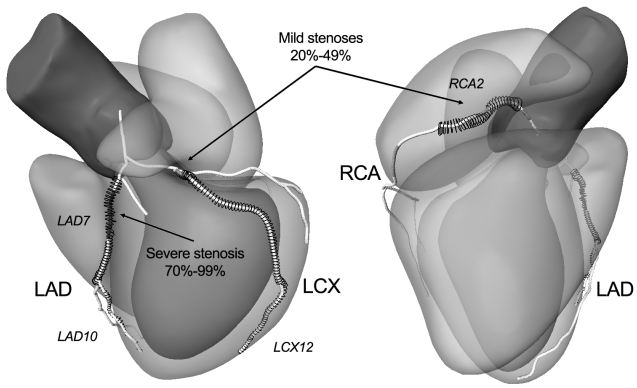



Figure 7: Example of CTA reference standard. Training dataset 08 presents five mild stenoses (one in RCA1, two in RCA2, one in RCA3, one in LCX12) and one severe stenosis (LAD7). Thus, segment LAD7 is selected to be segmented, as well as three other random segments in each of the main arteries, i.e. segments RCA2, LAD10 and LCX12.

### *3.4.2. Reference standard from CCA*

The reference standard from CCA for the detection and quantification of stenoses was obtained with quantitative coronary angiography (QCA). One experienced cardiologist (K.N.), unaware of the results of the CTA scoring results, identified and analyzed all coronary segments using the modified 17-segment AHA classification (Figure 2) on a separate workstation. Segments were visually classified as normal (smooth parallel or tapering borders, visually ≤20% narrowing) or as having coronary obstruction (visually ≥20% narrowing). The stenoses in segments visually scored as having ≥20% narrowing were quantified using the validated QCA algorithm (Cardiovascular Angiography Analysis System II, CAASII, Pie Medical Imaging Maastricht, the Netherlands) (Haase et al., 1993). Stenoses were evaluated in the worst (available) angiographic view (Figure 8) and classified as significant if the lumen diameter reduction exceeded 50%.

## *3.5. Evaluation measures*

The evaluation measures for the coronary artery stenosis detection and quantification are reported per coronary calcium category (Figure 4) and over all patients, as providing the errors per dataset may reveal information about the reference stenosis grades. The final evaluation measure (as reported in Table 9 and 10) is obtained over all patients. The evaluation measures for the lumen segmentation are communicated per patient, and the number and identity of evaluated segments remains hidden.

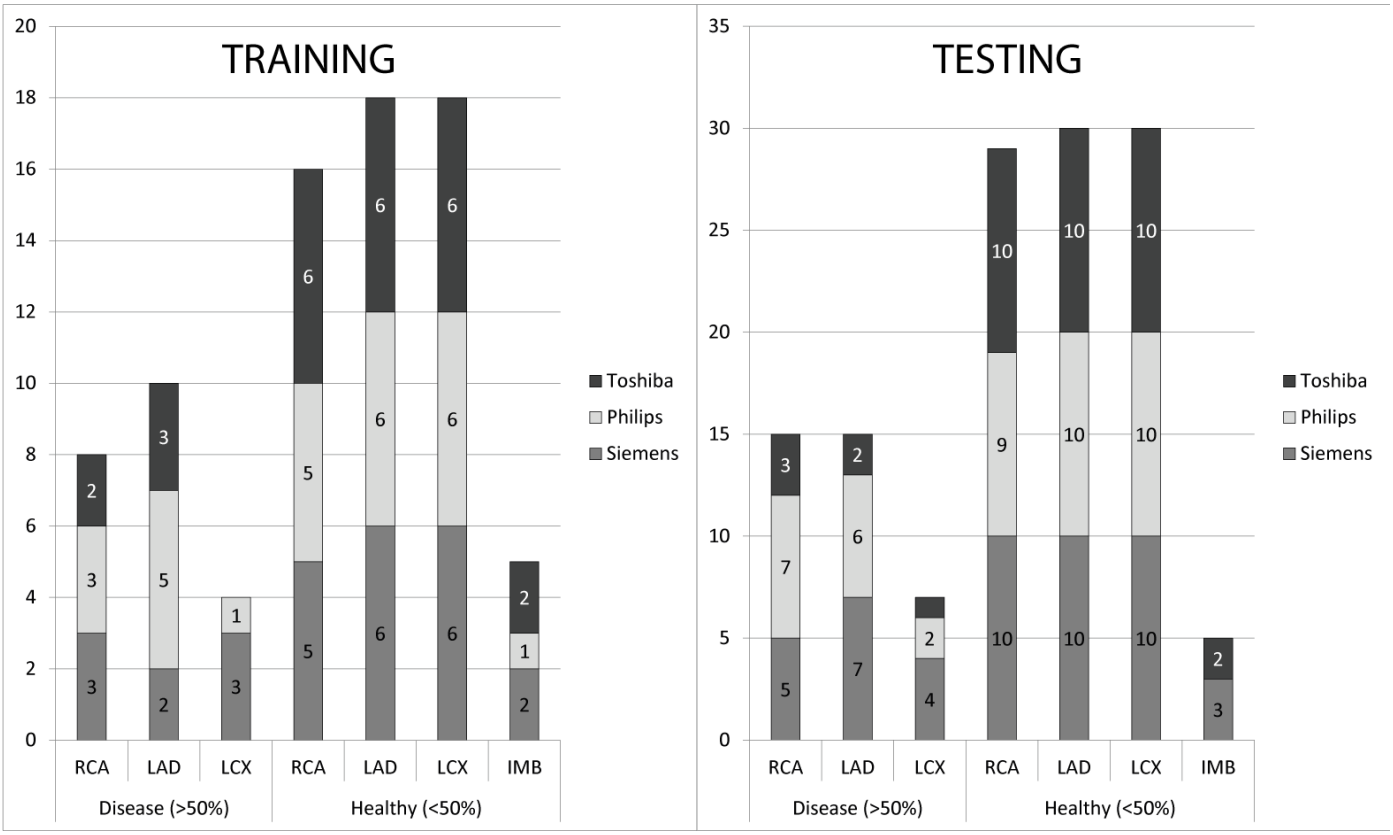



Figure 6: Overview of the segments considered for the lumen segmentation evaluation. *Diseased* segments are segments presenting in CTA consensus with at least one significant stenosis (≥50%). *Healthy* segments are segments presenting in CTA consensus with no significant stenosis (≤50%). Occluded segments were excluded from the lumen segmentation evaluation. The training set consists of 18 datasets and the testing set of 30 datasets.

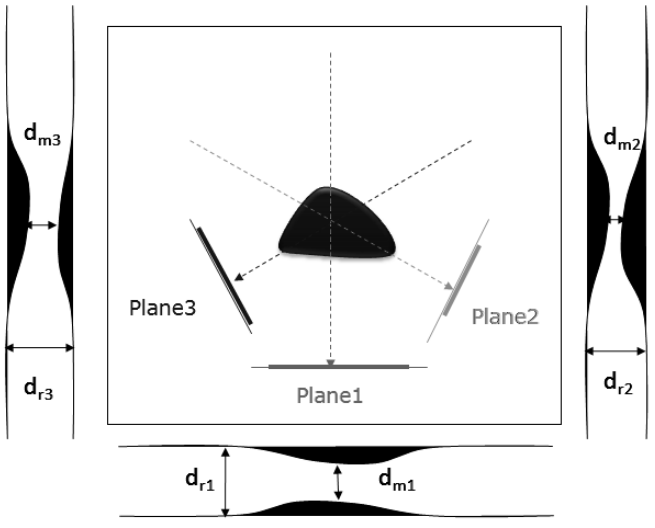



Figure 8: Quantitative coronary angiography (QCA). For each acquired X-ray imaging plane, the minimal luminal diameters ($d_m$) are measured and compared to the reference diameter ($d_r$) of the vessel immediately adjacent. Given the minimal (projected) diameter, the percentage of stenosis can be calculated, in plane 2 in the given example.

### *3.5.1. Stenosis detection*

Two metrics are used to evaluate the performance of the coronary artery stenosis detection algorithms: the sensitivity (Eq.(1)) and the positive predictive value (Eq.(2)).

$$S = \frac{TP}{TP + FN} \tag{1}$$

$$PPV = \frac{TP}{TP + FP} \tag{2}$$

where TP, FN, FP are the true positive, false negative and false positive detections, respectively. Table 6 defines the TP, FN, FP and TN.

The evaluation as compared to the CTA reference standard is *lesion-based*. The stenoses considered here are the union of the stenoses in the reference standard and in those detected by the algorithm. An example of stenosis detection is presented in Figure 9 for training dataset#10 and results of observer#1, and more details about the matching procedure can be found in the Appendix B.

The evaluation as compared to the CCA reference standard is *segment-based*. The segments considered here are all anatomically present segments from the modified 17-AHA-segments model (Fig. 2), with a minimal lumen diameter greater than 1.5 mm.

Table 6: Stenosis detection, as compared to CTA and CCA reference standard. Descriptions of true-positive (TP), false-negative (FP), false-positive (FP) and true-negative (TN) detection.

| **Detection** | **Description** for *segment-based* and *lesion-based* analysis |
|---|---|
| **TP** | Both the reference standard and the algorithm stenosis/segment have a grade ≥ 50%. |
| **FN** | The reference standard stenosis/segment has a grade ≥ 50% while the algorithm stenosis/segment has a grade < 50%. |
| **FP** | The reference standard stenosis/segment has a grade < 50% while the algorithm stenosis/segment has a grade ≥ 50%. |
| **TN** | Both the reference standard and the algorithm stenosis/segment have a grade < 50%. |
| | **Description** for *patient-based* analysis |
| **TP** | At least 1 significant stenosis in a patient detected by both the reference standard and the algorithm, regardless of location of stenosis |
| **FN** | No significant stenosis detected by the algorithm and at least 1 significant stenosis detected by the reference standard. |
| **FP** | Significant stenosis detected by the algorithm and no significant stenosis detected by the reference standard. |
| **TN** | No significant stenosis in a patient detected either by the reference standard and the algorithm. |

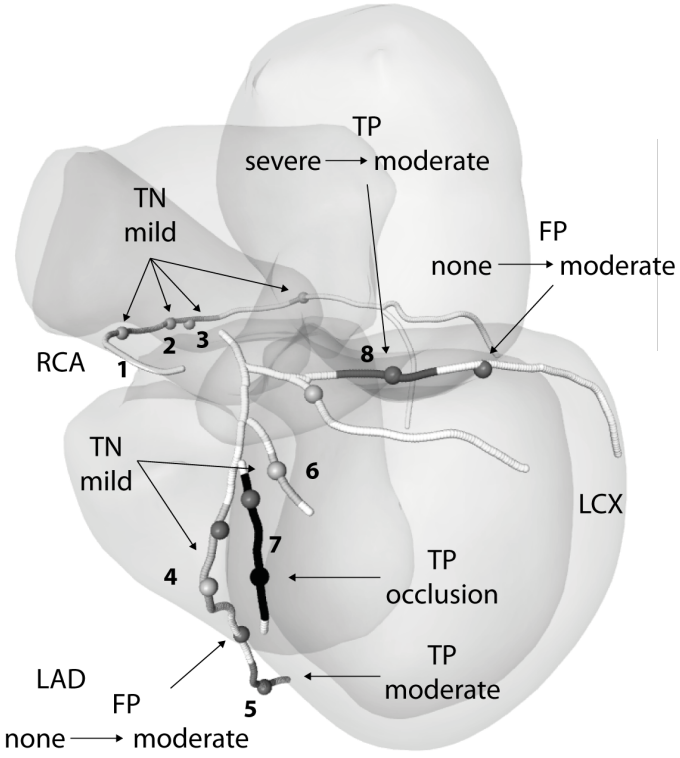



Figure 9: Detection example for training dataset10 and results of observer#1. The patient presents 8 stenoses (grey-scale segments) and the observer#1 detected 14 stenoses (grey-scale spheres). Reference stenoses #1, #2, #3, #4 and #6 are TN detections, i.e. the observer assigned a *mild* grade (≤ 50%). Reference stenoses #5 (severe) and #8 (moderate) are TP detections, i.e. the observer assigned a *moderate* grade (≥ 50%). Reference stenosis #7 (occlusion) is also a TP because the observer's average grade (*severe* + *occluded*) is ≥ 50%. The observer detected two FP *moderate* lesions in LAD and LCX. For more details about the grades, see Table 1; grey-scale from white to black correspond to *healthy*, *mild*, *moderate*, *severe* and *occlusion*.

### *3.5.2. Stenosis quantification*

As compared to the reference standard derived from CCA, two metrics are used to evaluate the performance of the coronary artery stenosis quantification algorithms, per segment: the absolute average difference (AAD, Eq.(3)) and the root mean squared difference (RMSD, Eq.(4)).

$$AAD = \frac{\sum_{i=1}^{S} \left| g_i - g_i^{ref} \right|}{S} \tag{3}$$

$$RMSD = \sqrt{\frac{\sum_{i=1}^{S} \left( g_i - g_i^{ref} \right)^2}{S}} \tag{4}$$

with $g^{ref}$ the reference standard stenosis grade, $g$ the estimated stenosis grade, and $S$ the number of considered segments in the evaluation.

When evaluating the performance of the coronary artery stenosis quantification algorithms per lesion as compared to the CTA reference standard, close misses (e.g. grading a stenosis as being mild while the reference standard indicates it is moderate) should be less heavily penalized than misses that are further apart (e.g. grading a stenosis as being severe or occluded while the reference standard indicates it is mild). Therefore, we use the linearly weighted Cohen's Kappa metric (Cohen, 1968). It measures how much different the observed agreement is from the expected agreement, and is standardized to take values between -1 and 1, where 1 is perfect agreement, 0 is exactly what would be expected by chance, and negative values indicate agreement less than chance, i.e. potential systematic disagreement between the reference standard and the algorithm.

To fairly compare the Kappa values of different algorithms, the Kappa must be computed using the same number of stenoses. However, in the CTA reference standard, true negative (TN) detections are not reported, while required to compute the Kappa value. We accommodate this issue by estimating an average TN per dataset as follows: given the 48 datasets, we measured a total arterial segments length of 22080 *mm*, among which 2120 *mm* are diseased sections (i.e. presenting lesions

with ≥ 20% obstruction). As there are 246 lesions, the average length of a lesion is of 8.6 *mm*. At the end, the 19960 *mm* of healthy vessel can be interpreted as 2321 TN, and thus, as 48 (TN+FP) per dataset. In the case that an algorithm would report (on average) more than 48 (TN+FP) lesions per dataset, its Kappa value is set to -1 (less than chance agreement), as exceeding this limit does not make sense. More details about the computation of the Kappa statistics are provided in Section 3.6.2.

#### *3.5.3. Lumen segmentation*

The segmentation is evaluated by comparing the result with the lumen contours that were manually drawn by three observers (see Section 3.4.1). The segmentations are compared to the reference standard using three measures: the Dice similarity index (Dice), the mean distance (MSD), and the Hausdorff distance (MaxD). Each metric is determined for each cross-sectional contour of the observer annotations, thus in 2D, and the results of all contours of a vessel segment are combined, yielding three scores per vessel segment per observer. To perform the 2D metric calculation, for each observer contour, the segmentation was intersected with the corresponding cross-sectional contour plane. From the intersection result (i.e. a set of segmentation contours), the segmentation contour closest to the observer contour and not too far away from it, is assumed to be the valid segmentation contour. If a valid segmentation contour is found, it is linearly resampled to ensure that the points along the contour are sufficiently close. Subsequently, the true positive area (overlap area), the false negative area (missed lumen) and the false positive area (segmentation outside lumen) are determined for this contour, by applying a 2D scan conversion algorithm to both contours. Also, the distances from each observer contour point to the segmentation contour, and vice versa, are determined and summed over the contours, and also the maximum distance is determined. True positive area, false negative area and false positive area are summed over all contours of the vessel segment, after which the Dice index is determined. Similarly, the mean squared distance is the average of all contours' mean squared distance of the vessel segment, and the Hausdorff distance is the maximum of the contours' Hausdorff distance. If no segmentation contour is sufficiently close to the observer contour, it is assumed that the segmentation does not contain this part of the vessel segment. In that case, the complete lumen area is counted as false negative area, the mean distance is equal to the mean radius of the manual contour, and the Hausdorff distance is equal to the maximum distance between the manual contour and its center.

### *3.6. Ranking the algorithms*

In order to rank the different algorithms for coronary artery stenosis detection, stenosis quantification and lumen segmentation, the evaluation metrics presented in the previous section have to be combined. This is achieved by first assigning to each algorithm a rank for each evaluation metric. The rank is between 1 (best) and $N$ (worst), $N$ being the number of observers and algorithms to be compared. The final rank is then obtained by averaging the ranks over the evaluation metrics.

It should be noted that it is possible for method A to have better average measures than method B, while still having a worse average rank.

#### *3.6.1. Stenosis detection*

The detection algorithms are ranked based on the overall sensitivity and positive predictive value achieved as compared to the CTA and CCA reference standards, as follows:

$$R_D = \frac{\text{rank}_{Sens}^{CCA} + \text{rank}_{PPV}^{CCA} + \text{rank}_{Sens}^{CTA} + \text{rank}_{PPV}^{CTA}}{4} \tag{5}$$

with $\text{rank}_{Sens}^{CCA}$ and $\text{rank}_{PPV}^{CCA}$ respectively the sensitivity and PPV ranks achieved over all data as compared to the CCA reference standard, and $\text{rank}_{Sens}^{CTA}$ and $\text{rank}_{PPV}^{CTA}$ respectively the sensitivity and PPV ranks achieved overall data as compared to CTA reference standard.

#### *3.6.2. Stenosis quantification*

The quantification algorithms are ranked based on their AAD and RMSD of the degree of stenosis as compared to the CCA reference standard (segment-based), and on their weighted Cohen's Kappa coefficient as compared to CTA reference standard (lesion-based), as follows:

$$R_Q = \frac{\text{rank}_{AAD}^{CCA} + \text{rank}_{RMSD}^{CCA} + 2 \cdot \text{rank}_{Kappa}^{CTA}}{4} \tag{6}$$

with $\text{rank}_{AAD}^{CCA}$ and $\text{rank}_{RMSD}^{CCA}$ respectively the AAD and RMSD ranks achieved over all data as compared to the CCA reference standard, and $\text{rank}_{Kappa}^{CTA}$ the linearly weighted Kappa rank achieved over all data as compared to the CTA reference standard. We added a weight of 2 to the $\text{rank}_{Kappa}^{CTA}$ to make the total weight for the CTA rank equal to the total weight for the CCA ranks.

#### *3.6.3. Lumen segmentation*

The segmentation algorithms are ranked based on the overlap, the mean distance, and the Hausdorff distance (average over the 3 observers' reference annotations), while making distinction between segments having non-significant and significant stenoses, as follows:

$$R_S = \frac{1}{N} \cdot \sum_{p=1}^{N} \left( \frac{\sum_{h=1}^{3} \text{rank}_h^p}{3} + \omega^p \cdot \frac{\sum_{d=1}^{3} \text{rank}_d^p}{3} \right) \tag{7}$$

where $\omega^p = 1$ if patient $p$ has evaluation metrics computed for diseased segments (i.e. segments with significant stenoses), and $\omega^p = 0$ otherwise, $N$ being the number of patients.

First, for each algorithm, a rank $\text{rank}_m^p$ is computed per patient $p$ and evaluation metric $m$. Then, an average rank $\text{rank}^p$ is obtained by averaging the three (if the patient does not present any significant stenosis) or six (if the patient presents at least one significant stenosis) $\text{rank}_m^p$ ranks; this leads to $N$ ranks. The final rank is obtained by averaging of the $N$ $\text{rank}^p$ patient ranks.

### *3.7. Algorithm categories*

Depending on the amount of user-interaction, we distinguish two different categories of algorithms:

*Fully automatic.* Fully automatic methods detect and quantify coronary artery stenosis and segment the lumen without user-interaction. The CTA image is the only input used by the method.

*Minimal user-interaction.* Methods with minimal-user interaction are allowed to use two additional points per vessel: 1) one point S at the ostium (start of the vessel), and 2) one point E at the end of each vessel. Points S and E are provided with the data.

### *3.8. Provided centerlines*

Coronary analysis methods often start with detecting a coronary centerline (see Section 2). To facilitate those methods that can do coronary analysis, but do not have a centerline extraction available, three teams of the centerline extraction challenge (Schaap et al., 2009a) were asked to provide centerline extraction results to the participants of this challenge: 1) automatic and manually corrected from the LKEB group (Leiden, the Netherlands) based on Yang et al. (2011, 2012), 2) automatic from Rcadia (Haifa, Israel) based on Goldenberg et al. (2012), and 3) automatic from VRVis (Vienna, Austria) based on Zambal et al. (2008). The participants can then use one of these set of centerlines as input for their method (as long as they use the same centerline extraction algorithm for all datasets) and submit the combined method to a category, depending on the automation of the used centerline extraction algorithm.

### *3.9. Web-based evaluation framework*

The proposed framework for coronary artery stenosis detection & quantification and lumen segmentation in CTA images is made publicly available through a web-based interface (*http://coronary.bigr.nl/stenoses/*). The 48 cardiac CTA datasets, as well as the corresponding stenosis detection, quantification and lumen segmentation reference standard of the training datasets, are available for download for anyone who wishes to validate their algorithm. Furthermore, the website provide several tools to inspect and compare the algorithms.

## 4. MICCAI 2012 workshop

The evaluation framework was launched during the "3D Cardiovascular Imaging: a MICCAI segmentation challenge" workshop that was organized in conjunction with the 15th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI), and was held on October 1st, 2012 in Nice Sophia Antipolis, Côte d'Azur, France. Around 200 groups from academia and industry were invited by e-mail to participate in the challenge. Seventy-five teams created an account on our website, fifty of which sent the data confidentiality form, which was required to download the datasets. Forty-four teams downloaded the training set and twenty-nine teams downloaded additionally the testing set. Eleven teams submitted results: eight of them participated in the quantification sub-challenge and five of them participated in the lumen segmentation sub-challenge. The 11 evaluated algorithms are described below and more details can be found in the full paper version that all authors submitted for the workshop, available on our website (*http://coronary.bigr.nl/stenoses/*).

### *4.1. Broersen et al.*

The algorithm by Broersen et al. (2012) has four stages. First, an automatic tree extraction (Yang et al., 2012) and segment labeling step is performed (Yang et al., 2011). Second, lumen and vessel wall contours are detected in each vessel (running from the ostia to the most distal point)(Boogers et al., 2012). Next, regions with potential lesions are automatically determined for each segment based on deviations from a regression on the lumen areas in the vessel representing normal vessel tapering. Additional lesions are detected in calcified regions as well as in regions with significant lumen intensity drops. Finally, the lesion information from all vessels is combined into a unique list of lesions.

### *4.2. Cetin et al.*

The algorithm by Cetin and Unal (2012) has four stages. First, the coronary vessels are extracted around the provided centerline coordinates (manually refined, based on Yang et al. (2011)) by the "vessel tractography" method presented in Cetin et al. (2013). Second, longitudinal vessel volumes are generated for each branch to provide rotation invariance. Third, the following features along the centerline of the vessel are extracted: intensity features based on a cylindrical sampling pattern with varying radii, length and position; and a geometric feature based on the energy of the vessel radius profile. Finally, a random forest based classifier is utilized to detect the stenosis coordinates along the vessel.

### *4.3. Eslami et al.*

The algorithm by Eslami et al. (2012) has three stages. First, the CTA is resampled with multiple cross sectional planes, employing the provided centerlines (manually refined, based on Yang et al. (2011)) to construct a cylindrical representation of the underlying vessel. Secondly, segmentation is performed using an inflating tube technique, starting from vessel centerline. Finally, stenoses are detected and quantified by comparing the extracted lumen area with the second order regression of the lumen area over the vessel length. Furthermore, the intensity of boundary voxels is contrasted against the intensity of left ventricle cavity and myocardium to take into account the partial volume effect.

### *4.4. Duval et al.*

The algorithm by Duval et al. (2012) has three stages. Firstly, five circular Regions-of-Interest (RoI) are extracted around each centerline point (automatic, based on (Goldenberg et al., 2012)). Secondly, for each extracted RoIs, thirteen features are calculated based on intensity and Haar-like features using suitably partitioned RoIs. To combat the inherent centerline detection error the same calculations are repeated on four neighbors of each centerline point. Thirdly, five independent random forests are used corresponding to the centerline point and its neighbors. The stenosis is deemed present if all the random forests are in agreement.

### 4.5. Flórez-Valencia et al.

The algorithm by Flórez Valencia et al. (2012) has three stages. First, an axis is extracted using Dijkstra's algorithm with costs calculated as in Gülsün and Tek (2008). Second, a tracking algorithm is used along this axis, smoothed by a Bezier curve, to construct a generalized cylindrical model of the artery from cross-sectional contours segmented using Fast-Marching (Baltaxe Milwer et al., 2007). The parameters of the model are deduced from these contours and corrected by a Kalman state estimator. Finally, stenoses are detected and quantified assuming that diameters of healthy arteries should decrease linearly from the ostium. Hence, the estimated diameters are compared to theoretical ones.

### 4.6. Lor et al.

The algorithm by Lor and Chen (2012) has three stages. First, a Gaussian distribution is utilized to adaptively locate the center of cross-sectional plaque with the variance of the posterior density as the plausible size. Such a concentric model is also applied to segment the vessel lumen. Second, the quantitative evaluation of diameter stenosis is determined using Kalman filtering. Finally, the stenosis degree is given using a Bayes classifier based on the posterior probability of severity conditioned on stenosis percentage and plaque type of the training data. The method was evaluated based on the centerline extracted by the vascular segmentation developed by Yang et al. (2012).

### 4.7. Melki et al.

The algorithm by Melki et al. (2012) has three main stages. First, the coronary arteries tree is constructed using the provided centerlines (automatic, based on (Goldenberg et al., 2012)). Common parts of the centerlines are merged in order to organize the whole tree in a set of disjoint segments. Second, a first step of stenosis candidate detection is applied using the vessel cross section area profile. Regions showing a deviation higher than 50% of a synthetic lumen area profile are flagged. Finally, they apply a false positive removal step in order to eliminate the erroneous candidates. This step is performed by inspecting the appearance properties of each flagged region.

### 4.8. Mohr et al.

The algorithm by Mohr et al. (2012) has four stages. First, initialization steps are performed including generating an orthogonal image stack from the provided centerlines (automatic, based on Goldenberg et al. (2012)) and estimating the lumen and wall segmentation. Second, calcium is identified using Bayes Information Criterion, and lumen and wall tissue is classified by Expectation-Maximization assuming Gaussian distributions in segments along the vessel. Third, the lumen segmentation is refined using a level-set driven by a speed function including the a posteriori probabilities obtained from the classification. Finally, stenoses are identified and quantified by estimating the expected vessel profile using line fitting.

### 4.9. Öksüz et al.

The algorithm by Öksüz et al. (2012) consists of five stages. First pulmonary vessels are removed with thresholding and morphological operations. Afterwards, Frangi vesselness filter (Frangi et al., 1998) is applied on the processed data. Vessel segmentation is realized by 3D region growing and fast marching, respectively. Plane fitting is performed on every centerline point (manually refined centerline, Yang et al. (2011)), where the corresponding vessel diameter is computed. Finally, a running window based median filtering followed by smoothing is applied, and nominal vessel diameter is estimated by linear regression. A positive difference between computed and estimated diameter values is defined as stenosis at that location.

### 4.10. Shahzad et al.

The algorithm by Shahzad et al. (2012) has three stages. First, centerlines are extracted using a two point minimum cost path approach (Metz et al., 2009) and are subsequently refined, after which bifurcations are detected. The resulting centerlines represent the vessel segments, and are used as an initialization for the lumen segmentation, performed using graph cuts and robust kernel regression (Schaap et al., 2009b). Finally, the expected diameter of the healthy lumen is estimated by applying robust kernel regression on the coronary artery lumen diameter profile; stenoses are subsequently detected and quantified by computing the difference between estimated and expected diameter profiles.

### 4.11. Wang et al.

The algorithm by Wang et al. (2012) has four stages. It uses an implicit model-guided level set method. First, a 3D vessel model from a set of initial centerlines (automatic, based on Goldenberg et al. (2012)) is generated. Second, this model is incorporated in the level set propagation to regulate the growth of the vessel contour. Third, new centerlines are extracted after evolving the level set and the diameter of vessels is re-estimated in order to generate a new vessel model. Finally, the propagation and re-modeling steps are repeated until convergence. For detecting and quantifying stenoses, the proposed method was run twice with different parameter settings to segment the outer wall and the lumen.

## 5. Results

The results presented in this section are based on the algorithms presented at the MICCAI'12 workshop (*http://coronary.bigr.nl/stenoses/*). Since the MICCAI'12 workshop, the ranking strategy has been modified. As a consequence, for each team that participated to the MICCAI'12 workshop, the public results available on our website are different from the one reported in their workshop paper.

Space limitations prevent us from incorporating more statistics here, but the on-line evaluation framework provides the possibilities to rank the methods on different measures or scores, and create statistics on a subset of the data (per vendor). The website also contains the most recent version of the results.

The on-line results can be different from the results reported in this paper, as new submissions or method improvements may have occurred.

In this section, result tables also contain the results of the observers, which have been scored in the same ways as the other evaluated methods. It should be noted that, as the CTA reference standard was derived from a consensus reading of the same 3 observers, the observers' performance for coronary stenoses detection and quantification as compared to CTA reference standard may be biased to their advantage.

### *5.1. Detection of stenosis*

The ability of a method to discriminate significant stenoses from non-significant ones is evaluated. Table 9 shows the average results and ranking of the 11 submissions (5 fully automatic, 6 semi-automatic), 3 observers and their consensus for stenosis detection measures: sensitivity and PPV. In the overall ranking, the algorithms of Cetin and Unal (2012) and Mohr et al. (2012) rank the first, in the semi-automatic and automatic category respectively.

As compared to QCA (*segment-based* analysis), the best sensitivity (68%) was achieved by the method proposed by Eslami et al. and the best PPV (50%) was obtained with the algorithm of Wang et al..

As compared to CTA (*lesion-based* analysis), the best sensitivity (55%) was achieved by the method proposed by Shahzad et al. and the best PPV (33%) was obtained with the algorithm of Wang et al.. Here, the results were worse than the average observers' performance (sensitivity of 73%, PPV of 67%). With respect to the CTA reference standard and over all calcium categories, the approach of Mohr et al. tends to over-estimate the degree of mild stenoses, thus increasing the number of FP detections, and under-estimate the degree of significant stenoses (especially moderate ones), thus increasing the number of FN detections and, consequently, penalizing the sensitivity. Their QCA sensitivity (57%) was less affected, probably because the degree of stenosis is generally over-estimated in CTA as compared to QCA (calcified lesions, due to blooming artifact), which compensates for under-estimation on CTA.

In addition, Table 7 presents the performance of the methods in terms of TP, FP, FN and TN detections, with respect to CCA and CTA reference standard. Overall, a good TP detection rate was achieved at the expense of FP and/or FN rates, and vice-versa. Note that for many of the methods, the number of reported FP was very large; these methods are therefore not yet suitable for implementation in clinical practice (risk of overwhelming the clinician). The current results highlight that discrimination between significant and non-significant lesions remains a challenge and that a trade-off between the ability to detect significant lesions and the ability of ruling out disease needs to be made.

Last, Table 8 presents the diagnostic performance of the methods, observers and consensus for the detection of significant stenosis on QCA and CTA in a *per-patient* analysis. The sensitivity, specificity, PPV and NPV, with respect to CCA and CTA reference standard are reported. The results indicate that half of the methods (Eslami et al.; Flórez Valencia et al.; Lor and Chen; Melki et al.; Mohr et al.) are not yet able to perform triage of the patients to rule out significant coronary artery disease, as they achieve very low specificity. Four methods (Broersen et al.; Duval et al.; Shahzad et al.; Wang et al.) perform relatively good as compared to the observers. The last two methods (Cetin and Unal; Öksüz et al.) have a diagnostic performance close to the observer's one, as well as sensitivity close to 90% and high specificity; they thus may be considered to be used as computer-aided triage systems, or as a second reader, where a very high sensitivity is required and false positives are reasonably acceptable.

### *5.2. Quantification of stenoses*

Less-obstructive plaques outnumber severely obstructive plaques (Falk et al., 1995), and most occlusions result from progression of the former plaques. It is thus as crucial to detect mildly to moderately obstructive lesions (20% to 70%) as to detect severely obstructive plaque ($\geq$ 70%). We therefore investigated the ability of a method to correctly estimate the degree of obstruction.

Table 10 shows the average results and ranking of 8 submissions (3 fully automatic, 5 semi-automatic) and 3 observers for stenosis quantification measures. Here, the quantification measures are computed using the union of the submitted lesions and the reference ones (i.e. including not only the TP, but also the FP and FN) to assess the whole system accuracy. For the Kappa statistic, a fixed number of negative detection $TN$ is used, and is determined as follows:

$$\mathrm{TN} = \mathrm{N} \times 48 = \mathrm{TN}_{\mathrm{algorithm}} + \mathrm{FP}_{\mathrm{algorithm}} \tag{8}$$

with $N$ the number of datasets, and $\mathrm{TN}_{\mathrm{algorithm}}$ and $\mathrm{FP}_{\mathrm{algorithm}}$ the true negative and false positive detections of the algorithm respectively.

First, as the observers quantified the coronary stenoses using semi-quantitative grades (see Section 3.4.1 and Table 1), their grades were converted to quantitative values (number between 0 and 100) for the comparison with QCA: a stenosis reported as being mild on CTA was assigned to be 35% on QCA, moderate to be 60%, severe to be 80% and occluded to be 100%. This explains the relatively large observers' errors with respect to QCA.

Second, some methods detect a large number of FP stenoses in CTA (Table 7); this is consequently expressed by negative or nul Kappa values (algorithms of Eslami et al., Lor and Chen and Flórez Valencia et al.). Identically, detecting too many FP in QCA would penalize the algorithm, by increasing their average absolute and root mean square differences.

The method proposed by Shahzad et al. achieves the best quantification results as compared to QCA, with an averaged absolute difference of 21% and a RMS difference of 29%. This method outperforms the observers and all the other methods. This is due to their low number of FP detections.

The method of Shahzad et al. also achieves the best performance with regard to the Kappa value ($\kappa = 0.28$). Though their Kappa value is positive, it remains relatively low. This may

Table 7: Performance of the 11 evaluated methods, 3 observers and their consensus for the detection of coronary artery stenoses (≥ 50% diameter reduction) on the 30 testing datasets. True positive (TP), false positive (FP), false negative (FN), true negative (TN), average false positive detection per patient (FP/pat). QCA analysis is *segment-based*; 394 segments evaluated, prevalence of disease is 7%. CTA analysis is *lesion-based*. The values in bold correspond to the best performance for each measure. Methods are listed by alphabetic order.

| **Method** | **QCA** | | | | **CTA** | | | |
|---|---|---|---|---|---|---|---|---|
| | TP | FP | FN | TN | TP | FP | FN | FP/pat |
| *CTA consensus* | *23* | *21* | *5* | *345* | *47* | *0* | *0* | *0* |
| *Observer 1* | *24* | *36* | *4* | *330* | *39* | *25* | *8* | *0.8* |
| *Observer 2* | *21* | *20* | *7* | *346* | *33* | *8* | *14* | *0.3* |
| *Observer 3* | *18* | *24* | *10* | *342* | *31* | *21* | *16* | *0.7* |
| Broersen et al. | 7 | 30 | 21 | 336 | 13 | 29 | 34 | 1.0 |
| Cetin and Unal | 15 | 63 | 13 | 303 | 25 | 71 | 22 | 2.4 |
| Duval et al. | 16 | 115 | 12 | 251 | 20 | 243 | 27 | 8.1 |
| Eslami et al. | **19** | 183 | **9** | 183 | 24 | 570 | 23 | 19 |
| Flórez Valencia et al. | 5 | 54 | 23 | 312 | 7 | 140 | 40 | 4.4 |
| Lor and Chen | 14 | 87 | 14 | 279 | 15 | 484 | 32 | 16.1 |
| Melki et al. | 13 | 94 | 15 | 272 | 20 | 196 | 27 | 6.5 |
| Mohr et al. | 16 | 95 | 12 | 271 | 24 | 129 | 23 | 4.2 |
| Öksüz et al. | 6 | 21 | 22 | 345 | 8 | 23 | 39 | 0.8 |
| Shahzad et al. | 1 | **7** | 27 | **359** | **26** | 71 | **21** | 2.4 |
| Wang et al. | 7 | **7** | 21 | **359** | 5 | **10** | 42 | 0.3 |

Table 8: Performance of the 11 evaluated methods, 3 observers and their consensus for the *per-patient* detection of coronary artery stenoses (≥ 50% diameter reduction) on the 30 testing datasets. Prevalence of disease: 60%. The values in bold correspond to the best performance for each measure (sensitivity, specificity, positive predictive value, negative predictive value). Methods are listed by alphabetic order. Results in percentage. **NA* in case (TN + FN) is null.

| **Method** | **Cat.** | **QCA** | | | | **CTA** | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | **Sens.** | **Spec.** | **PPV** | **NPV** | **Sens.** | **Spec.** | **PPV** | **NPV** |
| *CTA consensus* | *Manual* | *94* | *67* | *81* | *89* | *100* | *100* | *100* | *100* |
| *Observer 2* | *Manual* | *100* | *58* | *78* | *100* | *95* | *78* | *91* | *88* |
| *Observer 1* | *Manual* | *94* | *50* | *74* | *86* | *95* | *67* | *87* | *86* |
| *Observer 3* | *Manual* | *94* | *75* | *85* | *90* | *86* | *75* | *90* | *67* |
| Broersen et al. | Auto. | 72 | 42 | 65 | 50 | 71 | 44 | 75 | 40 |
| Cetin and Unal | Min. user | 94 | 50 | 74 | 86 | 90 | 33 | 75 | 60 |
| Duval et al. | Auto. | 94 | 33 | 68 | 80 | 86 | 25 | 76 | 40 |
| Eslami et al. | Min. user | **100** | 0 | 57 | NA | **100** | 0 | 70 | NA |
| Flórez Valencia et al. | Min. user | **100** | 8 | 62 | **100** | **100** | 0 | 63 | NA |
| Lor and Chen | Min. user | **100** | 7 | 55 | **100** | **100** | 0 | 70 | NA |
| Melki et al. | Auto. | **100** | 8 | 62 | **100** | 95 | 0 | 69 | 0 |
| Mohr et al. | Auto. | **100** | 0 | 63 | NA | **100** | 0 | 70 | NA |
| Öksüz et al. | Min. user | 74 | 73 | 82 | 62 | 76 | 67 | **84** | 55 |
| Shahzad et al. | Min. user | 28 | **92** | **83** | 46 | **100** | 44 | 81 | **100** |
| Wang et al. | Auto. | 39 | 83 | 78 | 48 | 33 | **78** | 78 | 33 |

Table 9: Performance of the 11 evaluated methods, 3 observers and their consensus for the detection of coronary artery stenoses (≥ 50% diameter reduction) on the 30 testing datasets. The values in bold correspond to the best performance for each measure. QCA analysis is *segment-based*; CTA analysis is *lesion-based*.

| **Method** | **Cat.** | **QCA** | | | | **CTA** | | | | **Avg. rank** |
|---|---|---|---|---|---|---|---|---|---|---|
| | | **Sensitivity** | | **PPV** | | **Sensitivity** | | **PPV** | | |
| | | % | Rank | % | Rank | % | Rank | % | Rank | |
| *CTA consensus* | *Manual* | *82* | *1.0* | *52* | *1.0* | *100* | *1.0* | *100* | *1.0* | *1.2* |
| *Observer 2* | *Manual* | *75* | *3.0* | *51* | *2.0* | *70* | *3.0* | *81* | *2.0* | *2.5* |
| *Observer 1* | *Manual* | *86* | *1.0* | *40* | *5.0* | *83* | *2.0* | *61* | *3.0* | *2.8* |
| *Observer 3* | *Manual* | *64* | *5.0* | *43* | *4.0* | *66* | *4.0* | *60* | *4.0* | *4.2* |
| Cetin and Unal | Min. user | 54 | 8.0 | 19 | 7.0 | 53 | 6.0 | 26 | 8.0 | 7.2 |
| Mohr et al. | Auto. | 57 | 6.0 | 14 | 9.0 | 51 | 7.0 | 16 | 10.0 | 8.0 |
| Wang et al. | Auto. | 25 | 11.0 | **50** | 3.0 | 11 | 15.0 | **33** | 5.0 | 8.5 |
| Broersen et al. | Auto. | 25 | 11.0 | 18.9 | 8.0 | 27.7 | 12.0 | 31 | 6.0 | 9.2 |
| Shahzad et al. | Min. user | 4 | 15.0 | 13 | 11.0 | **55** | 5.0 | 27 | 7.0 | 9.5 |
| Eslami et al. | Min. user | **68** | 4.0 | 9 | 14.0 | 51 | 7.0 | 4 | 14.0 | 9.8 |
| Duval et al. | Auto. | 57 | 6.0 | 12 | 12.0 | 43 | 9.0 | 8 | 12.0 | 9.8 |
| Öksüz et al. | Min. user | 21 | 13.0 | 22 | 6.0 | 17 | 13.0 | 26 | 9.0 | 10.2 |
| Melki et al. | Auto. | 46 | 10.0 | 12 | 13.0 | 43 | 9.0 | 9 | 11.0 | 10.8 |
| Lor and Chen | Min. user | 50 | 9.0 | 14 | 10.0 | 32 | 11.0 | 3 | 15.0 | 11.2 |
| Flórez Valencia et al. | Min. user | 18 | 14.0 | 9 | 15.0 | 15 | 14.0 | 5 | 13.0 | 14.0 |

Table 10: Performance of the 8 evaluated methods, 3 observers and their consensus for the quantification of coronary artery stenoses on the 30 testing datasets. The quantification measures are computed using the union of the submitted lesions and the reference ones, thus including not only the TP, but also the FP and FN, to assess the whole system accuracy. The values in bold correspond to the best performance for each measure.

| **Method** | **Cat.** | **QCA** | | | | **CTA** | | **Avg. rank** |
|---|---|---|---|---|---|---|---|---|
| | | **Avg. Abs. Diff.** | | **R.M.S. Diff.** | | **Weighted Kappa** | | |
| | | % | Rank | % | Rank | $\kappa$ | Rank | |
| *CTA consensus* | *Manual* | *28.8* | *3.0* | *34.4* | *3.0* | *1.00* | *1.0* | *2.0* |
| Shahzad et al. | Min. user | **21.1** | 1.0 | **29.1** | 1.0 | **0.28** | 5.0 | 3.0 |
| *Observer 1* | *Manual* | *30.1* | *4.0* | *35.2* | *4.0* | *0.74* | *3.0* | *3.5* |
| *Observer 2* | *Manual* | *31.1* | *6.0* | *36.5* | *5.0* | *0.77* | *2.0* | *3.8* |
| *Observer 3* | *Manual* | *30.6* | *5.0* | *36.9* | *6.0* | *0.73* | *4.0* | *4.8* |
| Wang et al. | Auto. | 28.8 | 2.0 | 33.7 | 2.0 | 0.18 | 8.0 | 5.0 |
| Broersen et al. | Auto. | 32.5 | 7.0 | 39.3 | 7.0 | 0.27 | 6.0 | 6.5 |
| Öksüz et al. | Min. user | 47.0 | 9.0 | 53.1 | 9.0 | 0.21 | 7.0 | 8.0 |
| Lor and Chen | Min. user | 38.6 | 8.0 | 42.7 | 8.0 | -0.03 | 12.0 | 10.0 |
| Mohr et al. | Auto. | 49.6 | 10.0 | 56.0 | 12.0 | 0.15 | 9.0 | 10.0 |
| Flórez Valencia et al. | Min. user | 51.6 | 12.0 | 55.6 | 11.0 | 0.01 | 10.0 | 10.8 |
| Eslami et al. | Min. user | 50.9 | 11.0 | 55.0 | 10.0 | -0.02 | 11.0 | 10.8 |

Table 11: Performance of the 5 evaluated methods for coronary artery lumen segmentation on the 30 testing datasets. The values in bold correspond to the best performance for each measure.

| **Method** | **Cat.** | **DICE** | | | | **MSD** | | | | **MaxD** | | | | **Avg. rank** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **Diseased** | | **Healthy** | | **Diseased** | | **Healthy** | | **Diseased** | | **Healthy** | | |
| | | % | Rank | % | Rank | mm | Rank | mm | Rank | mm | Rank | mm | Rank | |
| *Observer 3* | *Manual* | *79* | *1.6* | *81* | *1.3* | *0.23* | *2.0* | *0.21* | *1.5* | *3.00* | *5.1* | *3.45* | *4.9* | *2.7* |
| Mohr et al. | Auto. | **70** | 3.6 | **73** | 3.4 | **0.40** | 4.2 | **0.39** | 3.8 | **2.68** | 2.9 | **2.75** | 2.2 | 3.3 |
| *Observer 1* | *Manual* | *76* | *2.3* | *77* | *3.2* | *0.24* | *2.6* | *0.24* | *2.8* | *2.87* | *4.3* | *3.47* | *4.8* | *3.4* |
| *Observer 2* | *Manual* | *65* | *5.0* | *72* | *4.9* | *0.34* | *4.7* | *0.27* | *3.7* | *2.82* | *4.5* | *3.26* | *4.3* | *4.5* |
| Shahzad et al. | Min. user | 58 | 6.3 | 66 | 5.8 | 0.49 | 6.5 | 0.43 | 5.3 | 2.81 | 5.0 | 3.05 | 3.0 | 5.2 |
| Wang et al. | Auto. | 69 | 4.5 | 69 | 4.6 | 0.45 | 5.4 | 0.5 | 5.9 | 3.94 | 5.7 | 6.48 | 5.9 | 5.4 |
| Broersen et al. | Auto. | 67 | 4.5 | 69 | 4.9 | 0.50 | 5.8 | 0.70 | 5.9 | 3.89 | 5.4 | 5.86 | 5.7 | 5.4 |
| Flórez Valencia et al. | Min. user | 42 | 7.8 | 38 | 7.7 | 0.83 | 7.2 | 1.13 | 7.7 | 3.81 | 4.4 | 6.96 | 5.6 | 6.8 |

either be caused by a high number of FP or FN, or by a high number of lesions reported with more than one grade difference as compared to the CTA reference.

The quantification results show that current stenosis quantification algorithms are not sufficiently reliable to be used standalone in clinical practice, but could be used as a second-reader.

### *5.3. Lumen segmentation*

Table 11 shows the average results and ranking of 5 submissions (3 fully automatic, 2 semi-automatic) and 3 observers for coronary artery lumen segmentation. The method proposed by Mohr et al. outperforms all the other methods, as well as two of the observers.

Overall, though the Dice value obtained on healthy vessel segments is higher than the one obtained on diseased ones, the mean square distance and maximum distance obtained on healthy segments is higher than the ones obtained on diseased ones, which may be caused by the smaller scale of the diseased vessel.

Figure 10 provides a visual impression for segment LAD7 of dataset#08 of the reference standard of observer#1 (top) and evaluated algorithms of respectively Broersen et al., Flórez Valencia et al., Mohr et al., Shahzad et al., and Wang et al.. Dataset #08 presents a severe mixed plaque in segment LAD7. While the algorithms of Mohr et al., Broersen et al. and Wang et al. successfully segment the diseased vessel segment as compared to the reference from the observer #1, the last two algorithms tend to under-segment the soft plaque. In this case, both the algorithm of Flórez Valencia et al. and Shahzad et al. fail to segment the mixed plaque: the first include the calcified part of the lesion within the segmentation, while the second is attracted towards the calcium spot. Note for this particular view of the vessel, the method of Shahzad et al. and Broersen et al. fail to display segmentation at some vessel position; in fact, their segmentation lies in another plane, and thus, no intersection was available.

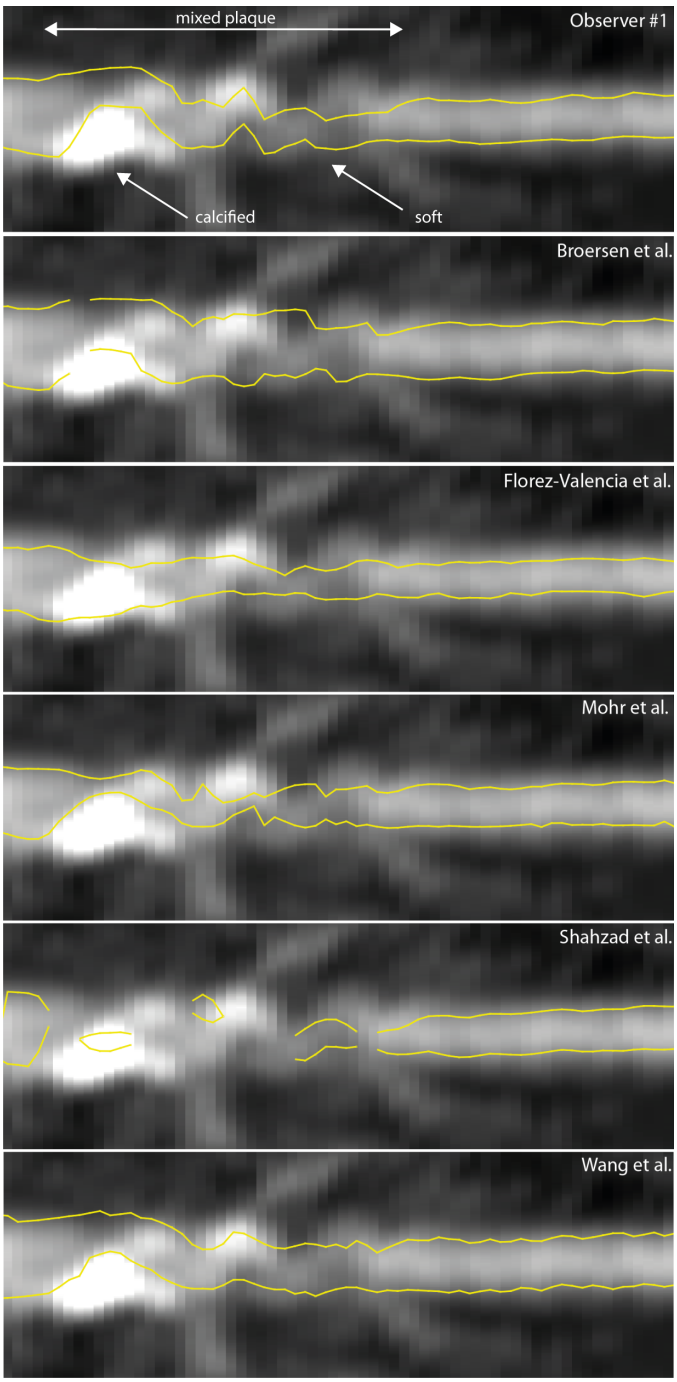



Figure 10: Lumen segmentation example for training dataset #08. Visual impression of the reference standard of observer#1 and evaluated algorithms of Broersen et al., Flórez Valencia et al., Mohr et al., Shahzad et al., and Wang et al.. Dataset #08 presents a severe mixed plaque in segment LAD7. Note that for this particular view of the vessel, the method of Shahzad et al. and Broersen et al. fail to display segmentation at some vessel position; in fact, their segmentation lies in another plane, and thus, no intersection was available.

## 6. Discussion

We presented a standardized evaluation framework allowing the effective comparison of coronary artery stenosis detection and quantification methods, and coronary lumen segmentation algorithms, on CTA images. The framework has been used to compare 11 algorithms as part of the "3D Cardiovascular Imaging: a MICCAI segmentation challenge" workshop at MICCAI'12, and remains publicly available via the website *http://coronary.bigr.nl/stenoses/*.

### *6.1. Evaluation framework*

The quality of an evaluation framework critically relies on the datasets that are made available for training and testing, and the quality of the reference standard. In our framework, currently, 48 cardiac CTA datasets with corresponding reference standard are available. Datasets were acquired at three different Dutch medical centers, with CT scanners from three different vendors (Siemens Healthcare, Philips Healthcare and Toshiba

Medical Systems), ensuring that algorithms would not be biased toward a specific scanner or acquisition protocol. The datasets were carefully selected in order to be representative of the types of pathologies which occur in clinical practice. Unfortunately, we were not able to include datasets from one of the remaining main CT scanner vendors (GE Healthcare) in the current framework. Also, it may be interesting to include CTA images acquired with 1) various acquisition protocols (retro/prospective ECG gating, flash mode, low-dose), 2) different reconstruction modes (different kernels), and 3) with different scanner models from a single vendor. Thus, the variety of CTA datasets provided in our framework could still be improved.

In our framework we utilized two reference standards: the performance of (semi-)automatic algorithms designed to detect and quantify stenoses was evaluated using both CTA consensus reading and QCA analysis of the angiograms.

Creating a reference standard with multiple observers is a tedious and complex task. To build the CTA reference standard, multiple observers annotated the datasets, whose annotations were then combined. Even though the same guidelines were given to all the observers, their annotations were not always consistent, thus making a merging step necessary. For instance, defining the start and end of a lesion can be challenging, especially in case of diffuse disease. One observer may indicate that a whole segment is diseased (leading to a single stenosis which extends over the whole segment), while another may indicate multiple smaller stenoses with varying degrees. Such disagreements between observers were solved during the consensus reading.

Another challenge is to build references from CTA and QCA which are consistent. As two different modalities are used, one providing 3D images and one providing 2D images, the observers' interpretation may considerably differ. Though the same coronary tree nomenclature was provided to CTA and QCA observers, mismatches between segments may occur: a stenosis detected at the end of the proximal LAD segment (LAD6) on the CTA image may be visualized as being in the proximal part of the mid LAD segment (LAD7) on QCA. To avoid such mismatch, the QCA reference has been corrected to match the CTA reference. Second, there can be detection/quantification mismatch: a significant stenosis may be reported in a certain segment on CTA, while no stenosis is reported on QCA, and vice-versa. Segments presenting mild stenoses on the CTA images usually do not present any obstruction on the CCA. Also, it may occur that, due to blooming artifacts caused by calcified plaque, motion artifacts or reduced image quality, a stenosis is overestimated on CTA, and is thus not present on the CCA image. Reversely, a stenosis detected on the CCA and not visible in CTA may be caused by an erroneous computation of the QCA, for instance by using wrong landmarks to estimate the "normal" vessel diameter immediately adjacent to the stenosis.

A limitation of the current evaluation framework is the point-based definition of stenosis location. Participants should provide a *single* point per stenosis, which is generally the central point. This may result in mismatches between the stenoses detected by the methods and the reference standard. If the method returns a series of smaller stenoses while the reference indicates a larger one, there is no mismatch: the large reference stenosis is detected, and will be assigned the average of the grades of the short stenoses. However, in the opposite case there will be mismatch: in the best case, only one of the reference stenoses will be correctly matched (i.e. if the provided point lies within the range of one of the small reference stenoses). Potentially, a better matching procedure could be implemented, which would take the *start* and *end* points of the stenoses as input. However, we believe that this situation occurs relatively infrequently. In addition, if a series of small stenoses has been reached by concensus, it could be argued that a method should detect this in a similar manner.

A second limitation is the use of hard classification into categories by the observers. For example, for a specific plaque, a method may yield a stenosis degree of 49%, while an observer grades the lesion as being moderate and the QCA reveals a 52% stenosis. Although the methods' stenosis degree estimation is close to the observers' one, the hard detection threshold at 50% would penalize the method, classifying its result as a FN. A potential solution would be to add a *borderline* category, so that the algorithm would not be punished neither for reporting nor for missing *borderline* stenoses. Another solution would be to use ROC curves for evaluating the algorithm performance. However, we believe that this limitation has had little impact in the overall evaluation, as in the current 48 datasets consisting of 637 coronary artery segments, only 15 of them had a QCA between 45% and 55%.

Last, as indicated in Section 5, the three observers performance for coronary stenoses detection and quantification as compared to CTA reference standard may be biased at their advantage, as the CTA reference standard was derived from a consensus reading of the same three observers. To allow a fair comparison of the observers performance with both the CTA reference standard and other evaluated methods, coronary artery detection and quantification in CTA should be obtained from different observers than the ones involved in the consensus reading.

### *6.2. Evaluated algorithms*

The aim of our standardized evaluation framework is to provide an objective methodology to compare the performance of different algorithms for certain clinical tasks. It is hence important, that the framework is adopted by all state-of-the-art algorithms. In the MICCAI challenge, 11 algorithms have been evaluated using the proposed framework, showing the potential of the framework to achieve this. However, not all recently published methods have yet been evaluated with our framework.

The authors of previously published stenosis detection algorithms, i.e. the ones presented in Table 2, were all invited by e-mail to participate in our MICCAI challenge, but none of them did and we did not further investigate why these groups did not participate.

Since the evaluation framework remains accessible, we hope and expect that an increasing number of algorithms will be evaluated. From the previous challenges we organized (Schaap

et al., 2009a; Hameeteman et al., 2011), we know that this indeed happens. Also, for newly published methods, reviewers of journals typically require a method to be evaluated using such standardized frameworks.

### *6.3. Evaluation results*

Nine of the eleven evaluated algorithms are developed following the work-flow of Figure 3, consisting of 1) the computation of an accurate lumen segmentation, either directly from the input CTA image or using previously extracted centerlines, and 2) the subsequent detection (and quantification) of coronary artery stenoses by estimate of the healthy vessel (Broersen et al., 2012; Eslami et al., 2012; Flórez Valencia et al., 2012; Mohr et al., 2012; Öksüz et al., 2012; Shahzad et al., 2012; Wang et al., 2012), or by analysis of intensity and geometry features (Lor and Chen, 2012; Melki et al., 2012). Though the nine methods actually segment the lumen, only five of them participated in the segmentation sub-challenge (exceptions are Eslami et al. (2012), Lor and Chen (2012), Melki et al. (2012) and Öksüz et al. (2012)). The lumen segmentation results show that the moderate detection and quantification performances of the algorithm proposed by Flórez Valencia et al. (2012) stem directly from the poor lumen segmentation results. To detect and quantify lesions, six of the algorithms estimated a "healthy" lumen radius using various regression approaches on the segmented lumen radius profile (linear for the approaches of Broersen et al. (2012), Flórez Valencia et al. (2012), Mohr et al. (2012) and Öksüz et al. (2012), second-order for the approach of Eslami et al. (2012), robust for the approach of Shahzad et al. (2012)). In the algorithm proposed by Wang et al. (2012) only, the outer vessel wall was segmented from the CTA image. Given similarly accurate lumen segmentation, the algorithm proposed by Shahzad et al. (2012) outperforms the approaches proposed by Broersen et al. (2012) and Wang et al. (2012) at the quantification stage. The results thus suggest that robust regression seems to be a good approach to quantify lesions from accurate lumen segmentation. Last, the algorithm proposed by Mohr et al. (2012) outperforms the three others at the segmentation stage, which suggests that tissue classification and calcium segmentation performed prior to lumen segmentation is a very promising approach.

Only one of the evaluated algorithms is not involving accurate lumen segmentation, but is using features extracted from the CTA image to detect plaques (Duval et al., 2012). Though reasonable sensitivity is achieved, the methods' performance is penalized by the important amount of reported FPs detections, and is therefore ranked just after the algorithms cited in previous paragraph (which make use of accurate lumen segmentation and regression) for detection.

The algorithm proposed by Cetin and Unal (2012), which makes use of both accurate lumen segmentation and feature extraction to detect lesions, seems very promising as it ranks first for detection.

Last, though the vessel lumen can be automatically segmented with a precision similar to the expert's one, detection and quantification of coronary artery stenosis is still not a solved problem; performance of the quantification is in general much worse than the observers.

The evaluation of the 11 different algorithms with the standardized evaluation framework provides useful directions for further investigations. First, it may be interesting to investigate the robustness of the segmentation methods with regard to the initial centerline used. Also, it would be interesting to combine the best segmentation method with the best detection and quantification method, and to combine the results of several algorithms and investigate whether a combination of algorithms outperforms the best single algorithm (Niemeijer et al., 2011).

Last, a clear limitation of our challenge remains that the evaluated algorithms are not available. In the future, the concept of a challenge would benefit from a framework where the evaluated algorithms become publicly available, such that it becomes possible to run the submitted algorithms on other datasets, without having to re-implement the complete pipeline, which is often tedious, if not possible at all, given that in literature often all information and/or data required to reproduce an algorithm is not available.

## 7. Conclusion

A publicly available evaluation framework to compare coronary artery stenosis detection and quantification methods, as well as lumen segmentation algorithms was presented in this article. The results showed that current stenosis detection/quantification algorithms are not sufficiently reliable to be used stand-alone in clinical practice, but that some could be used for triage or as a second-reader, and that automatic lumen segmentation is possible with a precision similar to the expert's one. The evaluation framework remains open for new submissions at *http://coronary.bigr.nl/stenoses/*.

## 8. Acknowledgments

Hortense Kirişli is supported by a grant from the Dutch Ministry of Economic Affairs (AgentschapNL) under the title "Het Hart in Drie Dimensies" (translated to "Heart In 3D", project PID06003). Coert Metz and Theo van Walsum are supported by a grant from the Information Technology for European Advancement (ITEA), under the title "Patient Friendly Medical Intervention" (project 09039, Mediate). Wiro Niessen is supported by the Stichting voor Technische Wetenschappen (STW) of the Netherlands Organization of Scientific Research (NWO).

## Appendix A. Reference and submission format

### *Appendix A.1. Reference format*

The CTA reference standard $\Lambda_{CTA}$ consists of a set of points in the coronary vessel tree. For every dataset, the following information is provided, per centerline point: 1) the $(x_{CTA}, y_{CTA}, z_{CTA})$ world coordinates, 2) the AHA-segment number $snr_{CTA}$ (between 1 and 17), 3) the patient's lesion number $lnr_{CTA}$ ($lnr_{CTA}$ =0 if the point does not belong to a stenosis, $lnr_{CTA}$ >0 if the point belongs to a stenosis, $lnr_{CTA} \leq N$, N

being the number of lesions present in the CTA reference standard), 4) the stenosis type $t$, and 5) the CTA diameter percentage stenosis $g_{CTA}$ (Table 1).

The CCA reference standard $\Lambda_{QCA}$ consists of couples $(snr_{QCA}, g_{QCA})$ for each of the 17-AHA segment $snr_{QCA}$ having a QCA diameter stenosis $g_{QCA}$ (between 0 and 100).

*Appendix A.2. Submission format*

The submitted results $\Lambda$ consist of the $(x, y, z)$ world coordinates position of each stenosis detected in the CTA image, and optionally the estimated CTA and QCA diameter stenosis $g_{CTA_{sub}}$ and $g_{QCA_{sub}}$, between 20 (mild) and 100 (complete occlusion).

## Appendix B. Matching procedure

As the evaluation is performed *lesion-based* with respect to the CTA reference standard, and *segment-based* with respect to the CCA reference standard, the reported stenoses need to be matched to one of the 17-AHA segment *snr* and to a reference stenosis *lnr*. The matching procedure is three-fold and is presented in Figure B.11. For all lesions $l^j \in \Lambda$, $j \in [1, S]$, $S$ being the number of detected stenoses, we determine 1) the 5-nearest neighbors of $l^j$ in $\Lambda_{CTA}$ ($k = 5$ was empirically determined, to cope with uncertainties at bifurcations for instance), 2) the segment number $snr^j$, and 3) the stenosis number $lnr^j$. Figure 9 presents a stenosis detection matching example for training dataset10 and results of observer#1.

## References


Achenbach, S., 2008. Can CT detect the vulnerable coronary plaque? International Journal of Cardiovascular Imaging 24, 311–312.

Achenbach, S., Schuhbaeck, A., Marwan, M., Bathina, R., Ovrehus, K., Anders, K., Hoffmann, U., Abbara, S., Aulbach, P., Ropers, D., Pflederer, T., Becker, C., Berman, D., Hausleiter, J., 2012. Multicenter evaluation of dual source CT coronary angiography in patients with intermediaite likelihood of coronary artery stenoses (MEDIC): Accuracy for the detection of individuals with significant coronary artery stenoses. Journal of American College of Cardiology 59 (1337), 61338–2.

Arnoldi, E., Gebregziabher, M., Schoepf, U. J., Goldenberg, R., Ramos-Duran, L., Zwerner, P. L., Nikolaou, K., Reiser, M. F., Costello, P., Thilo, C., May 2010. Automated computer-aided stenosis detection at coronary CT angiography: initial experience. European Radiology 20 (5), 1160–1167.

Austen, W. G., Edwards, J. E., Frye, R. L., Gensini, G. G., Gott, V. L., Griffith, L. S., McGoon, D. C., Murphy, M. L., Roe, B. B., Apr 1975. A reporting system on patients evaluated for coronary artery disease. report of the ad hoc committee for grading of coronary artery disease, council on cardiovascular surgery, american heart association. Circulation 51 (4 Suppl), 5–40.

Baltaxe Milwer, M., Flórez-Valencia, L., Hernández-Hoyos, M., Magnin, I., Orkisz, M., aug. 2007. Fast-marching contours for the segmentation of vessel lumen in CTA cross-sections. In: Proc. of the IEEE Engineering in Medicine and Biology Society. pp. 791 –794.

Boogers, M., Broersen, A., van Velzen, J., de Graaf, F., El-Naggar, H., Kitslaar, P., Dijkstra, J., Delgado, V., Boersma, E., de Roos, A., Schuijf, J., Schalij, M., Reiber, J., Bax, J., Jukema, J., 2012. Automated quantification of coronary plaque with computed tomography: comparison with intravascular ultrasound using a dedicated registration algorithm for fusion-based quantification. European Heart Journal 33 (8), 1007–1016.

Boogers, M. J., Schuijf, J. D., Kitslaar, P. H., van Werkhoven, J. M., de Graaf, F. R., Boersma, E., van Velzen, J. E., Dijkstra, J., Adame, I. M., Kroft, L. J., de Roos, A., Schreur, J. H. M., Heijenbrok, M. W., Jukema, J. W., Reiber, J. H. C., Bax, J. J., Jul 2010. Automated quantification of stenosis severity on 64-slice CT: a comparison with quantitative coronary angiography. JACC Cardiovasc Imaging 3 (7), 699–709.
URL http://dx.doi.org/10.1016/j.jcmg.2010.01.010

Broersen, A., Kitslaar, P., Frenay, M., Dijkstra, J., 2012. FrenchCoast: Fast, Robust Extraction for the Nice Challenge on COronary Artery Segmentation of the Tree. In: Proc. of MICCAI Workshop "3D Cardiovascular Imaging: a MICCAI segmentation Challenge".

Budoff, M. J., Achenbach, S., Blumenthal, R. S., Carr, J. J., Goldin, J. G., Greenland, P., Guerci, A. D., Lima, J. A. C., Rader, D. J., Rubin, G. D., Shaw, L. J., Wiegers, S. E., on Cardiovascular Imaging, A. H. A. C., Intervention, on Cardiovascular Radiology, A. H. A. C., Intervention, American Heart Association Committee on Cardiac Imaging, C. o. C. C., Oct 2006. Assessment of coronary artery disease by cardiac computed tomography: a scientific statement from the american heart association committee on cardiovascular imaging and intervention, council on cardiovascular radiology and intervention, and committee on cardiac imaging, council on clinical cardiology. Circulation 114 (16), 1761–1791.

Cetin, S., Demir, A., Yezzi, A., Degertekin, M., Unal, G., 2013. Vessel tractography using an intensity based tensor model with branch detection. IEEE Transactions on Medical Imaging 32 (2), 348–363.

Cetin, S., Unal, G., 2012. Automatic detection of coronary artery stenosis in cta based on vessel intensity and geometric features. In: Proc. of MICCAI Workshop "3D Cardiovascular Imaging: a MICCAI segmentation Challenge".

Cohen, J., 1968. Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. Psychological Bulletin 70 (4), 213–220.

den Dekker, M., de Smet, K., de Bock, G., Tio, R., Oudkerk, M., Vliegenthart, R., Dec 2012. Diagnostic performance of coronary CT angiography for stenosis detection according to calcium score: systematic review and meta-analysis. European Radiology 22 (12), 2688–2698.

Duval, M., Ouzeau, E., Precioso, F., Matuszewski, B., 2012. Coronary artery stenoses detection with random forest. In: Proc. of MICCAI Workshop "3D Cardiovascular Imaging: a MICCAI segmentation Challenge".

Eslami, A., Aboee, A., Hodaei, Z., Moghaddam, M. J., Carlier, S., Katouzian, A., Navab, N., 2012. Quantification of coronary arterial stenosis by inflating tubes in CTA images. In: Proc. of MICCAI Workshop "3D Cardiovascular Imaging: a MICCAI segmentation Challenge".

Falk, E., Shah, P. K., Fuster, V., Aug 1995. Coronary plaque disruption. Circulation 92 (3), 657–671.

Flórez Valencia, L., Orkisz, M., Corredor Jerez, R. A., Torres Gonzlez, J. S., Correa Agudelo, E. M., Mouton, C., Hernández Hoyos, M., 2012. coronary artery segmentation and stenosis quantification in ct images with use of a right generalized cylinder model,. In: Proc. of MICCAI Workshop "3D Cardiovascular Imaging: a MICCAI segmentation Challenge".

Frangi, A., Niessen, W., Vincken, K., Viergever, M., 1998. Multiscale vessel enhancement filtering. In: Proc. of MICCAI'98. Vol. .1496. pp. 130–137.

Goldenberg, R., Eilot, D., Begelman, G., Walach, E., Ben-Ishai, E., Peled, N., Apr 2012. Computer-aided simple triage (CAST) for coronary CT angiography (CCTA). International Journal of Computer Assisted Radiology and Surgery, 1–9.

Goldenberg, R., Peled, N., 2011. Computer-aided simple triage. International Journal of Computer Assisted Radiology and Surgery 6 (5), 705–711.

Gülsün, M. A., Tek, H., 2008. Robust vessel tree modeling. In: Proc. of the 11th international conference on Medical Image Computing and Computer-Assisted Intervention - Part I. MICCAI '08. Springer-Verlag, Berlin, Heidelberg, pp. 602–611.

Haase, J., Escaned, J., Montauban van Swijndregt, E., Ozaki, Y., Gronenschild, E., Slager, C., P.W., S., 1993. Experimental validation of geometric and densitometric coronary measurements on the new generation cardiovascular angiography analysis system (caasii). Catheterization and Cardiovascular Diagnosis 30, 104–114.

Halpern, E. J., Halpern, D. J. a., Mar 2011. Diagnosis of coronary stenosis with ct angiography comparison of automated computer diagnosis with expert readings. Academic Radiology 18 (3), 324–333.

Hameeteman, K., Zuluaga, M. A., Freiman, M., Joskowicz, L., Cuisenaire, O., Flórez Valencia, L., Gülsün, M. A., Krissian, K., Mille, J., Wong, W. C. K., Orkisz, M., Tek, H., Hernández Hoyos, M., Benmansour, F., Chung, A. C. S., Rozie, S., van Gils, M., van den Borne, L., Sosna, J., Berman, P., Cohen, N., Douek, P. C., Sánchez, I., Aissat, M., Schaap, M., Metz, C. T., Krestin, G. P., van der Lugt, A., Niessen, W. J., van Walsum, T., Aug

2011. Evaluation framework for carotid bifurcation lumen segmentation and stenosis grading. Medical Image Analysis 15 (4), 477–488.

Kelm, B. M., Mittal, S., Zheng, Y., Tsymbal, A., Bernhardt, D., Vega-Higuera, F., Zhou, S. K., Meer, P., Comaniciu, D., 2011. Detection, grading and classification of coronary stenoses in computed tomography angiography. Medical Image Computing and Computer-Assisted Interventions 14, 25–32.

Khan, M. F., Wesarg, S., Gurung, J., Dogan, S., Maataoui, A., Brehmer, B., Herzog, C., Ackermann, H., Assmus, B., Vogl, T. J., Aug 2006. Facilitating coronary artery evaluation in MDCT using a 3D automatic vessel segmentation tool. European Radiology 16 (8), 1789–1795.

Lesage, D., Angelini, E. D., Bloch, I., Funka-Lea, G., Dec 2009. A review of 3D vessel lumen segmentation techniques: models, features and extraction schemes. Medical Image Analysis 13 (6), 819–845.

Lor, K., Chen, C., 2012. Probabilistic model based evaluation of coronary artery stenosis on CTA. In: Proc. of MICCAI Workshop "3D Cardiovascular Imaging: a MICCAI segmentation Challenge".

Meijboom, W. B., Meijs, M. F. L., Schuijf, J. D., Cramer, M. J., Mollet, N. R., van Mieghem, C. A. G., Nieman, K., van Werkhoven, J. M., Pundziute, G., Weustink, A. C., de Vos, A. M., Pugliese, F., Rensing, B., Jukema, J. W., Bax, J. J., Prokop, M., Doevendans, P. A., Hunink, M. G. M., Krestin, G. P., de Feyter, P. J., Dec 2008. Diagnostic accuracy of 64-slice computed tomography coronary angiography: a prospective, multicenter, multivendor study. Journal of American College of Cardiology 52 (25), 2135–2144.

Melki, I., Talbot, H., Cousty, J., Pruvot, C., Knoplioch, J., Launay, L., Najman, L., 2012. Automatic coronary arteries stenosesdetection in 3D CTA. In: Proceedings of MICCAI Workshop "3D Cardiovascular Imaging: a MICCAI segmentation Challenge".

Metz, C., Schaap, M., Weustink, A., Mollet, N., van Walsum, T., Niessen, W., 2009. Coronary centerline extraction from ct coronary angiography images using a minimum cost path approach. Medical Physics 36 (12), 5568–5579.

Mittal, S., Zheng, Y., Georgescu, B., Vega-Higuera, F., Zhou, S., Meer, P., Comaniciu, D., 2010. Fast automatic detection of calcified coronary lesions in 3D cardiac CT images. In: Proc. of MICCAI Workshop "Machine Learning in Medical Imaging" (MLMI). Vol. 6357. pp. 1–9.

Mohr, B., Masood, S., Plakas, C., 2012. Accurate stenosis detection and quantification in coronary CTA. In: Proc. of MICCAI Workshop "3D Cardiovascular Imaging: a MICCAI segmentation Challenge".

Nieman, K., Galema, T. W., Neefjes, L. A., Weustink, A. C., Musters, P., Moelker, A. D., Mollet, N. R., de Visser, R., Boersma, E., de Feijter, P. J., Dec 2009. Comparison of the value of coronary calcium detection to computed tomographic angiography and exercise testing in patients with chest pain. American Journal Cardiology 104 (11), 1499–1504.

Niemeijer, M., Loog, M., Abràmoff, M. D., Viergever, M. A., Prokop, M., van Ginneken, B., 2011. On combining computer-aided detection systems. IEEE Transactions on Medical Imaging 30 (2), 215–223.

Öksüz, d., Ünay, D., Kadipaşaoğlu, K., 2012. A hybrid method for coronary artery stenosis detection and quantification. In: Proc. of MICCAI Workshop "3D Cardiovascular Imaging: a MICCAI segmentation Challenge".

Pryor, D. B., Shaw, L., McCants, C. B., Lee, K. L., Mark, D. B., Harrell, F. E., Muhlbaier, L. H., Califf, R. M., Jan 1993. Value of the history and physical in identifying patients at increased risk for coronary artery disease. Annals of Internal Medicine 118 (2), 81–90.

Raff, G. L., Abidov, A., Achenbach, S., Berman, D. S., Boxt, L. M., Budoff, M. J., Cheng, V., DeFrance, T., Hellinger, J. C., Karlsberg, R. P., 2009. SCCT guidelines for the interpretation and reporting of coronary computed tomographic angiography. Journal of Cardiovascular Computed Tomography 3 (2), 122–136.

Roger, V. L., Go, A. S., Lloyd-Jones, D. M., Benjamin, E. J., Berry, J. D., Borden, W. B., Bravata, D. M., Dai, S., Ford, E. S., Fox, C. S., Fullerton, H. J., Gillespie, C., Hailpern, S. M., Heit, J. A., Howard, V. J., Kissela, B. M., Kittner, S. J., Lackland, D. T., Lichtman, J. H., Lisabeth, L. D., Makuc, D. M., Marcus, G. M., Marelli, A., Matchar, D. B., Moy, C. S., Mozaffarian, D., Mussolino, M. E., Nichol, G., Paynter, N. P., Soliman, E. Z., Sorlie, P. D., Sotoodehnia, N., Turan, T. N., Virani, S. S., Wong, N. D., Woo, D., Turner, M. B., Committee, A. H. A. S., Subcommittee, S. S., Jan 2012. Heart disease and stroke statistics–2012 update: a report from the american heart association. Circulation 125 (1), e2–e220.

Saur, S. C., Alkadhi, H., Desbiolles, L., Székely, G., Cattin, P. C., 2008. Automatic detection of calcified coronary plaques in computed tomography data sets. Medical Image Computing and Computer-Assisted Interventions 11, 170–177.

Schaap, M., Metz, C. T., van Walsum, T., van der Giessen, A. G., Weustink, A. C., Mollet, N. R., Bauer, C., Bogunović, H., Castro, C., Deng, X., Dikici, E., O'Donnell, T., Frenay, M., Friman, O., Hernández Hoyos, M., Kitslaar, P. H., Krissian, K., Kühnel, C., Luengo-Oroz, M. A., Orkisz, M., Smedby, Ö., Styner, M., Szymczak, A., Tek, H., Wang, C., Warfield, S. K., Zambal, S., Zhang, Y., Krestin, G. P., Niessen, W. J., Oct 2009a. Standardized evaluation methodology and reference database for evaluating coronary artery centerline extraction algorithms. Medical Image Analysis 13 (5), 701–714.

Schaap, M., Neefjes, L., Metz, C., van der Giessen, A., Weustink, A., Mollet, N., Wentzel, J., van Walsum, T., Niessen, W., July 2009b. Coronary lumen segmentation using graph cuts and robust kernel regression. In: Jerry L. Prince, Dzung L. Pham, K. J. M. (Ed.), Information Processing in Medical Imaging. pp. 528–539.

Shahzad, R., van Walsum, T., Kirişli, H., Tang, H., Metz, C., Schaap, M., van Vliet, L., Niessen, W., 2012. Automatic detection, quantification and lumen segmentation of the coronary arteries using two-point centerline extraction scheme. In: Proc. of MICCAI Workshop "3D Cardiovascular Imaging: a MICCAI segmentation Challenge".

Shaw, L., Marwick, T., Zoghbi, W., Hundley, W., Kramer, C., Achenbach, S., Dilsizian, V., Kern, M., Chandrashekhar, Y., Narula, J., 2010. Why all the focus on cardiac imaging? Journal of American College of Cardiology 3 (7), 789–794.

Teßmann, M., Vega-Higuera, F., Fritz, D., Scheuering, M., Greiner, G., 2009. Multi-scale feature extraction for learning-based classification of coronary artery stenosis. In: Proc. of SPIE, Medical Imaging 2009: Computer-Aided Diagnosis.

Virmani, R., Burke, A., Farb, A., Kolodgie, F., 2006. Pathology of the vulnerable plaque. Journal of American College of Cardiology 47 (8), 13–18.

Wang, C., Moreno, R., Smedby, Ö., 2012. Vessel segmentation using implicit model-guided level sets. In: Proc. of MICCAI Workshop "3D Cardiovascular Imaging: a MICCAI segmentation Challenge".

Wesarg, S., Khan, M. F., Firle, E. A., Sep 2006. Localizing calcifications in cardiac CT data sets using a new vessel segmentation approach. Journal of Digital Imaging 19 (3), 249–257.

Weustink, A. C., de Feyter, P. J., Aug 2011. The role of multi-slice computed tomography in stable angina management: a current perspective. Netherlands Heart Journal 19 (7-8), 336–343.

Xu, Y., Liang, G., Hu, G., Yang, Y., Geng, J., Saha, P. K., Jan 2012. Quantification of coronary arterial stenoses in CTA using fuzzy distance transform. Computerized Medical Imaging and Graphics 36 (1), 11–24.

Yang, G., Broersen, A., Petr, R., Kitslaar, P., de Graaf, M., Bax, J. J., Reiber, J. H. C., Dijkstra, J., 2011. Automatic coronary artery tree labeling in coronary computed tomographic angiography datasets. Computing in Cardiology 38, 109–112.

Yang, G., Kitslaar, P., Frenay, M., Broersen, A., Boogers, M. J., Bax, J. J., Reiber, J. H. C., Dijkstra, J., Apr 2012. Automatic centerline extraction of coronary arteries in coronary computed tomographic angiography. International Journal of Cardiovascular Imaging 28 (4), 921–933.

Zambal, S., Hladuvka, J., Kanitsar, A., Bühler, K., 2008. Shape and appearance models for automatic coronary artery tracking. In: Proc. of MICCAI Workshop 3D Segmentation in the Clinic: A Grand Challenge II.

Zhou, C., Chan, H.-P., Chughtai, A., Patel, S., Hadjiiski, L. M., Sahiner, B., Wei, J., Kazerooni, E. A., 2010. Automated segmentation and tracking of coronary arteries in cardiac CT scans: comparison of performance with a clinically used commercial software. In: Proc. of SPIE, Medical Imaging 2010: Computer-Aided Diagnosis. Vol. 7624.

Zuluaga, M. A., Magnin, I. E., Hernández Hoyos, M., Delgado Leyton, E. J. F., Lozano, F., Orkisz, M., Mar 2011. Automatic detection of abnormal vascular cross-sections based on density level detection and support vector machines. International Journal of Computer Assisted Radiology and Surgery 6 (2), 163–174.

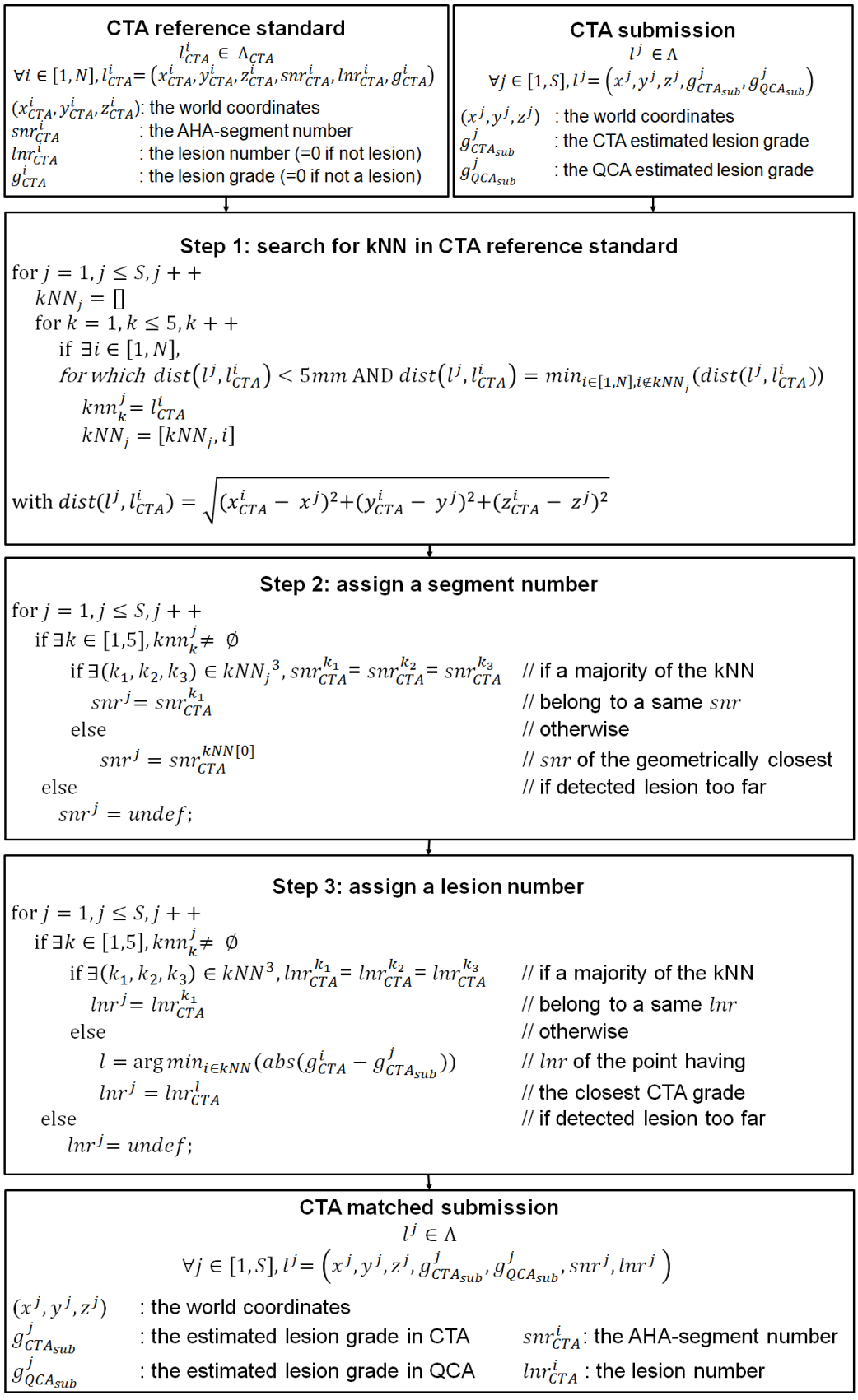



Figure B.11: References and submission - Stenoses and segment matching procedure