Deep convolutional neural networks for automatic coronary calcium scoring in a screening study with low-dose chest CT

Nikolas Lessmann[a], Ivana Išgum[a], Arnaud A.A. Setio[b], Bob D. de Vos[a], Francesco Ciompi[b], Pim A. de Jong[c], Matthijs Oudkerk[d], Willem P.Th.M. Mali[c], Max A. Viergever[a], and Bram van Ginneken[b]

[a]Image Sciences Institute, University Medical Center Utrecht, The Netherlands
[b]Diagnostic Image Analysis Group, Radboud University Medical Center Nijmegen, The Netherlands
[c]Department of Radiology, University Medical Center Utrecht, The Netherlands
[d]Center for Medical Imaging North East Netherlands, University Medical Center Groningen, The Netherlands

ABSTRACT

The amount of calcifications in the coronary arteries is a powerful and independent predictor of cardiovascular events and is used to identify subjects at high risk who might benefit from preventive treatment. Routine quantification of coronary calcium scores can complement screening programs using low-dose chest CT, such as lung cancer screening. We present a system for automatic coronary calcium scoring based on deep convolutional neural networks (CNNs).

The system uses three independently trained CNNs to estimate a bounding box around the heart. In this region of interest, connected components above 130 HU are considered candidates for coronary artery calcifications. To separate them from other high intensity lesions, classification of all extracted voxels is performed by feeding two-dimensional 50 mm × 50 mm patches from three orthogonal planes into three concurrent CNNs. The networks consist of three convolutional layers and one fully-connected layer with 256 neurons.

In the experiments, 1028 non-contrast-enhanced and non-ECG-triggered low-dose chest CT scans were used. The network was trained on 797 scans. In the remaining 231 test scans, the method detected on average 194.3 mm^3 of 199.8 mm^3 coronary calcifications per scan (sensitivity 97.2 %) with an average false-positive volume of 10.3 mm^3. Subjects were assigned to one of five standard cardiovascular risk categories based on the Agatston score. Accuracy of risk category assignment was 84.4 % with a linearly weighted κ of 0.89.

The proposed system can perform automatic coronary artery calcium scoring to identify subjects undergoing low-dose chest CT screening who are at risk of cardiovascular events with high accuracy.

Keywords: Automatic coronary calcium scoring, cardiovascular risk estimation, low-dose chest CT, lung cancer screening, convolutional neural networks, deep learning

1. INTRODUCTION

Cardiovascular disease (CVD) is the most common cause of death world-wide.[1] Smoking is an important risk factor for CVD and as such strongly related to fatal and nonfatal cardiovascular events. In the so far largest lung cancer screening trial in heavy smokers, cardiovascular illness was the leading cause of death, even before lung cancer.[2] Quantification of coronary artery calcifications (CAC) using dedicated cardiac CT is well-established as an independent predictor of cardiovascular events and mortality. It has recently been shown that CAC quantified in non-ECG-synchronized low-dose chest CT also enables identification of subjects at high cardiovascular risk.[3–5] In a lung cancer screening setting, coronary calcium quantification ("calcium scoring") therefore has the potential to identify subjects at high risk who might benefit from preventive treatment and for whom referral to a cardiologist may be advised.[6]

Eligibility criteria for lung cancer screening with low-dose chest CT are typically based on age and smoking history, two major risk factors for coronary artery disease. The prevalence and amount of coronary calcifications is therefore relatively high in the screened subjects compared to the general population. Moreover, a large

Send correspondence to n.lessmann@umcutrecht.nl

Medical Imaging 2016: Computer-Aided Diagnosis, edited by Georgia D. Tourassi, Samuel G. Armato III,
Proc. of SPIE Vol. 9785, 978511 · © 2016 SPIE · CCC code: 1605-7422/16/$18 · doi: 10.1117/12.2216978

number of scans have to be analyzed given that these criteria are met by a substantial part of the population.[7] For these reasons, manual calcium scoring would add a considerable burden to the screening process and, hence, automatic calcium detection and quantification would be desirable.

The coronary arteries are typically not well visible in CT without contrast enhancement. In addition, the lack of ECG-synchronization results in motion artifacts in the heart and the low radiation dose leads to increased image noise. Together, this makes automatic detection of coronary calcifications a challenging task. Previous approaches used hand-crafted features[8] or decision rules[9] to detect lesions in the coronary arteries. In recent years, deep convolutional neural networks[10] (CNNs) have shown outstanding performance in many image classification contests.[11,12] An interesting property of these networks is that they do not rely on expert designed features or decision rules but learn themselves which features of the image are relevant for classification purposes, which makes them potentially more robust and generic than traditional machine learning approaches. We therefore exploited the use of CNNs for detection of coronary calcifications in a screening study with low-dose chest CT.

2. DATA DESCRIPTION

This study included 1028 current or former heavy smokers between the age of 50 and 75 who participated in a screening study with low-dose chest CT. Participants were screened between 2004 and 2006 at the UMC Utrecht (710), UMC Groningen (159) and Kennemer Gasthuis Haarlem (159). CT scans were acquired without contrast enhancement and ECG triggering at inspiration with 16 or 64 detector-row CT scanners (Mx8000 IDT or Brilliance 16P, Philips Medical Systems; Sensation 16 or Sensation 64, Siemens Medical Solutions). Slice thickness was 0.75 mm or 1.0 mm with 0.7 mm increment, and in-plane resolution ranged from 0.52 mm to 0.89 mm. Peak voltage was, depending on the patient weight, 120 kVp or 140 kVp at a tube current of 30 mAs. The images were reconstructed with medium-soft kernels (Philips B, Siemens B30f). Because calcium scoring is commonly performed in images with a slice thickness of 3 mm, thick-slice images were reconstructed from the original thin-slice reconstructions by averaging four consecutive slices with two overlapping slices.[13]

The reference standard was defined by manual annotation of coronary calcifications in the thick-slice images by one experienced observer. Only voxels with attenuation values above the clinical standard threshold for calcium scoring of 130 HU were considered. The observer manually marked all calcified lesions in the coronary arteries. In order to reduce the influence of image noise, only lesions larger than 1.5 mm^3 were considered coronary calcifications. Smaller lesions were ignored as they consisted of only one or two voxels so that the observer could not be certain in distinguishing them from noise.

3. METHOD

The automatic system first estimates the position of the heart using three deep CNNs that predict the presence of the heart in axial, coronal and sagittal slices of the image volume.[14] The combination of these per-slice probabilities yields a 3D bounding box around the heart. Inside this region of interest, candidates for CAC are identified by thresholding at the standard threshold of 130 HU followed by 3D connected-component analysis. Components smaller than 1.5 mm^3 or larger than 5000 mm^3 are discarded from further analysis as they are likely noise or bony structures. Unlike previous approaches that analyzed lesions, all voxels contained in the extracted candidates are analyzed individually. This increases the robustness of the predictions as many perspectives of each lesion are considered compared to analyzing the lesions at only a single location, e.g., their center of mass. Centered at each analyzed voxel, three two-dimensional 50 mm $\times$ 50 mm patches (axial, sagittal, and coronal; 64×64 pixels) are extracted from the respective image and subsequently classified by a convolutional neural network. The posterior probability of each candidate lesion is obtained by averaging the predictions of all voxels within that lesion.

The network is a feed-forward neural network with three convolutional layers and one fully connected layer (Figure 1). Every convolutional layer is followed by 2×2 max-pooling to reduce the size of the feature maps and thus the number of network parameters. Furthermore, half of the neurons in the fully connected layer are randomly deactivated during training ("Dropout") to avoid over-fitting. Because coronary calcifications appear in a large variety of orientations and do not have consistent alignment with the patches, the three orthogonal patches per voxel are fed into networks with identical architecture and shared weights. The outputs of the final

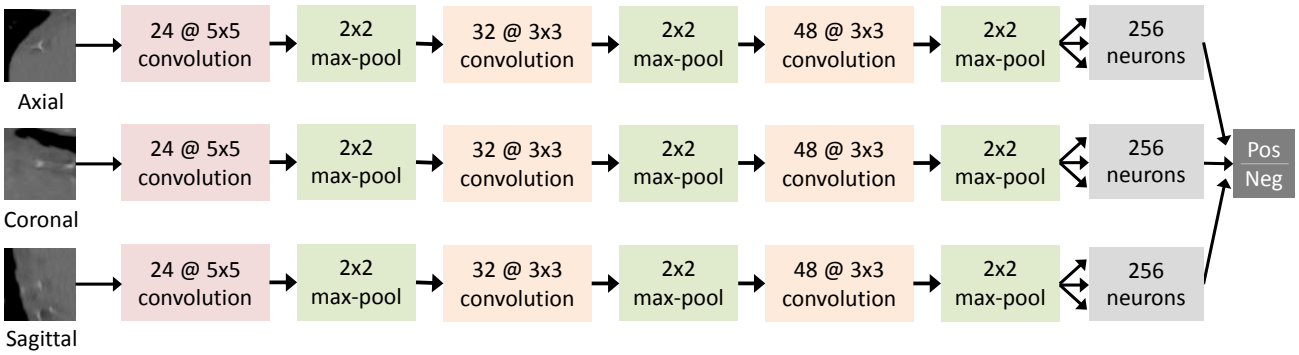

Figure 1. Architecture of the convolutional neural network. Weights are shared between the three concurrent networks. The convolutional steps are followed by 2×2 subsampling with max-pooling, and are finally connected to one fully-connected layer consisting of 256 neurons. These are fully-connected to two output nodes.

hidden layers of the three networks are jointly connected to two output nodes, which correspond to positive (coronary calcification) and negative (other high density candidate) classification.

The class label of each candidate lesion was determined by thresholding the posterior probability. Because the total volume of calcium in the coronary arteries is a better predictor of cardiovascular events than the number of calcified lesions,[5] this threshold was defined to minimize the average volume error per scan. For each subject, the detected coronary calcifications were quantified using volume and Agatston scores.[15] Thereafter, all subjects were assigned to one of five standard cardiovascular risk categories based on the Agatston score (I–V: 0, 1–10, 11–100, 101–400, >400). These results were then compared to the manual reference annotations.

4. RESULTS AND DISCUSSION

From the 1028 images in our data set, 797 were used to train the network. These images contained on average 680.5 mm^3 coronary calcium per scan and in total 5322 coronary calcifications with 347 717 voxels. An equal number of samples from the negative class were randomly selected from all 249 665 negative candidate lesions within the cardiac region of interest because preliminary experiments showed that our network learned better from a balanced training set. Ten percent of the positive and negative training samples were used to optimize the network architecture and for validation during training.

The remaining 231 test images contained 657 coronary calcifications and 16 840 negative candidates. The automatic method detected on average 194.3 mm^3 of 199.8 mm^3 coronary calcifications (sensitivity 97.2 %) per scan. The average false-positive volume was 10.3 mm^3. Since the classification threshold was optimized for minimal volume error, the lesion-wise sensitivity was slightly lower, namely 88.9 %, with on average 0.48 false-positives per scan. The correct cardiovascular risk category was assigned to 84.4 % of the subjects (Table 1). 33 subjects (14.3 %) were one category off, two subjects (0.87 %) were two categories off, and one subject (0.43 %) was three categories off. No subject was four categories off. The reliability of the risk category assignment was excellent[16] with a linearly weighted κ of 0.89.

The high detection sensitivity led to slight overestimation of cardiovascular risk due to false-positive detections. As the average false-positive volume was small, this mostly affected low-risk categories (I, II) where small absolute score changes are more likely to affect risk categorization. However, reliable classification of these subjects is hardly feasible in low-dose chest CT due to the compromised image quality and the often severe motion artifacts. In addition, preventive efforts are likely not indicated for these subjects. Less than 6% of the subjects with very-low and low cardiovascular risk were incorrectly assigned a higher risk (III or higher). On contrary, it is clinically important to correctly classify subjects at intermediate to high risk (III–V) as those could benefit from primary and secondary preventive efforts.[6] More than 99% of the subjects with elevated cardiovascular risk were identified, and the correct risk category was assigned in 95.5% of the cases.

The proposed method consists of two major stages. In the first stage, a bounding box around the heart is estimated, limiting subsequent analysis to a region of interest (Figure 2). Preliminary experiments with the

Table 1. Agreement between the automatically assigned CVD risk category and the manually determined reference category. Automatic method and reference agreed in 195 of 231 images (84.4 %). The network over-estimated risk in 26 (11.3 %) and under-estimated in 10 (4.3 %) of the subjects. 24 of these 36 misclassifications (66 %) occurred between very-low (I) and low risk (II).

Ref \ Auto	I 0	II 1–10	III 11–100	IV 101–400	V >400
I (0)	**71**	17	2	1	0
II (1–10)	7	**19**	4	0	0
III (11–100)	0	1	**43**	1	0
IV (101–400)	0	0	2	**31**	1
V (>400)	0	0	0	0	**31**

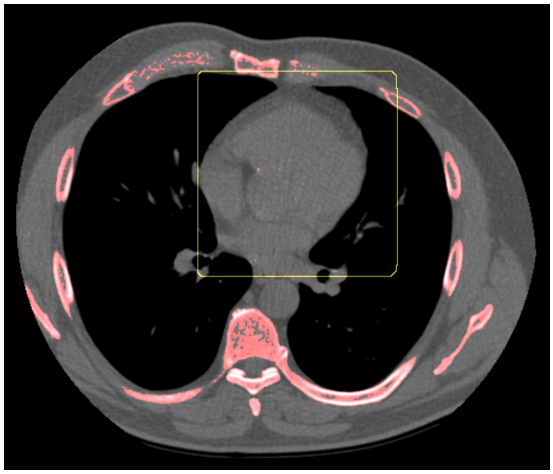

Figure 2. CT scan overlayed with potential CAC candidates, i.e., voxels above 130 HU (red), and the automatically detected region of interest (yellow). The majority of candidates is not within the region of interest.

whole images resulted in often large false-positives in the heart and in the ascending and descending aorta, which strongly influenced the accuracy of cardiovascular risk category assignment (Figures 3 and 4). When classification was performed inside the bounding box only, the performance improved substantially as false-positives outside the heart were discarded and the network focused on distinguishing positive and negative candidates inside the heart. Notably, proximal coronary calcifications were reliably distinguished from non-coronary calcifications located in the ascending aorta in direct neighborhood to the coronary arteries (Figure 5). Most false-positives represented image noise in or in proximity of the coronary arteries. Generally, the network learned to infer plausible locations of coronary calcifications from three small orthogonal patches alone, without any explicitly provided spatial information. In contrast, spatial features were crucial for correct classification in previous approaches.[8,17] However, such features rely on the definition of a common reference space, which typically requires complex non-elastic registration steps.

The presented method outperformed previously published techniques for automatic CAC detection in non-ECG-synchronized chest CT. Xie et al.[9] reported results for 41 scans using only four Agatston-based risk categories (0–10, 11–100, 101–400, >400), to which 58.5 % of the subjects were correctly assigned. Using the same metric, the here proposed method achieves 94.8 % risk categorization accuracy. In comparison to the method published by Išgum et al.,[8] which was evaluated on the same set of 231 scans as in this work, our convolutional neural network performed only slightly better in terms of accuracy of risk category agreement (84.4 % vs. 82.2 %). However, even though both approaches were optimized for minimal volume error, the CNN performed much better in terms of lesion-wise sensitivity (89 % vs. 59 %) and volume-wise sensitivity (97 % vs. 79 %). However, the CNN had a higher false-positive rate (10 mm^3 vs. 4 mm^3).

In other studies with low-dose chest CT scans for screening purposes, images are acquired with a wide variety of imaging protocols and exhibit a large range of pathologies. The transfer of automatic calcium scoring methods to other data sets is therefore not trivial. However, the here presented method exclusively relies on convolutional neural networks, which learn to extract and combine features based on the appearance of the provided training images. It will therefore likely adapt well to characteristics of different populations and imaging protocols. In future work, we are planning to evaluate the performance on images from other studies and for a larger number of subjects.

5. CONCLUSIONS

We presented an automatic approach for coronary artery calcium scoring in low-dose chest CT. This method employs deep convolutional neural networks for detection of coronary calcifications and achieved high detection

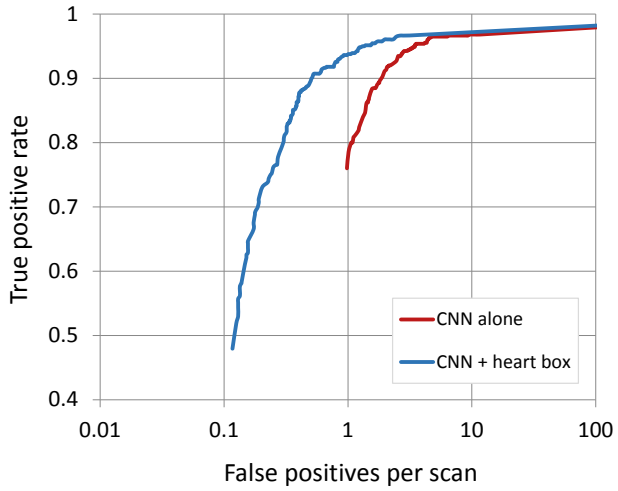

Figure 3. FROC curves for the CNN with and without removal of candidates outside the heart prior to training and testing. Detection performance is here reported in terms of detected lesions.

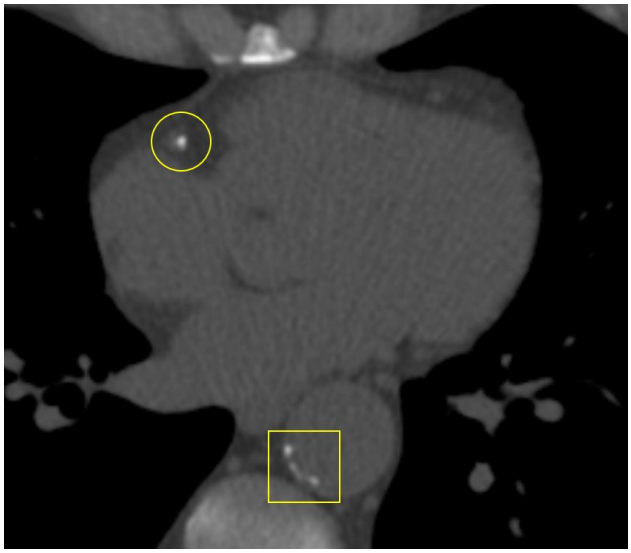

Figure 4. Example of classification result without using a cardiac bounding box. The CNN falsely detected calcification in the descending aorta (box), but still correctly detected CAC in the right coronary artery (circle).

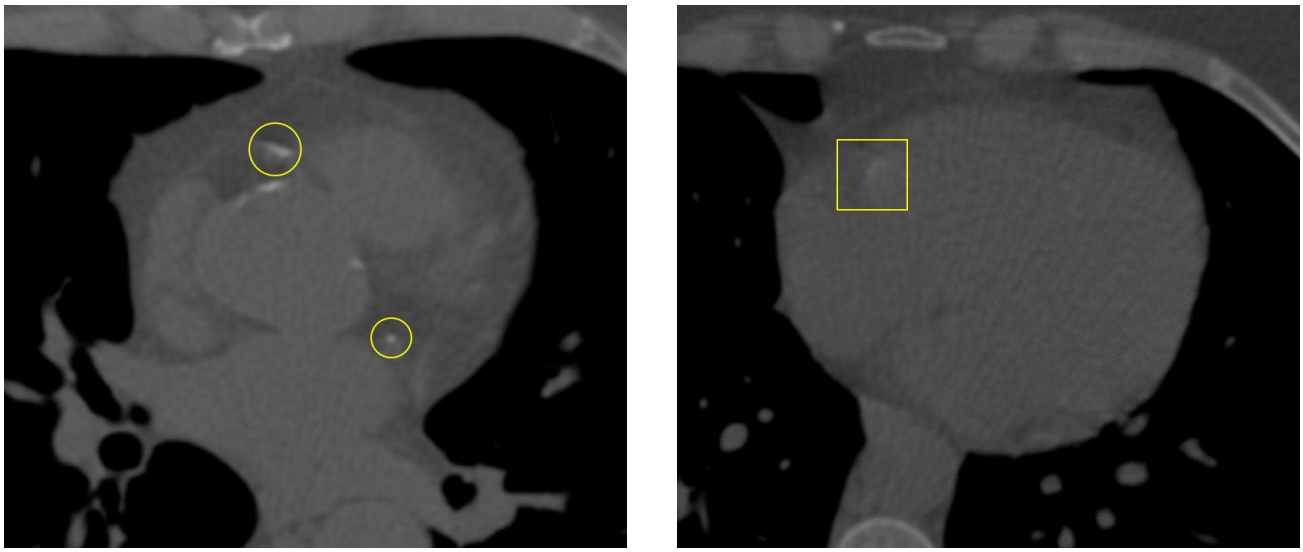

Figure 5. Examples of typical true-positive and false-positive detections. On the left, correctly detected calcifications in the coronary arteries (circles). Aortic calcifications in direct proximity were correctly classified as negatives. On the right, a false-positive detection of an imaging artifact (box) close to the right coronary artery.

accuracy and excellent agreement with manually determined cardiovascular risk categories. Because it relies exclusively on self-learned features and does not rely on prior knowledge or hand-crafted features, it likely adapts well to low-dose chest CT scans from other studies with different acquisition protocols and study population.

ACKNOWLEDGMENTS

This research was kindly supported by NVIDIA Corporation with the donation of a Tesla K40 GPU.

REFERENCES

[1] World Health Organization, "Global status report on noncommunicable diseases 2014," (2014).

[2] The National Lung Screening Trial Research Team, "Reduced lung-cancer mortality with low-dose computed tomographic screening," *New England Journal of Medicine* **365**, 395–409 (2011).

[3] Jacobs, P. C., Gondrie, M. J. A., van der Graaf, Y., de Koning, H. J., Isgum, I., van Ginneken, B., and Mali, W. P. T. M., "Coronary artery calcium can predict all-cause mortality and cardiovascular events on low-dose CT screening for lung cancer," *American Journal of Roentgenology* **198**, 505–511 (2012).

[4] Chiles, C., Duan, F., Gladish, G. W., Ravenel, J. G., Baginski, S. G., Snyder, B. S., DeMello, S., Desjardins, S. S., and Munden, R. F., "Association of coronary artery calcification and mortality in the National Lung Screening Trial: A comparison of three scoring methods," *Radiology* **276**, 82–90 (2015).

[5] Takx, R. A. P., Išgum, I., Willemink, M. J., van der Graaf, Y., de Koning, H. J., Vliegenthart, R., Oudkerk, M., Leiner, T., and de Jong, P. A., "Quantification of coronary artery calcium in nongated CT to predict cardiovascular events in male lung cancer screening participants: Results of the NELSON study," *Journal of Cardiovascular Computer Tomography* **9**, 50–57 (2015).

[6] Mets, O. M., Vliegenthart, R., Gondrie, M. J., Viergever, M. A., Oudkerk, M., de Koning, H. J., Mali, W. P., Prokop, M., van Klaveren, R. J., van der Graaf, Y., Buckens, C. F., Zanen, P., Lammers, J.-W. J., Groen, H. J., Isgum, I., and de Jong, P. A., "Lung cancer screening CT-based prediction of cardiovascular events," *JACC: Cardiovascular Imaging* **6**, 899–907 (2013).

[7] Pinsky, P. F. and Berg, C. D., "Applying the National Lung Screening Trial eligibility criteria to the US population: what percent of the population and of incident lung cancers would be covered?," *Journal of Medical Screening* **19**, 154–156 (2012).

[8] Išgum, I., Prokop, M., Niemeijer, M., Viergever, M. A., and van Ginneken, B., "Automatic coronary calcium scoring in low-dose chest computed tomography," *IEEE Transactions on Medical Imaging* **31**, 2322–2334 (2012).

[9] Xie, Y., Cham, M. D., Henschke, C., Yankelevitz, D., and Reeves, A. P., "Automated coronary artery calcification detection on low-dose chest CT images," *Proc. SPIE* **9035**, 90350F (2014).

[10] Schmidhuber, J., "Deep learning in neural networks: An overview," *Neural Networks* **61**, 85–117 (2015).

[11] Krizhevsky, A., Sutskever, I., and Hinton, G. E., "ImageNet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems* **25**, 1097–1105 (2012).

[12] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A., "Going deeper with convolutions," *Proc. CVPR* (2015).

[13] Rutten, A., Isgum, I., and Prokop, M., "Calcium scoring with prospectively ECG-triggered CT: Using overlapping datasets generated with MPR decreases inter-scan variability," *European Journal of Radiology* **80**, 83–88 (2011).

[14] de Vos, B. D., Wolterink, J. M., de Jong, P. A., Viergever, M. A., and Išgum, I., "2D image classification for 3D anatomy localization; employing deep convolutional neural networks," *Proc. SPIE* **9784** (2016).

[15] Agatston, A. S., Janowitz, W. R., Hildner, F. J., Zusmer, N. R., Viamonte, Jr, M., and Detrano, R., "Quantification of coronary artery calcium using ultrafast computed tomography," *Journal of the American College of Cardiology* **15**, 827–832 (1990).

[16] Landis, J. R. and Koch, G. G., "The measurement of observer agreement for categorical data," *Biometrics* **33**, 159–174 (1977).

[17] Wolterink, J. M., Leiner, T., Takx, R. A. P., Viergever, M. A., and Isgum, I., "Automatic coronary calcium scoring in non-contrast-enhanced ECG-triggered cardiac CT with ambiguity detection," *IEEE Transactions on Medical Imaging* **34**, 1867–78 (2015).