Deep learning for automated exclusion of cardiac CT examinations negative for coronary artery calcium

van den Oever, Leonardus B.; Cornelissen, Ludo; Vonder, Marleen; Xia, Congying; van Bolhuis, Jurjen N.; Vliegenthart, Rozemarijn; Veldhuis, Raymond N. J.; de Bock, Geertruida H.; Oudkerk, Matthijs; van Ooijen, Peter M. A.

Published in:
European Journal of Radiology

DOI:
10.1016/j.ejrad.2020.109114

Publication date:
2020

Deep learning for automated exclusion of cardiac CT examinations negative for coronary artery calcium

Authors:

Leonardus B. van den Oever[1], MSc

Ludo Cornelissen[1], PhD

Marleen Vonder[2], PhD

Congying Xia[3], MD

Jurjen N. van Bolhuis[4], MSc

Rozemarijn Vliegenthart[3], MD PhD

Raymond N. J. Veldhuis[5], PhD

Geertruida H. de Bock[2], PhD

Matthijs Oudkerk[6], MD PhD

Peter M. A. van Ooijen[1], PhD

Affiliations:

1. University of Groningen, University Medical Center Groningen, Department of Radiation Oncology.
2. University of Groningen, University Medical Center Groningen, Department of Epidemiology, Groningen, The Netherlands.
3. University of Groningen, University Medical Center Groningen, Department of Radiology.
4. Lifelines Cohort Study, Groningen, The Netherlands
5. University of Twente, Department of Electrical Engineering, Computer Science and Mathematics.
6. University of Groningen, Faculty of Medical Sciences, Groningen, The Netherlands

Corresponding author: P.M.A. van Ooijen, p.m.a.van.ooijen@umcg.nl, Hanzeplein 1, 9713GZ Groninginge, EB45

Abstract

Purpose: Coronary artery calcium (CAC) score has shown to be an accurate predictor of future cardiovascular events. Early detection by CAC scoring might reduce the number of deaths by cardiovascular disease (CVD). Automatically excluding scans which test negative for CAC could significantly reduce the workload of radiologists. We propose an algorithm that both excludes negative scans and segments the CAC.

Method: The training and internal validation data were collected from the ROBINSCA study. The external validation data were collected from the ImaLife study. Both contain annotated low-dose non-contrast cardiac CT scans. 60 scans of participants were used for training and 2 sets of 50 CT scans of participants without CAC and 50 CT scans of participants with an Agatston score between 10 and 20 were collected for both internal and external validation. The effect of dilated convolutional layers was tested by using 2 CNN architectures. We used the patient-level accuracy as metric for assessing the accuracy of our pipeline for detection of CAC and the Dice coefficient score as metric for the segmentation of CAC.

Results: Of the 50 negative cases in the internal and external validation set, 62% and 86% were classified correctly, respectively. There were no false negative predictions. For the segmentation task, Dice Coefficient scores of 0.63 and 0.84 were achieved for the internal and external validation datasets, respectively.

Conclusions: Our algorithm excluded 86% of all scans without CAC. Radiologists might need to spend less time on participants without CAC and could spend more time on participants that need their attention.

Introduction

Cardiovascular disease (CVD) is one of the major causes of death in the western world. In Europe, 19.9 million new cases of CVD were diagnosed in 2017[1]. 65,5 percent of patients in the Netherlands present with a severe CVD at first diagnosis[2]. Screening on CVD in an early stage is being investigated, since the survival rate for CVD is better at earlier stages than at later stages. [2–5]. It can be done by quantifying coronary artery calcium (CAC), because the amount of CAC is a strong risk marker for future cardiac events, related to underlying coronary atherosclerosis[6]. Screening programs would add a large number of additional scans to be seen by radiologists, due to the large number of eligible participants. The ROBINSCA trial showed that 40% of participants who are at elevated risk for CVD, have no CAC, suggesting even higher percentages of negative scores in general population screening[7]. CAC scoring is currently done *semi*-automatically by selecting calcifications in the coronaries with a CT density ≥130 HU. Automatically excluding the participants without CAC from the workflow would result in an enormous reduction of the total screening workload. The objective of this study is to determine the feasibility to reduce the workload of radiologists and technicians by automatically detecting participants without CAC on non-contrast cardiac CT images without having false negatives by applying deep learning methods. Two different model architectures, with and without dilated convolutional layers, are developed and assessed.

Materials and Methods

Nomenclature

To avoid confusion in the use of nomenclature, we adopt the terminology coined by Liu and Faes et al.[8] for the description of our dataset: a training set is used directly for optimization of the model weights. The internal and external validation sets are used to independently assess model performance.

Study population

The data used for training and internal validation were acquired from the ROBINSCA trial. The ROBINSCA trial was focused on reducing the morbidity and mortality of CVD by detecting the disease in an early stage so that treatment could be started earlier[9]. 13,000 Dutch participants who underwent a non-contrast cardiac CT scan were included in this study. Inclusion criteria were a waist circumference of ≥88cm and ≥102 for women and men, respectively; body mass index of ≥30; family history of CVD, defined as myocardial infarctions or cardiac arrests in first or second degree relatives before the age of 65 years; or current smokers and an age between 45 and 74 for men, and 55 and 74 for women[7,10]. In the present study, 60 participants with a calcium score higher than zero for the training dataset were randomly selected. The low calcium scores are usually the most difficult to detect since the calcifications are small and have a relatively low density close to the threshold of 130 HU. Therefore, we selected 50 participants with a calcium score of zero and 50 with a calcium score between 10 and 20 to check for false negatives for the internal validation.

The external validation dataset was selected from an independent cohort to ensure that the proposed algorithm also works on other data and that it is not over fitting on the training dataset. The external validation dataset was collected from the ImaLife study[11]. The ImaLife study is an embedded study of the Lifelines cohort[12], which was designed to establish references values of imaging biomarkers in the Dutch population for the early stage of the big three diseases : lung cancer, chronic obstructive pulmonary disease and coronary artery disease. Lifelines participants, with age ≥ 45 years, were invited for a low dose CT scan for the heart. The ImaLife study used a different CT system than the ROBINSCA study. The current study sampled 100 ImaLife participants. Similar to the internal validation dataset, we selected 50 participants with a calcium score of zero and 50 with a calcium score between 10 and 20 to test for false negatives.

Scan protocols

The complete scan protocols and designs of the used studies can be found in the design papers of ROBINSCA[9] and ImaLife[11]. We will mention the main differences between the two. In the ROBINSCA trial, second-generation dual source CT was used, while in the ImaLife study, a third-generation dual source CT was used. The Tube voltage was set to 120 kVp in both studies. The tube current was 80 reference mAs in ROBINSCA and 64 reference mAs in ImaLife. The reconstruction kernels were B35f (sharp) and Qr36 (medium-sharp) for the ROBINSCA trial and ImaLife study respectively.

Data annotation and processing

All scans were semi-automatically annotated for CAC by experienced analysts in the University Medical Center Groningen using dedicated software (CaScoring, Syngo.via, version VB30A, Siemens Healthineers). The software colour coded all voxels above the threshold of ≥130 HU. The reader could then select objects that were coronary calcifications and select which coronary it is in (**Figure 1**). These CAC related colour masks were later extracted to binary masks to be used as labels for the convolutional neural network. For the classification task, If the binary mask contained any positive voxels, the participants were classified as positive for CAC. For the segmentation tasks, the binary mask contained the CAC lesions as found by the analysts. The binary mask was, therefore, used in both tasks as the reference.

Automatic Heart segmentation

The CT images are slices of 512 by 512 voxels. To reduce the memory usage of the neural network and reduce the redundant information in the images, we developed an image processing pipeline to crop the images to 320 by 320 voxels around the heart. The algorithm was based on the work of Larrey-Ruiz et al. [13] (**Figure 2)**. By thresholding the image, a binary map of the slice was acquired **(Figure 2B)**. By creating a one-dimensional profile of each column, the thorax wall was deleted. In the one-dimensional profile for each column, when two large objects (>150 voxels) were found near the edges of the image, both were removed. When three large objects were found, i.e., the heart

and the chest wall and back, the outer objects near the edge of the image were removed. Small objects were then removed and the central object was selected as the heart. A bounding box was then selected around the center of gravity of the heart segmentation (**Figure 2C)**. A 3D volume was created by stacking the cropped slices and adding empty slices until the volume consists of 120 slices in the axial direction. From this volume, the axial, sagittal and coronal slices were selected. Therefore, the dimensions of the sagittal and coronal slices were 120x320 voxels.

Architecture

To segment CAC, three dilated convolutional neural networks were used for axial, sagittal and coronal images to reduce the number of false positives. This 2.5D method makes use of more spatial information than only a single network. The design was based on U-net[14] with a number of adaptations (**Figure 3**). First, we included dilation in the first two and last two convolutional layers. Two architectures, one with and one without dilated convolutional layers, were tested to investigate the effect of these layers. Second, the amount of up- and down sampling layers was reduced.
By thresholding the cropped CT at 130HU, we created a binary image. The input of the network consisted of the cropped image slice concatenated with the binary image created by the threshold of 130HU. This focused the network on high intensity regions in which CAC potentially is present. The network output consisted of a binary segmentation of CAC per slice. After prediction, the slices were again combined to create a volume. The procedure was repeated for each of the three view planes, and the three resulting output volumes were then multiplied to create an intersection that served as output for further processing.

Model Training

The three CNNs were trained separately on their respective axial, coronal and sagittal slices. 60 CT image volumes were used to generate 2029 axial, 6544 sagittal and 6423 coronal slices that all contain CAC. The CNNs were trained for 30 epochs and the Dice coefficient was used for the loss

function[15,16]. We used the Adam[15] learning rate starting at $1e^{-5}$. The Adam algorithm helps updating the weight parameters to minimize the loss function. Drop-out was set to 50% of the features after each convolutional layer, meaning that 50% of learned features get randomly dropped to help regularize the CNN and prevent overfitting. The activation function of the convolutional layers is a rectified linear unit (ReLU). The networks were implemented in Keras v2.1.5[17]. Training was performed on the Peregrine high-performance computing cluster of the Center for Information Technology of the University of Groningen.

Post processing

After the predictions of the CNNs, the final volume was created by intersection of the output volumes. This final volume was then post-processed by removing any 2D objects, meaning that objects only visible in one slice in one orientation were removed.

Analysis

Several analysis steps were performed to determine the feasibility of the use of the proposed pipeline for CAC detection. The first analysis was performed to evaluate the effect of the dilated convolutional layers by comparing the accuracies of the two architectures compared to the reference by manual scoring that was done earlier by trained experts. Participants were first classified as either having CAC or not having CAC based on the presence of lesions. If any object present was classified as CAC, the participant was classified as having CAC. Accuracy, specificity, sensitivity and precision were calculated based on these results compared to the reference.

The second analysis was done to evaluate the segmentation of the CAC lesions. The annotations made by the experts were used as a reference for this evaluation. This evaluation was done voxel based, meaning each voxel was classified either as CAC or not CAC. For each participant, a volume containing the segmentations was created. From these volumes, we calculated the false positive

(FP), false negative (FN) and true positive (TP) voxels. These values were then used to calculate the Dice coefficient[18], specificity, precision and sensitivity of the pipeline on segmenting calcification in the internal and external validation dataset. The average volume of TP, FP and FN calcium plaques per scan was also calculated.

The results of the classification step were compared to recent similar studies on CAC[19–22]. Although there are more studies on deep learning algorithms for CAC, many originate from the same imaging research groups. Only the latest results are discussed. These studies usually have different outputs, such as the Agatston score class[20]. Confusion matrices were built based on these studies by using the group that had no CAC as negative group and the group that had the lowest calcium score as the positive class. The confusion matrices from both our own work and the other authors were then used for calculating the accuracy metrics. Bootstrapping with 1000 iterations was used to estimate confidence intervals.

Results

The pipeline without the dilated convolutional layers in the participants without CAC predicted 28 of 50 (56%) and 35 of the 50 (70%) correctly on the internal and external validation datasets, respectively. Of the participants with CAC, 50 of 50 (100%) were classified as positive on the internal validation dataset. On the external dataset, 48 of 50 (96%) participants were classified correctly. Two false negative cases were found in the external validation dataset. After adding the dilated convolutional layers to the CNN, of the participants without CAC in the internal and external validation set, 31 of 50 (62%) and 43 of 50 (86%) were predicted correctly as having no CAC. Of the participants with a 10 to 20 score, no participants were categorized as false negatives. The use of dilated convolutions resulted in an increase in precision, sensitivity, negative predictive value and specificity of the network (**Table 1**). Especially for the external validation, the number of false positives was reduced from 15 to 7 when using dilated convolutional layers. This resulted in an

increase of precision from 0.76 to 0.88 and an increase in specificity of 0.70 to 0.86. The processing pipeline found 19 false positives in the internal validation set and 7 in the external dataset (**Table 2**). **Figure 4-6** show examples of correct and incorrect classification and segmentation results. The processing pipeline achieved a DC score of 0.63 and 0.84 on the internal and external datasets, respectively. The volume of the total CAC lesions in the internal and external datasets was similar, but the number of FP lesions was higher in the internal validation dataset (**Table 3**). With this implementation, which was not optimized for speed, the software pipeline needed approximately 30 seconds to predict the presence of CAC.

Discussion

This research shows that artificial intelligence (AI) might be used for automatically excluding patients without CAC. By using dilated convolutional layers to reduce the number of false positives significantly, a hypothetical workload of 100 CAC scans be reduced by 34, based on a model specificity of 86% and a prior probability for a scan to be CAC positive of 60 with very little to no false negatives. Our algorithm takes 30 seconds and can run without supervision.

A large gap in accuracy between our internal and external validation is found. The internal validation set contained more high CT density spots (HU ≥ 130) actually not being CAC, such as calcium in the valves of the heart or in the aorta. This occurred because it was acquired from a high-risk population, whereas our external validation dataset came from a low- to medium-risk population. Visual inspection showed that the processing pipeline was sensitive to these high CT density spots. Therefore, a higher number of false positives was found in the internal validation dataset. The same was seen in the segmentation results in **Table 3**. Although the number of CAC lesions was similar between datasets, the number of false positive lesions was higher in the internal validation.

Cano-Espinosa et al.[21] use a two step method to directly predict Agatston scores on non-ECG gated chest CT scans. The first step is an object detector for cropping the image around the heart and the second step is a 3D CNN for the regression. They have trained on 5973 scans and used 1000 scans for validation. They categorized the participants into 5 classes depending on the score. The lowest class contains both participants with a zero score as participants with a score lower than 10. Therefore, we can still make a comparison between participants under 10 and between 10 and 100, but we cannot make the comparison zero score and non-zero score. They reach similar precision (0.88) and specificity (0.92) as our work. However, they find 114 false negatives in their predictions on the lowest 2 classes. Therefore, their sensitivity (0.61), Cohen's kappa (0.53) and negative predictive values (NPV) (0.71) are lower than our work.

Wang et al. used also used a 3D deep learning algorithm trained on 530 ECG triggered CT scans to make segmentations of the CAC and then calculate the Agatston scores based on these segmentations. In the validation set of 54 patients with scores between 1 and 99, five were classified as false negative, yielding negative and positive predictive values of 84% and 88%[19]. Overall, on their lowest 2 classification groups, a Cohen's kappa of 0.70 was reached. De Vos et al., using deep learning to directly predict CAC scores from chest or cardiac CT, obtained negative and positive predictive values of 97% and 100%, in line with our work[20], but with higher precision and specificity, but with false negative predictions. Their pipeline uses 3D atlas registration to align the cardiac and chest CT's FOV. Van Velzen et al. used a combination of six datasets containing a combination of 7240 CT scans[23]. They have used a combination of cardiac PET, radiation therapy treatment planning, diagnostic chest, ECG gated CAC screening and low-dose chest CT scans. Validation of the algorithm was done on each dataset containing one of the specific types of CT scans. The algorithm was trained in three different ways for further validation. Either on only the cardiac CT scans, on all the scan protocols, or on the same scan protocol that the validation was done on. By training on the combined dataset of 2563 scans, a kappa of 0.92 was reached. Two false negatives were found on the validation dataset of 323 cardiac CT scans.

A detailed comparison between our results and those reported in the literature is given in **Table 4**. The other groups all used patients with no CAC next to patients with CAC for the training stage of their network. In the other papers, this might have resulted in the false negative predictions. By training on only positive cases, we managed to reduce our false negative predictions to zero. This might make implementation of our software, once validated on a larger dataset, more likely.

Limitations

There are a number of limitations to this research. We only used a limited number of screening scans for training the data. This does not allow the pipeline to learn much anatomical variation. However, even with this small amount of data, the results are promising.

We did not include any participants with a score between 0 and 10 in our cohort. In general, the reproducibility of these cases is low, making them less suitable for this pilot study. In the future, these cases will be included.

During post-processing, 2D objects were removed. This might mean that objects thinner than 3 millimetre in axial direction might have gotten removed. However, no such cases were seen in our validation process. A larger validation study might have to proof whether calcium spots are usually larger than 3 millimetres or whether simply no such cases were present in our validation datasets.

Cases that were misclassified are often participants with calcium in different places than the heart, for instance, the bronchi or the liver. Improving the algorithm for the automatic heart segmentation would help to mitigate this. In **Figure 5**, such an example can be seen.

Other cases often misclassified are participants with calcium in the valves of the heart, as shown in **Figure 6**. For these cases, combining our algorithm with an algorithm for automatic segmentation of the substructures in the heart might result in improved performance.

Our external validation set only consists of 100 participants. Although no false negatives were found, we need to do more validation on a larger dataset to proof that our pipeline can be safely used for detection of participants with no CAC. We will then also include cases with a CAC score <10. Larger validation tests are currently ongoing.

Implications

Assuming a prevalence for CAC of 60% in a screening population at elevated risk, deploying our model would allow for a direct CAC negative classification of 34 out of 100 scans. That implies a 34% reduction in the number of scans due for manual evaluation, and represents a considerable reduction in radiologists' workload in such a screening setting. However, we expect CAC screening may become combined with lung cancer screening[24–26]. This would shift the screening population from elevated risk for CVD to medium risk for CVD, more like as in the ImaLife study. Results of a sample of 3111 male participants of the NELSON study (a trial for lung a cancer screening in a population of heavy smokers)[27] indicate that 79% of the participants would have CAC, so our pipeline might exclude fewer participants than in CAC screening. With these numbers, 17 out of 100 scans might be excluded. Potentially, only a chest CT scan would be made for lung cancer screening, instead of a cardiac ECG-triggered CT, so for future work, we will develop our pipeline for use in thorax scans.

Conclusion

We proposed an automated pipeline for automatically detecting scans containing CAC. The results show that our pipeline in a screening population might be used to exclude scans with no CAC

without the risk of false negatives, and thus might be used to reduce the workload for radiologists in CAC screening.

References

[1] A. Timmis, N. Townsend, C.P. Gale, A. Torbica, M. Lettino, S.E. Petersen, E.A. Mossialos, A.P. Maggioni, D. Kazakiewicz, H.T. May, D. De Smedt, M. Flather, L. Zuhlke, J.F. Beltrame, R. Huculeci, L. Tavazzi, G. Hindricks, J. Bax, B. Casadei, S. Achenbach, L. Wright, P. Vardas, L. Mimoza, G. Artan, D. Aurel, M. Chettibi, N. Hammoudi, H. Sisakian, S. Pepoyan, B. Metzler, P. Siostrzonek, F. Weidinger, T. Jahangirov, F. Aliyev, Y. Rustamova, N.M.A. Mrochak, P. Lancellotti, A. Pasquet, M. Claeys, Z. Kusljugic, L.D. Hudic, E. Smajic, M.P. Tokmakova, P.M. Gatzov, D. Milicic, M. Bergovec, C. Christou, H.H. Moustra, T. Christodoulides, A. Linhart, M. Taborsky, M. Abdelhamid, K. Shokry, P. Kampus, M. Viigimaa, E. Ryödi, M. Niemela, T.T. Rissanen, J.Y. Le Heuzey, M. Gilard, A. Aladashvili, A. Gamkrelidze, M. Kereselidze, A. Zeiher, H. Katus, K. Bestehorn, C. Tsioufis, J. Goudevenos, Z. Csanádi, D. Becker, K. Tóth, P.J. Hrafnkelsdóttir, J. Crowley, P. Kearney, B. Dalton, D. Zahger, A. Wolak, D. Gabrielli, C. Indolfi, S. Urbinati, G. Imantayeva, S. Berkinbayev, G. Bajraktari, A. Ahmeti, G. Berisha, M. Erkin, A. Saamay, A. Erglis, I. Bajare, S. Jegere, M. Mohammed, A. Sarkis, G. Saadeh, R. Zvirblyte, G. Sakalyte, R. Slapikas, K. Ellafi, F. El Ghamari, C. Banu, J. Beissel, T. Felice, S.C. Buttigieg, R.G. Xuereb, M. Popovici, A. Boskovic, M. Rabrenovic, S. Ztot, S. Abir-Khalil, A.C. Van Rossum, B.J.M. Mulder, M.W. Elsendoorn, E. Srbinovska-Kostovska, J. Kostov, B. Marjan, T. Steigen, O.C. Mjølstad, P. Ponikowski, A. Witkowski, P. Jankowski, V.M. Gil, J. Mimoso, S. Baptista, D. Vinereanu, O. Chioncel, B.A. Popescu, E. Shlyakhto, R. Oganov, M. Foscoli, M. Zavatta, A.D. Dikic, B. Beleslin, M.R. Radovanovic, P. Hlivak, R. Hatala, G. Kaliska, M. Kenda, Z. Fras, M. Anguita, A. Cequier, J. Muniz, S. James, B. Johansson, P. Platonov, M.J. Zellweger, G.B. Pedrazzini, D. Carballo, H.E. Shebli, S. Kabbani, L. Abid, F. Addad, E. Bozkurt, M. Kayikçioğlu, M.K. Erol, V. Kovalenko, E. Nesukay, A. Wragg, P. Ludman, S. Ray, R. Kurbanov, D. Boateng, G. Daval, V. De Benito Rubio, D. Sebastiao, P.T. De Courtelary, I. Bardinet, European society of cardiology: Cardiovascular disease statistics 2019, Eur. Heart J. 41 (2020) 12–85. https://doi.org/10.1093/eurheartj/ehz859.

[2] M.A. Heuvelmans, M. Vonder, M. Rook, H.J.M. Groen, G.H. De Bock, X. Xie, M.J. Ijzerman, R. Vliegenthart, M. Oudkerk, H. M.A., V. M., R. M., G. H.J.M., D.B. G.H., X. X., I. M.J., V. R., M.A. Heuvelmans, M. Vonder, M. Rook, H.J.M. Groen, G.H. De Bock, X. Xie, M.J. Ijzerman, R. Vliegenthart, M. Oudkerk, Screening for Early Lung Cancer, Chronic Obstructive Pulmonary Disease, and Cardiovascular Disease (the Big-3) Using Low-dose Chest Computed Tomography: Current Evidence and Technical Considerations, J. Thorac. Imaging. (2018) 1. https://doi.org/http://dx.doi.org/10.1097/RTI.0000000000000379.

[3] R. Detrano, A.D. Guerci, J.J. Carr, D.E. Bild, G. Burke, A.R. Folsom, K. Liu, S. Shea, M. Szklo, D.A. Bluemke, D.H. O'Leary, R. Tracy, K. Watson, N.D. Wong, R.A. Kronmal, Coronary calcium as a predictor of coronary events in four racial or ethnic groups, N. Engl. J. Med. 358 (2008) 1336–1345. https://doi.org/10.1056/NEJMoa072100.

[4] U. Hoffmann, J.M. Massaro, R.B.S. D'Agostino, S. Kathiresan, C.S. Fox, C.J. O'Donnell, Cardiovascular Event Prediction and Risk Reclassification by Coronary, Aortic, and Valvular Calcification in the Framingham Heart Study., J. Am. Heart Assoc. 5 (2016). https://doi.org/10.1161/JAHA.115.003144.

[5] R. Erbel, S. Mhlenkamp, S. Moebus, A. Schmermund, N. Lehmann, A. Stang, N. Dragano, D. Grnemeyer, R. Seibel, H. Klsch, M. Brcker-Preuss, K. Mann, J. Siegrist, K.H. Jckel, Coronary risk stratification, discrimination, and reclassification improvement based on quantification of Subclinical coronary atherosclerosis: The Heinz Nixdorf Recall study, J. Am. Coll. Cardiol. 56 (2010) 1397–1406. https://doi.org/10.1016/j.jacc.2010.06.030.

[6] A.S. Agatston, W.R. Janowitz, F.J. Hildner, N.R. Zusmer, M. Viamonte Jr., R. Detrano,

Quantification of coronary artery calcium using ultrafast computed tomography, J Am Coll Cardiol. 15 (1990) 827–832.

[7] M. Vonder, R. Vliegenthart, M.A. Kaatee, C.M. van der Aalst, P.M.A. van Ooijen, G.H. de Bock, J.W. Gratama, D. Kuijpers, H.J. de Koning, M. Oudkerk, High-pitch versus sequential mode for coronary calcium in individuals with a high heart rate: Potential for dose reduction, J. Cardiovasc. Comput. Tomogr. 12 (2018) 298–304. https://doi.org/10.1016/j.jcct.2018.02.005.

[8] X. Liu, L. Faes, A.U. Kale, S.K. Wagner, D.J. Fu, A. Bruynseels, T. Mahendiran, G. Moraes, M. Shamdas, C. Kern, J.R. Ledsam, M.K. Schmid, K. Balaskas, E.J. Topol, L.M. Bachmann, P.A. Keane, A.K. Denniston, A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis, Lancet Digit. Heal. (2019). https://doi.org/10.1016/S2589-7500(19)30123-2.

[9] M. Vonder, C.M. van der Aalst, R. Vliegenthart, P.M.A. van Ooijen, D. Kuijpers, J.W. Gratama, H.J. de Koning, M. Oudkerk, Coronary Artery Calcium Imaging in the ROBINSCA Trial: Rationale, Design, and Technical Background, Acad. Radiol. 25 (2018) 118–128. https://doi.org/10.1016/j.acra.2017.07.010.

[10] S.J. Denissen, C.M. van der Aalst, M. Vonder, M. Oudkerk, H.J. de Koning, Impact of a cardiovascular disease risk screening result on preventive behaviour in asymptomatic participants of the ROBINSCA trial, Eur. J. Prev. Cardiol. (2019) 204748731984339. https://doi.org/10.1177/2047487319843396.

[11] C. Xia, M. Rook, G.J. Pelgrim, G. Sidorenkov, H.J. Wisselink, J.N. van Bolhuis, P.M.A. van Ooijen, J. Guo, M. Oudkerk, H. Groen, M. van den Berge, P. van der Harst, H. Dijkstra, M. Vonder, M.A. Heuvelmans, M.D. Dorrius, P.P. De Deyn, G.H. de Bock, A. Dotinga, R. Vliegenthart, Early imaging biomarkers of lung cancer, COPD and coronary artery disease in the general population: rationale and design of the ImaLife (Imaging in Lifelines) Study, Eur. J. Epidemiol. (2019). https://doi.org/10.1007/s10654-019-00519-0.

[12] S. Scholtens, N. Smidt, M.A. Swertz, S.J.L. Bakker, A. Dotinga, J.M. Vonk, F. Van Dijk, S.K.R. Van Zon, C. Wijmenga, B.H.R. Wolffenbuttel, R.P. Stolk, Cohort Profile: LifeLines, a three-generation cohort study and biobank, Int. J. Epidemiol. 44 (2015) 1172–1180. https://doi.org/10.1093/ije/dyu229.

[13] J. Larrey-ruiz, J. Morales-sánchez, M.C. Bastida-jumilla, R.M. Menchón-lara, R. Verdú-monedero, J.L. Sancho-gómez, Automatic image-based segmentation of the heart from CT scans, EURASIP J. Image Video Process. 52 (2014) 1–13.

[14] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: Lect. Notes Comput. Sci. (Including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), 2015: pp. 234–241. https://doi.org/10.1007/978-3-319-24574-4_28.

[15] D.P. Kingma, J.L. Ba, Adam: A method for stochastic optimization, in: 3rd Int. Conf. Learn. Represent. ICLR 2015 - Conf. Track Proc., International Conference on Learning Representations, ICLR, 2015.

[16] L.R. Dice, Measures of the Amount of Ecologic Association Between Species, Ecology. 26 (1945) 297–302. https://doi.org/10.2307/1932409.

[17] F. Chollet, Keras, (2015). https://github.com/fchollet/keras.

[18] W.R. Crum, O. Camara, D.L.G. Hill, Generalized overlap measures for evaluation and validation in medical image analysis, IEEE Trans. Med. Imaging. 25 (2006) 1451–1461.

https://doi.org/10.1109/TMI.2006.880587.

[19] W. Wang, H. Wang, Q. Chen, Z. Zhou, R. Wang, H. Wang, N. Zhang, Y. Chen, Z. Sun, L. Xu, Coronary artery calcium score quantification using a deep-learning algorithm., Clin. Radiol. (2019). https://doi.org/10.1016/j.crad.2019.10.012.

[20] B.D. de Vos, J.M. Wolterink, T. Leiner, P.A. de Jong, N. Lessmann, I. Isgum, Direct Automatic Coronary Calcium Scoring in Cardiac and Chest CT, IEEE Trans. Med. Imaging. 0062 (2019) 1–12. https://doi.org/10.1109/TMI.2019.2899534.

[21] C. Cano Espinosa, G. González, G.R. Washko, M. Cazorla, R.S.J. Estépar, Automated Agatston score computation in non-ECG gated CT scans using deep learning, in: Proc. SPIE--the Int. Soc. Opt. Eng., SPIE-Intl Soc Optical Eng, 2018: p. 91. https://doi.org/10.1117/12.2293681.

[22] S.A.M. Gernaat, S.G.M. van Velzen, V. Koh, M.J. Emaus, I. Išgum, N. Lessmann, S. Moes, A. Jacobson, P.W. Tan, D.E. Grobbee, D.H.J. van den Bongard, J.I. Tang, H.M. Verkooijen, Automatic quantification of calcifications in the coronary arteries and thoracic aorta on radiotherapy planning CT scans of Western and Asian breast cancer patients, Radiother. Oncol. 127 (2018) 487–492. https://doi.org/10.1016/j.radonc.2018.04.011.

[23] S.G.M.M. van Velzen, N. Lessmann, B.K. Velthuis, I.E.M.M. Bank, D.H.J.G.J.G. van den Bongard, T. Leiner, P.A. de Jong, W.B. Veldhuis, A. Correa, J.G. Terry, J.J. Carr, M.A. Viergever, H.M. Verkooijen, I. Išgum, Deep learning for automatic calcium scoring in CT: Validation using multiple cardiac CT and chest CT protocols, Radiology. 295 (2020) 66–79. https://doi.org/10.1148/radiol.2020191621.

[24] H.S. Hecht, C. Henschke, D. Yankelevitz, V. Fuster, J. Narula, Combined detection of coronary artery disease and lung cancer, Eur. Heart J. 35 (2014) 2792–2796. https://doi.org/10.1093/eurheartj/ehu296.

[25] P.C. Jacobs, I. Isgum, M.J. Gondrie, W.P. Mali, B. van Ginneken, M. Prokop, Y. van der Graaf, Coronary artery calcification scoring in low-dose ungated CT screening for lung cancer: interscan agreement, AJR Am J Roentgenol. 194 (2010) 1244–1249. https://doi.org/10.2214/ajr.09.3047.

[26] L. Fan, K. Fan, Lung cancer screening CT-based coronary artery calcification in predicting cardiovascular events: A systematic review and meta-analysis, Med. 97 (2018) e10461. https://doi.org/10.1097/md.0000000000010461.

[27] R.A.P. Takx, I. Išgum, M.J. Willemink, Y. van der Graaf, H.J. de Koning, R. Vliegenthart, M. Oudkerk, T. Leiner, P.A. de Jong, Quantification of coronary artery calcium in nongated CT to predict cardiovascular events in male lung cancer screening participants: Results of the NELSON study, J. Cardiovasc. Comput. Tomogr. 9 (2015) 50–57. https://doi.org/10.1016/j.jcct.2014.11.006.

Figure 1: An example of an annotation in Syngo.via. In pink are the pixels above the 130 HU threshold. The blue, yellow and red colors indicate CAC in different coronaries in the heart.

Figure 2: Automated image cropping by image post-processing as developed by the authors. (A) Original CT slice. (B) A threshold is applied to create a binary image. (C) Region selection after removal of chest wall by removing object near the edges of the image and small objects. This image is used for finding the center of mass of the heart. The bounding box is then selected by using the center of mass as center for the bounding box. (D) Cropped image after processing.

Figure 3:Schematic overview of the Architecture of the axial neural network. On the left, two examples of the input, the upper image is the cropped CT image and beneath that is the thresholded binary image. The grey boxes represent the feature maps, with the x and y sizes corresponding with the number of channels and the size of the image respectively. The coloured arrows show the operations used. The number of features for the coronal and sagittal CNNs are the same, but he input images are 120x320. This is down sampled to 60x160 and 30x80 in the lower layers.

Figure 4: Example of a correct segmentation. From left to right, the cropped CT image (A), the thresholded image (B), the segmentation as made by an expert reader (C) and the prediction of the pipeline(D) are shown.

Figure 5: Example of an incorrect segmentation result. The pipeline predicts calcium in the aorta and in the cartilage of the bronchi to be CAC. From left to right, the cropped CT image (A), the thresholded image (B), the segmentation as made by an expert reader (C) and the prediction of the pipeline(D) are shown.

Figure 6: Example of an incorrect segmentation result. The pipeline predicts calcium in the aortic valve to be CAC. From left to right, the cropped CT image (A), the thresholded image (B), the segmentation as made by an expert reader (C) and the prediction of the pipeline(D) are shown.

Table 1: Effect of using dilated convolutions in the CNN on the results of the pipeline. The confidence intervals are given after the plusminus symbol. Confidence intervals are acquired by bootstrapping to 1000 participants and calculating the confidence intervals over the cohort.

	Internal Validation		**External Validation**	
	Without Dilation layers	With Dilation layers	Without Dilation layers	With Dilation layers
Precision	0.69±0.10	0.72±0.10	0.76±0.10	0.88±0.10
Sensitivity	1.00±0.00	1.00±0.00	0.96±0.04	1.00±0.00
Negative predictive value	1.00±0.00	1.00±0.00	0.95±0.14	1.00±0.00
Specificity	0.55±0.14	0.61±0.14	0.70±0.12	0.86±0.11

Table 2: Confusion matrix of the positive or negative for CAC analysis.

	Internal Validation		**External Validation**	
	Predicted: No CAC	Predicted: CAC	Predicted: No CAC	Predicted: CAC
Reference: No CAC	31	19	43	7
Reference: CAC	0	50	0	50
	κ = 0.62		κ = 0.86	

CAC = coronary artery calcium, κ = Cohen's kappa coefficient

Table 3: Summary of the validation and testing performance of the proposed pipeline on lesion segmentation in participants with a positive CAC.

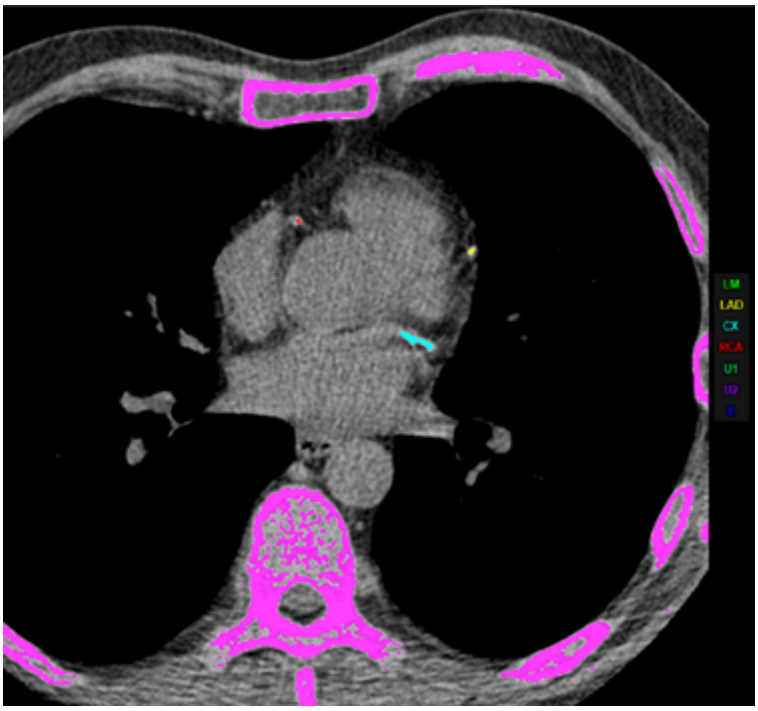
	Internal Validation	**External Validation**
DC score	0.63	0.84
TP (mm^3)/scan	31.76	34.75
FP (mm^3)/scan	91.29	39.78
FN (mm^3)/scan	7.76	5.77

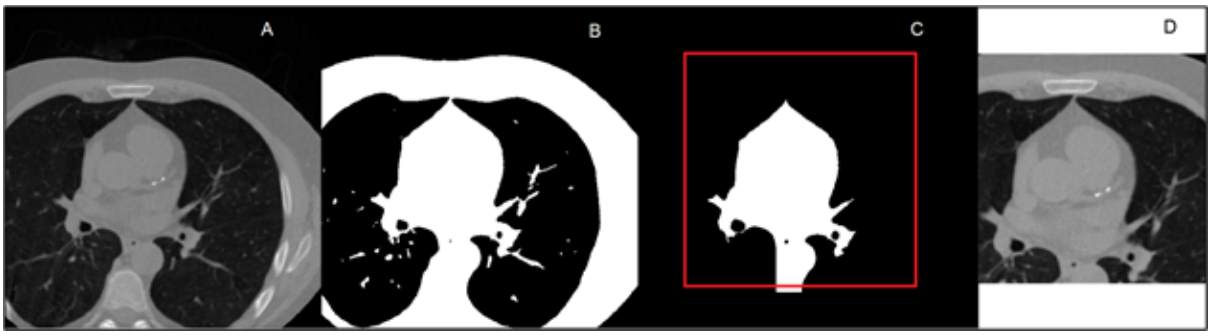
DC = Dice Coefficient, TP = True Positive, FP = False Positive, FN = False Negative

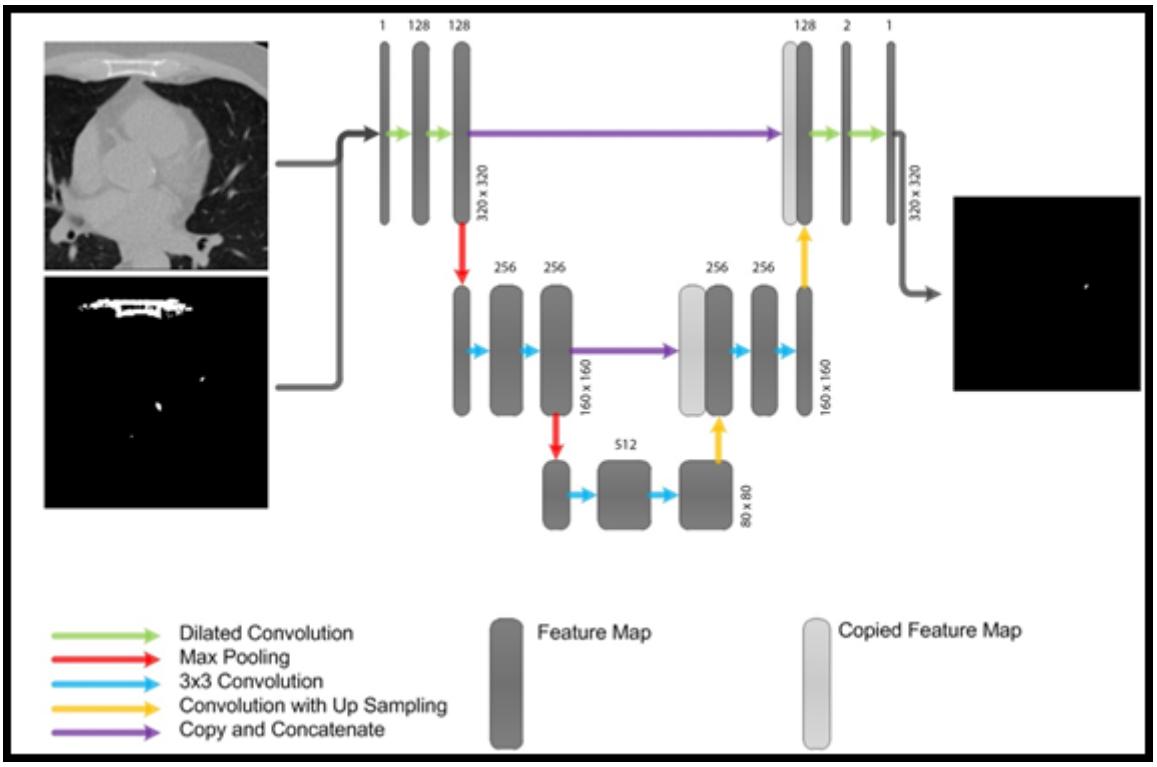
Table 4: results of the pipeline compared to similar works.

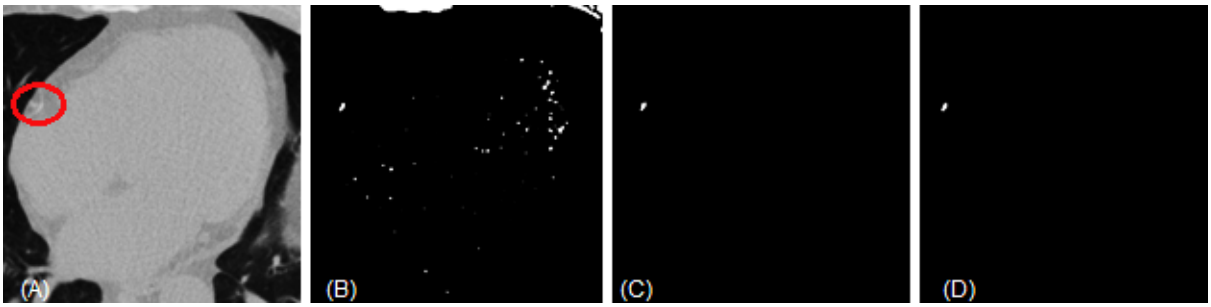
	Proposed work	**Cano-Espinosa et al.**[21]	**Wang**[19]	**De Vos**[20]	**Van Velzen**[23]
Precision	0.88±0.10	0.88±0.04	0.88±0.08	1.00±0.00	0.90±0.08
Sensitivity	1.00±0.00	0.62±0.06	0.91±0.08	0.80±0.12	0.95±0.05
NPV	1.00±0.00	0.70±0.05	0.84±0.14	0.97±0.02	0.99±0.01
Specificity	0.86±0.11	0.92±0.03	0.79±0.14	1.00±0.00	0.98±0.02
TP	50	180	49	36	47
FP	7	25	7	0	5
FN	0	114	5	9	2
TN	43	279	26	259	269
kappa	0.86	0.53	0.7	0.87	0.92

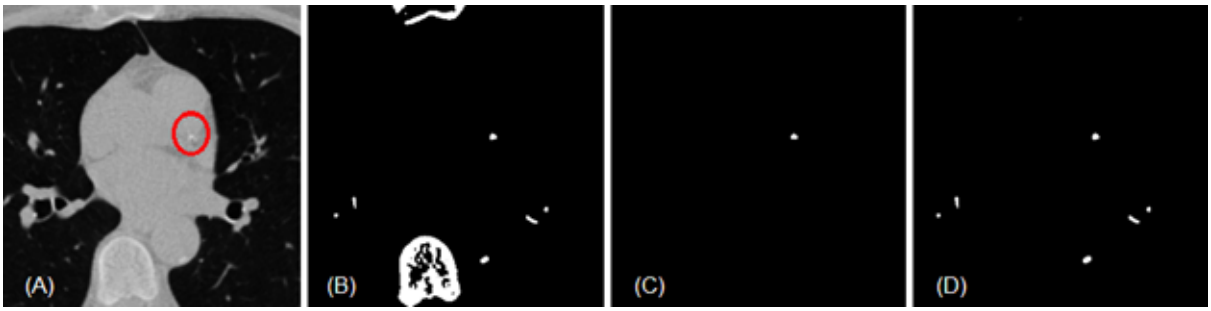
NPV = negative predictive value, TP = True Positive, FP = False Positive, FN = False Negative, kappa = Cohen's kappa coefficient

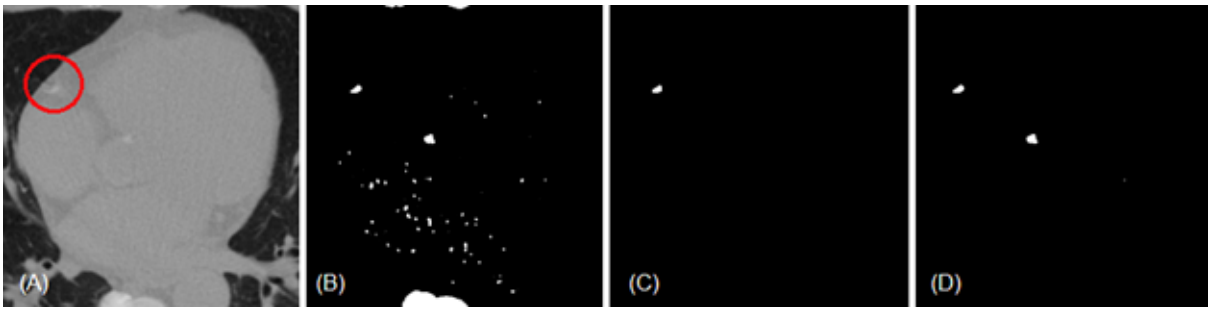

LM
LAD
CX
RCA
U1
U2


A
B
C
D


1
128
128
128
2
1
320 x 320
320 x 320
256
256
256
256
160 x 160
160 x 160
512
80 x 80
Dilated Convolution
Max Pooling
3x3 Convolution
Convolution with Up Sampling
Copy and Concatenate
Feature Map
Copied Feature Map


(A)
(B)
(C)
(D)


(A)
(B)
(C)
(D)


(A)
(B)
(C)
(D)