Deep learning-based detection of coronary artery calcification in non-contrast and contrast-enhanced CT scans

Salman Mohammadi
salman.mohammadi@mre.medical.canon

Canon Medical Research Europe

Shadia Mikhael
Canon Medical Research Europe

Keith Goatman
Canon Medical Research Europe

Sonia Dahdouh
Canon Medical Research Europe

Olivier Jaubert
Canon Medical Research Europe

Jeremy Voisey
Canon Medical Research Europe

Rebecca Hughes
NHS Golden Jubilee National Hospital

Conor Bradley
University of Glasgow, NHS Golden Jubilee National Hospital

Richard Good
University of Glasgow, NHS Golden Jubilee National Hospital

Article

Keywords:

Posted Date: April 29th, 2024

DOI: https://doi.org/10.21203/rs.3.rs-4281908/v1

License: © This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: No competing interests reported.
Abstract

Coronary artery calcification (CAC) assessed using computed tomography (CT) scans is a clinically-validated biomarker that is highly prognostic for coronary heart disease (CHD) and adverse cardiac events. Clinical assessment of CAC relies on a dedicated coronary electrocardiogram (ECG)-synchronised non-contrast CT scan. However, millions of CT scans are acquired every year for various indications that include the heart in the field-of-view yet visible CAC is often not reported in these scans. This is a significant missed opportunity for incidental detection of a powerful cardiac risk factor. Our study was conducted on a set of 295 unselected, consecutive CT scans from the National Health Service (NHS) Golden Jubilee Hospital. These were annotated for CAC and used for model training and testing. We developed and validated a deep learning model to accurately quantify CAC on any CT scan including the heart, regardless of the presence or phase of contrast agent, reason for the scan, or use of ECG-synchronisation. The model achieved substantial agreement with the manual human assessment (Cohen's Kappa=0.61, Bland-Altman mean difference=−40.8mm³). Additionally, we found no correlation between arterial brightness (a surrogate metric for the level of contrast agent present) and agreement between manual and automated measurements (Spearman correlation R=−0.005). Early intervention is vital to improve patient outcomes. The automated CAC scoring method demonstrated here could be applied to all chest CT scans that include the heart, greatly expanding the opportunities for early detection of subclinical cardiovascular disease and the potential for large-scale implementation to reduce the burden of coronary heart disease through systematic, opportunistic CAC screening.

Introduction

Coronary heart disease (CHD) remains the leading cause of death worldwide. According to the World Health Organisation, an estimated 17.9 million deaths in 2019 were due to CHD, accounting for 32% of global deaths. Many effective treatments are available for CHD, but these require timely intervention. Primary prevention strategies targeting CHD are largely based on opportunistic assessment of 10 year cardiovascular risk using risk scoring models such as QRISK or ASSIGN, which rely on pooled risk factor equations. However, these risk scores suffer from poor risk stratification and poor risk discrimination in diverse groups.

Coronary artery calcification (CAC) confirms atherosclerotic disease, and predominately occurs within the arterial intimal layer as a result of smooth muscle cell apoptosis and macrophage infiltration. CAC may be the strongest predictor of future cardiovascular events and a better guide to the introduction of primary preventative therapies than traditional risk factor models. The “gold standard” method to assess CAC is a computed tomography (CT) scan specifically calibrated for the detection of calcium: non-contrast, electrocardiogram (ECG) synchronised cardiac scans with heart rate control, and breath holding during the scan to reduce motion of the heart. These scan acquisition parameters reduce variation and artefacts in images and ensure reliable quantitation of calcium volume. Conventionally, the scan is reconstructed into 3mm thick slices and calcium is annotated semi-automatically using a threshold of 130 Hounsfield units (HUs) to produce a cumulative Agatston Score, which is the current standard prognostic measure of risk from CAC burden.

Despite the prognostic value of Agatston scores, image reconstruction parameters and the nominal cut-off of 130 HU were derived in the early 1990s. CT scanner technology and post processing techniques have markedly changed since this time, including multidetector scanners now capable of imaging the whole heart in a fraction of a heartbeat. As a result, CAC can also be detected on non-cardiac CT scans that include the heart in the field of view, which were not acquired for coronary indications. Such “incidental findings” are often not reported but they have significant implications for patients. For example, it could be used to prompt clinical review, including screening for risk factors (e.g., hypertension, dyslipidaemia, diabetes mellitus, and obesity), to provide advice regarding lifestyle measures proven to reduce the risk of cardiovascular events (e.g., smoking cessation, weight loss, diet, exercise etc.), and prompt consideration of disease modifying therapies such as statins. Such intervention following detection of incidental coronary calcification on ungated, non-cardiac chest CT scans has been shown to improve patient engagement with secondary prevention therapies.

Automated CAC scoring techniques have historically focused on speeding up the calculation of an Agatston score in routine calcium screening workflows. However, recent studies have shown excellent agreement for automated detection of CAC in scans for non-coronary indications on non-ECG-synchronised, non-contrast chest CT scans (including positron emission tomography [PET-CT]). However, these scans represent a small portion of CT scans. Many CT scans are acquired following administration of exogenous intravenous contrast agent for non-cardiac indications and may also have a wider field of view than just the chest (e.g., chest, neck, abdomen, and pelvis). Contrast agent alters the brightness of vessels and other tissues rendering the traditional threshold of 130 HU inapplicable. Though several techniques have been demonstrated on contrast-enhanced ECG-gated, dedicated cardiac CT angiography (CCTA) scans, there are currently no automated methods for CAC quantification on contrast-enhanced, non-ECG-synchronised, non-Coronary CT scans. Furthermore, while several studies have successfully developed automated CAC quantification techniques for both ungated, non-contrast chest CT and CCTA scans separately, there is currently no reported method for automated CAC quantification on both ungated non-coronary non-contrast chest CT, and ungated non-coronary contrast-enhanced CT using a single deep learning model.

In this paper, we describe the development of a deep learning model to detect and quantify CAC on any ungated CT scan that includes the whole heart in the field of view, irrespective of the presence of contrast. We developed our model using only reference ground truth of CAC, without the use of additional ground truth annotation methods.
of annotations of coronary arteries or coronary structures and reported accurate risk stratification of patients. We demonstrate that deep learning models can reliably quantify CAC across a spectrum of contrast phases in contrast-enhanced scans without being adversely affected by relative brightness levels in coronary vessels. This study supports opportunistic cardiovascular risk screening by expanding the pool of CT scans that can be automatically screened for evidence of cardiovascular disease. This could enable early, targeted intervention for at risk, asymptomatic populations who otherwise remain undetected.

Results

Cohort Details

In this study, we developed a single deep learning method that automates the quantification of CAC using a single model for both non-contrast and contrast-enhanced CT. The model does not require ECG-synchronisation or scans acquired for coronary indications and is agnostic to contrast phase. Our study was conducted on a set of 295 unselected CT scans from the National Health Service (NHS) Golden Jubilee Hospital that included the heart in the field of view. Scans were annotated for calcification in the coronary arteries, aortic root, aortic valve, and mitral annular by three clinical annotators.

A summary of the CT scan dataset is provided in Table 1. The majority of scans were undertaken for the investigation of cancer (165/295, 56%) and lung disease (107/295, 36%). The field-of-view varied from chest only (104/295, 35%) to neck, chest, abdomen, and pelvis (10/295, 3%). The dataset comprised both non-contrast scans and contrast scans with varying brightness in the vasculature and tissues. At the time of data export, 221 (75%) were documented on the radiology report as containing contrast.
Table 1
Patient demographics. All continuous values are reported as mean ± standard deviation. Imaging protocol abbreviations are: Chest Abdomen Pelvis (CAP), Chest High Resolution (HR), Chest Abdomen (CA), Neck Chest Abdomen Pelvis (NCAP). **Other imaging protocols included 3 CT Chest Neck, and 1 CT Neck Chest Abdomen.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>All (N = 295)</th>
<th>Train/validation (N = 214)</th>
<th>Test (N = 81)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (years)</td>
<td>64 ± 12</td>
<td>65 ± 12</td>
<td>62 ± 1</td>
</tr>
<tr>
<td>Female Sex (%)</td>
<td>54.2%</td>
<td>56%</td>
<td>49.3%</td>
</tr>
<tr>
<td>CT Scanner Manufacturer</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Canon</td>
<td>124 (42%)</td>
<td>94</td>
<td>30</td>
</tr>
<tr>
<td>Siemens</td>
<td>170 (57.6%)</td>
<td>120</td>
<td>50</td>
</tr>
<tr>
<td>GE Medical Systems</td>
<td>1 (&lt; 1%)</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Imaging Protocol</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CT CAP</td>
<td>128 (43%)</td>
<td>99</td>
<td>29</td>
</tr>
<tr>
<td>CT Chest</td>
<td>73 (25%)</td>
<td>56</td>
<td>17</td>
</tr>
<tr>
<td>CT Chest HR</td>
<td>32 (11%)</td>
<td>18</td>
<td>14</td>
</tr>
<tr>
<td>CT CA</td>
<td>47 (16%)</td>
<td>33</td>
<td>14</td>
</tr>
<tr>
<td>CT NCAP</td>
<td>11 (4%)</td>
<td>7</td>
<td>4</td>
</tr>
<tr>
<td>Other**</td>
<td>4 (1%)</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>Reconstruction Kernel</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Body Sharp</td>
<td>106 (35%)</td>
<td>82</td>
<td>24</td>
</tr>
<tr>
<td>Lung</td>
<td>18 (5%)</td>
<td>12</td>
<td>6</td>
</tr>
<tr>
<td>Br40f</td>
<td>170 (57.6%)</td>
<td>120</td>
<td>50</td>
</tr>
<tr>
<td>Standard</td>
<td>1 (&lt; 1%)</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Tube Voltage (kV)</td>
<td>105.2 ± 14.5</td>
<td>105.7 ± 14.6</td>
<td>104 ± 14</td>
</tr>
<tr>
<td>80</td>
<td>4 (1%)</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>90</td>
<td>123 (42%)</td>
<td>88</td>
<td>35</td>
</tr>
<tr>
<td>100</td>
<td>19 (7%)</td>
<td>11</td>
<td>8</td>
</tr>
<tr>
<td>110</td>
<td>12 (4%)</td>
<td>8</td>
<td>4</td>
</tr>
<tr>
<td>120</td>
<td>137 (46%)</td>
<td>104</td>
<td>33</td>
</tr>
<tr>
<td>Current (mA)</td>
<td>290.7 ± 164.8</td>
<td>290.9 ± 163.5</td>
<td>290.2 ± 169</td>
</tr>
<tr>
<td>Contrast present (%)</td>
<td>75%</td>
<td>78%</td>
<td>65%</td>
</tr>
<tr>
<td>Slice thickness (mm)</td>
<td>0.74 ± 0.04</td>
<td>0.74 ± 0.04</td>
<td>0.73 ± 0.05</td>
</tr>
<tr>
<td>Annotator</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R.I.S.G.</td>
<td>202 (68.5%)</td>
<td>139</td>
<td>63</td>
</tr>
<tr>
<td>C.P.B.</td>
<td>48 (16.3%)</td>
<td>38</td>
<td>10</td>
</tr>
<tr>
<td>R.K.H.</td>
<td>45 (15.2%)</td>
<td>37</td>
<td>8</td>
</tr>
</tbody>
</table>

Assessment of automated CAC detection model and annotator agreement

We investigated agreement between and within annotators, and between annotators and our model, which we summarise in Table 2. Inter-observer and intra-observer variability was assessed on 18/295 (6%) and 30/295 (10%) of scans, respectively. The model was evaluated using a held-out test set of 81 patients.
Table 2
Agreement between and within annotators, and between the model and annotators. Values are reported as mean ± standard deviation where relevant.

<table>
<thead>
<tr>
<th>Annotator</th>
<th>Intra-Observer Variability</th>
<th>Cohen's Kappa</th>
<th>ICC</th>
<th>F1 (%)</th>
<th>Sensitivity (%)</th>
<th>PPV (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>-48.9 [-878.3, 780.48]</td>
<td>0.91</td>
<td>0.92</td>
<td>80.1 ± 12.5</td>
<td>66 ± 31.5</td>
<td>79.3 ± 27.4</td>
</tr>
<tr>
<td>B</td>
<td>-92.1 [-342.9, 160.5]</td>
<td>0.10</td>
<td>0.77</td>
<td>66.3 ± 10</td>
<td>43.5 ± 15.1</td>
<td>80.9 ± 35.2</td>
</tr>
<tr>
<td>C</td>
<td>-150.9 [-1024.5, 772.5]</td>
<td>0.73</td>
<td>0.80</td>
<td>75 ± 19.2</td>
<td>72.6 ± 24.3</td>
<td>84.4 ± 16.2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Annotators</th>
<th>Inter-Observer Variability</th>
<th>F1 (%)</th>
<th>Sensitivity (%)</th>
<th>PPV (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A and B</td>
<td>163.5 [-517.8, 844.8]</td>
<td>0.58</td>
<td>0.83</td>
<td>70.6 ± 20.5</td>
</tr>
<tr>
<td>B and C</td>
<td>-265.6 [-1136.7, 605.5]</td>
<td>0.65</td>
<td>0.76</td>
<td>60 ± 15.1</td>
</tr>
<tr>
<td>A and C</td>
<td>-102.0 [-1063.4, 859.2]</td>
<td>0.72</td>
<td>0.76</td>
<td>67.2 ± 14.7</td>
</tr>
<tr>
<td>Model</td>
<td>-40.8 [-519.4, 437.7]</td>
<td>0.61</td>
<td>0.94</td>
<td>58.6 ± 19.1</td>
</tr>
</tbody>
</table>

Detection of CAC was assessed using a combination of Bland-Altman analysis and Interclass Coefficients (ICC) ratings. Bland-Altman mean and 95% limits of agreement (LOA) are reported. Cohen's Kappa was used to assess agreement of volume score risk category classification of patients into five volume score risk categories (0–4, for volume scores of < 1, 1–112, 112–400, 401–999, >= 1000, respectively). We report the F1-score, sensitivity, and positive predicted value (PPV), which are aggregated over each voxel for a given patient and averaged over all coronary labels (i.e., the intersection between true calcified voxels and false positive voxels). Figure 1 and Fig. 2 show Bland-Altman agreement plots for intra-observer and inter-observer agreement, respectively. Figure 3 shows Bland-Altman agreement plots between our model and annotators.

To better understand model performance, we assessed agreement between the model and annotators Bland-Altman analysis of the volume score (in mm³). For all coronary territories, we found a mean offset of -40.8 and 95% LOA as [-519.4, 437.7]. For each coronary territory: left main stem (LMS) -4.96 [-190.1, 180.2], left anterior descending (LAD) -53.1 [-447.16, 340.8], circumflex (Cx) 0.12 [-162.5, 162.7], right coronary artery (RCA) -18.5 [-772, 734.8], and non-coronary regions - 277.5 [-2731.4, 2176.3]. We provide per-vessel Bland-Altman plots in Supplementary Fig. 4.

Fourteen (17%) patients, initially reported without any calcification, had calcium predicted by the model. After manual review, two patients were found to have calcification that was missed by the annotators but correctly predicted by the model. The twelve remaining false positive cases included four with metal artefacts in the heart, and two where only non-coronary calcification was reported. The remaining six patients presented with features including noisy scans and swirling contrast, bright artefacts near the heart. We include qualitative examples of false positive predictions in Fig. 4. Our method reported CAC in all patients where calcium was present.

We additionally report our voxel-wise metrics and across volume score risk categories in Table 3. For classification of patients into volume score risk categories, we found substantial agreement (Kappa = 0.61) between predicted CAC and reference CAC using Cohen's Kappa statistic, with a specificity of 94.8% (± 2.37), sensitivity of 72.3% (± 20.8) PPV of 71.5% (± 21.9), and NPV of 94.6% (± 2.68). We provide qualitative examples of accurate model predictions in Fig. 5. We provide a confusion matrix for agreement between the model and annotators for volume score risk classification in Supplementary Fig. 2.

Table 3
Metrics across the volume score risk-categories. Mean values are reported with standard deviations.

<table>
<thead>
<tr>
<th>Automated Model Voxel-wise Metrics</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Volume Score risk category</strong></td>
</tr>
<tr>
<td>No risk (&lt; 1)</td>
</tr>
<tr>
<td>Average Risk (1-112)</td>
</tr>
<tr>
<td>Moderate Risk (112–400)</td>
</tr>
<tr>
<td>High Risk (401–999)</td>
</tr>
<tr>
<td>Very High Risk (&gt; 1000)</td>
</tr>
</tbody>
</table>

Effect of contrast on automated calcium detection

To explore how our method will perform in the incidental detection of CAC in any CT scan including the heart, we investigated the effect of contrast agent on the ability of the model to accurately detect CAC. We report the difference in reported CAC between annotators and our automatic
We found that model performance was not affected by contrast agent concentration in the coronary arteries. No significant correlation was found between brightness in organs and the volume score agreement between annotators and our model. Spearman's rank correlation for the brightness in the aorta and pulmonary artery was -0.005 (p=0.62), and spleen and liver, 0.02 (p=0.8). Bland-Altman analysis agreement in volume score between annotators and our model for contrast patients shows a mean offset and LOAs in mm$^2$ of -54.8 [-465.1, 355.5], and for non-contrast patients, -14.4 [-597.3, 568.5]. A Mann-Whitney U-test did not show a significant difference (p=0.84) in mean difference between contrast (N=53) and non-contrast (N=28) scans.

Discussion

There is rising interest in incidental identification of biomarkers to opportunistically screen for risk of future adverse events\textsuperscript{19,47}. Significant evidence supports CAC as a predictor of future atherosclerotic events\textsuperscript{14,48–50}, particularly in otherwise asymptomatic populations, but remains under-reported\textsuperscript{17,18}. This is supported by a SCCT/STR guideline which recommends reporting moderate and severe CAC on all patients undergoing lung cancer screening\textsuperscript{51}. Several studies have developed automated techniques for CAC scoring in non-contrast CT scans acquired for non-coronary indications\textsuperscript{27,29–31,33,32,52,53}. A recent study from Teng et al.\textsuperscript{14} manually reported the prevalence of CAC in both contrast and non-contrast, non-ECG synchronised, non-coronary chest CT. Whilst multiple studies have developed capabilities for automatically quantifying CAC in CCTA scans\textsuperscript{25,34,35,37–44,54}, our study is the first automated technique capable of quantifying CAC in both contrast and non-contrast, non-ECG synchronised, non-coronary CT using a single model.

This study shows that a single deep learning model can reliably detect and quantify CAC in any CT scan that include the heart in the field of view, irrespective of the presence of contrast. We evaluated the model's performance on a single held-out test set, demonstrating robust detection of CAC and stratification of patients into established risk categories\textsuperscript{46}. We found that agreement between the model and annotators was not did not correlate with brightness in the heart (Spearman coefficient $\rho = 0.005$), and found no significant difference in annotator and model agreement between contrast and non-contrast scans. Therefore, our model is invariant to contrast levels and is capable of quickly detecting CAC on scans which may otherwise require careful thresholding to differentiate calcium from surrounding tissue.

In the absence of the established 130HU threshold for calcium in non-contrast scans, there is currently no standardised procedure for manually annotating CAC in contrast scans. A common method is to estimate a per-scan threshold using HU statistics within a region of interest in the ascending or descending aorta, followed by manual annotation of voxels above this threshold, or semi-automatic labelling of whole calcium lesions identified by thresholding and connected component analysis\textsuperscript{34,38,41,44,55}. Other studies have used free-form voxel annotation without the use of thresholds using arbitrary manual adjustment of windowing levels\textsuperscript{36,38,43}. Our annotation procedure involves first manually identifying a per-scan threshold, and then manually annotating voxels above this threshold. This reduces the time spent on tedious editing and/or rejection of calcium lesions from metal artefacts identified by semi-automatic thresholding. Similarly, using connected-component analysis to identify calcium lesions makes it impossible to differentiate calcium lesions which span multiple anatomical regions (e.g., across the aortic valve, LCA, and RCA). This work benefits from precise reference annotations for calcium in specific coronary regions.

Compared with other techniques which automate CAC quantification on CCTA, our method did not require ground truth for the coronary arteries\textsuperscript{37,38} or coronary regions\textsuperscript{25}, large datasets for domain adaptation learning\textsuperscript{41}, or rely on reference ground truth generated from virtual non-contrast scans\textsuperscript{42}. More generally, our method was developed without the use of cardiac atlases for registration\textsuperscript{22,25}, thus able to implicitly identify coronary anatomy and discern between calcification in different locations in both non-contrast and contrast scans. To reduce the FOV of scans to just the coronary region, we used the open-source TotalSegmentator\textsuperscript{56} tool. Additionally, to eliminate false positive calcium detections outside the heart, a heart region mask generated using an in-house multi-atlas heart segmentation tool was used.

There are important limitations to this study. The small size of our dataset which was acquired using only two scanners from a single institution, constrains the scope of our generalizations. The retrospective design of this study also limits the clinical applicability of our findings. However, our cohort consists of unselected sequential scans. Furthermore, our model was trained on a set of homogenous high-resolution and thin-slice scans, which are typical of acquisitions from modern scanners. In a preliminary investigation on lower resolution, thick-slice data, we found similar results indicating strong generalizability on lower resolution scans which are traditionally used for calcium scoring. The inclusion criteria for this study included scans with the whole heart in the field of view. In practice, any calcium detected in scans where the heart is partially visible presents a valuable opportunistic finding.

The automatic method developed in this study reported CAC in fourteen scans where no calcium was reported. Two of these scans detected only non-coronary calcification, which does not adversely affect CAC risk scores. In the remaining twelve scans, false positive calcium detection was due to artefacts including pacemaker wires, poor quality scans (e.g. high noise or motion), abnormal brightness near the heart (e.g. pericardial calcification), or calcification missed by annotators which was retrospectively found to be correctly identified by the model. Since the model had few examples of pacemaker wires during training, they were more prevalent in the test set, it was not expected that the model would robustly generalize to these cases. Additionally, in our intra-observer study we found two cases (Supplementary Fig. 1) where annotators had reported no calcification in
one attempt, and positively identified calcification in another, for the same scan. Robust identification of coronary calcium can be difficult in the presence of noisy scans, significant artefacts, and non-coronary calcification.

In conclusion, we have developed the first automatic technique capable of detecting CAC invariant to contrast in non-ECG synchronised CT scans. The general applicability of our method further reduces barriers for large-scale incidental screening which can utilise the greater prognostic value of CAC. This work can utilise CT scans across a variety of indications and modalities, thus can take advantage of the millions of CT scans acquired annually for any indication. Patients who are asymptomatic, or missed by traditional risk models, would benefit from early detection and preventative therapies, which could help significantly reduce the global health burden of CHD.

Methods

Ethics and governance approval

The study was undertaken at the NHS Golden Jubilee Hospital, a tertiary cardiothoracic centre that provides regional CT imaging for a population of over two million patients in the west of Scotland. This study was conducted in accordance with guidance from the UK Policy Framework for Health and Social Care Research. The project was approved by Yorkshire and The Humber – Leeds East Research Ethics Committee (REC Ref 21/YH/0217).

Scan selection

Using the Picture Archiving and Communication System (PACS) in our institution, we identified CT scans between March 2022 and March 2023 that included the heart in the field of view. The clinical indication for each scan was recorded from the electronic care record together with patient demographics (age and sex). The scan field of view and use of contrast were also recorded. All scans were anonymised and the DICOM file transferred to a remote workstation. Details of additional scan parameters, including CT reconstruction kernel, x-ray tube voltage, and reconstructed slice thickness, were noted from the relevant DICOM tags. When multiple scans were available for a single patient, we chose the thinnest-slice scan (for greater ground-truth annotation accuracy) with the softest reconstruction kernel (for minimum noise), and the patient in a supine position.

Ground Truth Annotations

Following curation, scans were uploaded to a dedicated annotation platform with a custom workflow and image annotation tool which enabled automatic and smooth transition of cases between curator, annotator, reviewer, and quality assurer. Three clinicians (R.H., C.B., R.G.), each with more than two years of experience annotating coronary calcification on CT scans, underwent training on both the annotation protocol and annotation platform, on at least 15 datasets each. The remaining scans were then randomly assigned to each of the three. The most senior of the three annotators had more than ten years of Cardiac CT reporting experience (R.G.) and acted as reviewer. Annotators and reviewer were blinded to radiology reports, patient details such as medical history, scan indication, and scan parameters during the annotation process.

Annotators identified calcium in seven defined regions of interest: the left main stem artery (LMS), left anterior descending artery (LAD), left circumflex artery (Cx), right coronary artery (RCA), aortic valve (AV), aortic root (AR, assumed below the sinotubular junction), and the mitral valve annulus (MV). Annotation was performed manually on axial slices using a threshold-assisted brush. No maximum threshold was advised but rather annotators gradually increased the lower threshold until there were few highlighted pixels within the main chambers (atria and ventricles) to minimise annotation of contrast in the vasculature.

Annotators additionally assessed for the presence of contrast in the scan (yes/no), whether the scan was of sufficient quality for interpretation, and the presence of any artefacts which may affect scan interpretation, e.g., motion, permanent pacemakers (including biventricular pacemaker), valve replacement, central venous lines, or evidence of previous cardiac surgery (including coronary artery bypass grafting). The senior annotator reviewed a random 40% of the image annotations to confirm consistency. Quality assurance was conducted on all cases by S.S.M and S.M. to ensure completeness of the data and compliance to the annotation protocol. Annotations which contained overlapping labels for a given voxel were merged by assigning voxels to labels in the following order of priority: LMS, LAD, RCA, Cx, AR, AV, MA. This ordering was consistent across all patients.

Scan summary data

An unselected series of 320 CT scans, acquired between March 2022 and March 2023, with the heart in the field of view (FOV) were identified, anonymised, and downloaded to the annotation platform. A total of 17 scans were excluded during curation (Fig. 7) due to unsuitable acquisition series, unusable reconstructions, or incomplete coverage of the heart, which made manual annotations of scans difficult. The remaining 303 scans were assigned for annotation, of which annotators excluded a further 4 scans due to severe artefact and/or extensive non-coronary calcification (Fig. 8). The reviewer excluded a further two scans: one scan with situs inversus and one scan with a biventricular pacemaker lead. Two further scans were excluded during the quality assurance (QA) process due to ambiguity around the presence of contrast and severe artefact, leaving a total of 295 scans available for development and testing. All exclusion criteria were applied before model training and testing. Our final cohort represents a wide variety of scan quality, imaging artefacts, and foreign bodies (Supplementary Fig. 3, Supplementary Table 1). Supplementary Table 2 details the various indications for the scans.
Scans were divided into training and validation sets (214 scans) and a testing set (81 scans) (Table 1). Scans were primarily acquired from two scanners, a Canon Medical Systems Aquilion ONE scanner, and a Siemens SOMATOM go Top scanner. A single scan was acquired from a GE Medical Systems Revolution EVO scanner. Scans were reconstructed using soft and sharp kernels, including the soft-tissue Br40f and Standard kernels, as well as sharper Lung and Body Sharp kernels. This indicates the variety in the dataset which may be encountered in incidental findings.

We used stratified random sampling based on the presence of calcium in specific coronary territories to assign patients to the training, validation, and testing splits, thus ensuring territory classes were balanced. We initially split the dataset between training and testing, then repeated the procedure in the training split to create three folds for cross-validation, resulting in three folds of the same dataset with 143 training and 71 validation scans each, and 81 testing scans.

Presence of contrast was indicated on the radiology reports of 221/295 (75%) scans. To understand the effect of contrast agent on CAC detection, we estimated the levels of contrast in cardiac, vascular, and the solid organ components of the scan. We used the open-source TotalSegmentator tool to automatically segment the spleen, liver, aorta, and main pulmonary artery. The median HU values in these regions were used as a surrogate for contrast phase as follows. Early-stage contrast scans can be identified by higher HU values in the aorta and pulmonary arteries than in the spleen and liver (early arterial phase). As contrast diffuses through the vascular system, the liver and spleen parenchyma become brighter. Thus, later stage contrast scans can be recognised by higher HU values in these organs as well (portal venous phase).

The variation in brightness of these predefined elements in our scans is shown in Fig. 9. We estimate the majority (199/295, 67%) of our contrast scans were acquired during the portal venous phase, as we observe enhanced attenuation in both the liver and spleen, and the aorta and pulmonary artery, compared to the non-contrast baseline (dashed lines, Fig. 9). A small number of scans (8/295, 2%) were indicated as contrast scans but presented without enhancement in the organs or arteries considered; we identified these as delayed contrast scans. We identified the remaining contrast scans as early arterial phase (14/295, 4.7%), as they showed enhanced attenuation in the aorta and pulmonary artery over non-contrast but did not show significant enhancement in the liver and spleen.

Importantly, we observe a wide variety of attenuation in the coronary region, which presents challenges for the detection of coronary calcification. Each scan will have varying thresholds at which calcium is significantly differentiated from coronary vessels and surrounding tissue. This makes manual annotation difficult and time-consuming, particularly early-stage contrast scans where calcium may not be differentiated from contrast-enhanced vessels. This highlights the value of an automated technique for detecting calcium that is robust to contrast levels.

Manual annotation results

Figure 10 illustrates the distribution of calcification across coronary territories in our cohort. No coronary calcification was found in 90/295 (30%) scans. Of these, 80/295 (27%) showed no evidence of any calcification, including non-coronary calcification. The remaining 10/295 (11%) had only non-coronary calcification. Of the 205 scans with coronary artery calcification, the most common artery was the LAD (191/295, 64%), followed by the RCA (135/295, 45%), the circumflex artery (125/295, 42%), and the LMS (102/295, 34%). Aortic root calcification was the most common non-coronary calcified region (133/295, 45%), followed by the aortic valve (76/295, 25%), and the mitral annular (58/295, 19%).

To assess annotator agreement, we assigned 18/295 (6%) cases to annotators for inter-observer variability, and 30/295 (10%) for intra-observer variability. Annotators were blinded to previous attempts and followed the same annotation procedure for both attempts. For intra-observer attempts, we ensured at least a three-month gap between an annotator’s first and second attempt at annotating a case.

Artificial Intelligence Algorithm Development

CT scans acquired for non-coronary indications can have wide FOVs, including the neck and pelvis. To reduce the unnecessarily large FOV we crop the images to the heart region in the axial-plane using the TotalSegmentator tool. This significantly improves efficiency and reduces the time for training the model.

Our algorithm uses two CNNs successively in stages, processing the CT volume in 3D patches. The first stage CNN produces a segmentation of calcium, including coronary and non-coronary territories. The second stage CNN refines the output from the first stage network and classifies each calcium voxel to its coronary territory (LMS, LAD, Cx, RCA, or non-coronary arteries AR, AV, and MA). While there is evidence for the prognostic value of calcification in non-coronary structures, our primary use for detecting calcification in these regions was to help eliminate the possibility of the model including confusing CAC for non-coronary artery calcification. We reduce false-positive predictions by post-processing model predictions. Firstly, we used an in-house multi-atlas heart segmentation tool to generate heart masks to reduce false positives, by excluding any voxels predicted by the models outside of the heart mask. Secondly, we use 3D connected-component labelling and discard lesions smaller than 3mm³. We tuned the minimum voxel volume size on the validation sets. Finally, whole-volume CAC segmentations are obtained by stitching adjacent 3D model output patches.

Algorithm Architecture and Training

We used an encoder-decoder based U-Net architecture for both the first and second CNN stages (Fig. 12). The encoder layers consist of two convolutional blocks which down-sample the input. The decoder layers consist of two convolutional blocks followed by a transposed convolution
which use skip-connections from the encoder layers at the corresponding levels. Input CT scans are intensity windowed (window level and width of 200 and 1200 HU respectively), and then cropped to the coronary region. During training we randomly sample input with and without calcium present, and augment patches using random rotation, and random adjustment of voxel intensity values. Both models used early stopping and the Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$, $\epsilon = 1e^{-8}$, and a learning rate of $3x10^{-3}$.

Coronary calcium is very sparse. To prevent the model simply predicting no calcium, we use a combination of Dice and focal loss to downregulate the contribution of the background voxels. The Dice loss was weighted by 0.5. For the focal loss hyperparameters, we used a $\gamma$ of 2 for both stages. The first stage model was trained with $\alpha = 6900$, and the second stage with $\alpha = [1.5, 1800, 1400, 1600, 1400, 1400]$, for background, LMS, LAD, Cx, RCA, and non-coronary classes, respectively. We calculated the per-class inverse voxel frequency as the weighting term for each class. Final model predictions during inference were generated by optimising the models for each cross-validation fold, and ensembling the best performing models by voxel-wise probability averaging.

Statistics

All analysis was done in Python (Version 3.9.15) using numpy (Version 1.23.5). Bland-Altman means and 95% limits of agreement were used to evaluate per-vessel and total coronary agreement between annotators, and between the automated model and annotators. Spearman's rank correlation or Mann-Whitney U tests were used to compare continuous data using the scipy package (Version 1.9.3). Cohen's Kappa was used to evaluate agreement in classification of patients in the five volume risk categories also using the scipy package. Cohen's Kappa was interpreted as: (< 0.00): poor; (0.00–0.20): slight; (0.21–0.40): fair; (0.41–0.60): moderate; (0.61–0.80): substantial; (0.81–1.00): almost perfect agreement. Volume score classification confusion matrices were created using scikit-learn (Version 1.1.3). We report the voxel-based F1-Score, positive predictive value by averaging over all voxels for a given scan, for all coronary labels. Voxel based metrics were calculated using the MONAI package (Version 1.1.0). All statistical tests were two-sided. Statistical significance was defined as a $p$ value of < 0.05. Intraclass Coefficients were calculated using the Pingouin package (Version 0.5.3).

Declarations

Author Contributions

This work would not have been possible without dedication and support from all authors. Salman Mohammadi designed, implemented, and executed all experiments. Richard Good was the primary clinical contributor, oversaw annotator training and data collection and contributed to the manuscript. Keith Goatman provided scientific oversight. Richard Good and Keith Goatman conceived and proposed the study. Shadia Mikhael provided clinical oversight over the whole annotation process, including curation, review, quality assurance, administration of scans across the annotators, annotator training, and review of model results. Jeremy Voisey and Salman Mohammadi supported data curation and processing. Rebecca Hughes and Conor Bradley supported annotations of reference ground truth. Sonia Dahdouh provided scientific support and extensive manuscript reviews. Olivier Jaubert contributed to development of the pre-processing methods used and provided manuscript reviews.

Competing Interests

Six of the authors (S.M, S.S.M, K.A.G, S.D, J.P.V, O.J) are employees of Canon Medical Research Europe. There are no competing interests in this study.

Data Availability

The raw patient data is not publicly available to protect patient confidentiality. However, reasonable data requests will be considered. After approval from the hospital and the corresponding authors, de-identified CT data may be provided. Requests to access the datasets should be directed to corresponding author.

Code Availability

The underlying code for this study is not publicly available as it includes proprietary code.

References


49. Peng, A. W. et al. Very High Coronary Artery Calcium (≥1000) and Association With Cardiovascular Disease Events, Non–Cardiovascular Disease Outcomes, and Mortality: Results From MESA. Circulation 143, 1571–1583 (2021).


**Figures**

**Figure 1**

Bland-Altman plots for intra-observer variability between each of the 3 annotators’ attempts. Outliers are indicated as orange points.

**Figure 2**

Bland-Altman plots for inter-observer variability between the 3 annotators. Outliers are indicated as orange points.

**Figure 3**

Bland-Altman plot showing agreement between annotators and our automatic method across the held-out test set. Agreement is reported on the volume score in 100mm$^3$. Outliers are indicated in orange.
Figure 4

Examples of model false positives caused by artefacts due to: pacemaker leads (a, b) as non-coronary calcification (red) and right coronary artery calcification (green), and doubling of the LCA from motion falsely reported as circumflex calcification (c).

Figure 5

Model outputs (top) compared with manually annotated ground truth (bottom) in three different patients (a, b, c) showing calcification in the LAD (yellow), RCA (green), Cx (blue), and non-coronary arteries (red). Accurate detection is shown in both non-contrast (left, right), and contrast-enhanced (middle) scans.
Figure 6

Agreement between annotators and our automatic technique against brightness of organs: the aorta and pulmonary artery (a), and spleen and liver (b). Separate 95% limits of agreement and mean values for contrast (orange) and non-contrast (blue) are shown as dashed lines. Contrast scans include portal venous, early arterial, and delayed contrast phase scans.

Figure 7

Flow diagram summarizing the process by which scans were excluded from the study.
Figure 8

Examples of excluded scans, for: situs inversus (a), pacemaker artefact and air present in aorta (b), heavy calcification, artefact due to pacemaker, heavy calcification and artefact (c), and coronary artery bypass graft (d).

Figure 9

Distribution of contrast phase across the training and test datasets. Median brightness (HU) is measured in the pulmonary artery and aorta and summed and compared against the sum of median brightness in the spleen and liver. Dashed lines indicate the highest brightness for each of these regions in the non-contrast set.
Figure 10
Total and coronary calcium volumes per-patient, for all patients with calcification present (a); each patient is indicated by a bar, black bars indicate total volume, and shaded bars indicate coronary volume. Prevalence of calcification per coronary-territory in patients (b) - bars indicate number of patients with any volume of calcification indicated in the coronary territory.

Figure 11
Pipeline for automated detection of CAC. CT scans are initially cropped to the coronary region using TotalSegmentator to localise the heart. Cropped scans are then used to generate heart masks, which are used to remove false positives from model predictions. The first-stage network segments all calcification and the second-stage network generates the final per-vessel segmentations. We created an end-to-end deep learning pipeline for fully automated detection and quantification of CAC in coronary and non-coronary territories in any chest CT scan (Figure 11). Working on both contrast-enhanced and unenhanced CT scans, the method was developed without the use of cardiac atlases or additional manual segmentations of the coronary arteries or coronary structures. The method is based on two convolutional neural networks (CNNs) that produce precise segmentations of calcium labelled per-voxel with their territory (LMS, LAD, Cx, RCA, or non-coronary structures).
Figure 12

Model architecture diagram. Dashed lines indicate skip connections, which are concatenated with the output from up-sampling connections. Model output channels, N, indicate number of classes predicted (N=1 for first stage, and N=6 for second stage). Shared parameters for convolution blocks are indicated by the grey cuboid. The output size is indicated adjacent to each level.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- DeeplearningbaseddetectionofcoronaryarterycalcificationinnoncontrastandcontrastenhancedCTscansSupplementarySubmission08042024.docx