# Reviewer #1 Comments

*Overview:*

This study is focused on an important limitation of Agatston score, which is a calcium thresholding defined as 130 HU. Black et al utilizes an integrated intensity and a volume fraction methods to overcome the above mentioned limitation. Topic is of significant interest, as due development of CT technology a new method of calcium scoring is needed. I have the following questions, concerns and/or suggestions for the authors to consider:

*Methods:*

1. Why were simulation phantoms used? And if Authors have an opportunity to use simulation phantom, why only one CT vendor was simulated? Wouldn’t it be more informative to present results from different CT systems? It is essential to increase the number of CT vendors to show clinical relevance. As already known, there is a discrepancy between CAC scored from different CT systems, therefore an integrated intensity and a volume fraction should be validated on different CT systems from different vendors.
   1. Thanks for this question. Simulation allows for a more controlled environment to test the limits of these new CAC scoring techniques. A simulation study was chosen as a precursor to physical phantom studies and, eventually, patient studies to limit the number of potential confounding variables involved in the comparison. The simulation package we have developed over the past few years does not include multiple different CT vendors. However, we are working on upgrading to a new simulation system that includes many different vendors for future studies.
2. I do believe that slice thickness should be the same. I do understand that Agatston methodology was created on 3.0 mm slice thickness, but just for the clarity of the results there should be added one column (Table 1.) presenting the same very same acquisition and reconstruction settings for all investigated methods. Otherwise, Authors enhance the results of simulated phantom.
   1. Thank you for pointing this out. Slice thickness in this simulation is limited to 0.5 mm, which makes us unable to include multiple slice thickness parameters without significant adjustments to the simulation package. Although, we are currently continuing this work on physical phantoms with varied slice thicknesses, including 3.0 mm. The discussion section was adjusted further to highlight this study's limitation of slice thickness. We also added a reference in the discussion, demonstrating good correlation between Agatston scoring at 0.5 mm and the original Agatston scoring at a slice thickness of 3.0 mm.

Slice thickness plays an important role in calcium scoring, and traditional Agatston scoring is only defined at a slice thickness of 3 mm. Recent studies have shown that the accuracy and sensitivity of Agatston scoring are improved when slice thickness is decreased 29. Hou et al. demonstrated good correlation between Agatston scoring on a 3 mm slice thickness image compared to Agatston scoring on a 0.5 mm slice thickness image 30. Our simulation was limited to 0.5 mm slice thickness which is expected to provide more accurate and sensitive comparisons for Agatston scoring. Nonetheless, future studies might provide insights by varying the slice thickness, and this study is limited without a direct comparison to Agatston scoring at the gold standard slice thickness of 3 mm.

1. What is the size of simulated calcifications? This is not explained in the manuscript, therefore it is difficult to assess the performance of presented methods.
   1. Thank you for addressing this. The insert diameters are listed in the methods section as 1, 3, and 5 mm. The inserts were each 1.5 mm in length. This detail was not included in the original manuscript, and a sentence has been added to the methods section explaining this.

Three calcification inserts of different diameters (1, 3, and 5 mm), each with a length of 1.5 mm, and different hydroxyapatite (HA) densities were placed within each phantom.

1. In the point 2.3 Agatston scoring. Did authors also correct for the weighting factor when changing kV? Or only for the HU threshold for calcium detectability? The factor for 120 kV is defined as follows: 130‐199 HU, factor 1; 200‐299 HU, factor 2; 300‐399 HU, factor 3; and ≥ 400 HU, factor 4. If the threshold is changed, the factor also should be adjusted.
   1. Thank you for pointing this out. We added more details in the methods section explaining the correction for the weighting factor related to the kV-dependent threshold.

Additionally, the weighting factor was adjusted in a similarly kV-dependent manner, according to criteria outlined in Gräni, C. et al., and extrapolated for a tube voltage of 135 kV 19.

1. Is there any possibility to utilise the dynamic phantom for this study? As we know, motion creates motion artefacts which strongly affects Agatston score calculation. As motion was only simulated and was not reflected in physical phantom.
   1. The dynamic phantom images acquired by Praagh et al. are also limited. The motion is linear and therefore does not produce realistic motion artifacts like one would see in the heart. Also, the calcium rods included in the motion images were limited to large inserts (5 mm) and higher densities (> 200 mg/cc). This simulation study focuses more on the low-density regime since Agatston scoring is already reasonably effective for large, high-density calcifications. Because of the limitations of the previously acquired dynamic images, we are currently working on a dynamic phantom study using the CIRS Dynamic Cardiac Phantom. This future study will include much more realistic cardiac motion and inserts within the size and density regime that Agatston scoring is ineffective at measuring.

*Results:*

1. How did author compare 25 and 50, as by definition, these densities cannot be detected with 130 HU?
   1. Thanks for this question. We automatically segmented a region of interest, using a technique proposed by Praagh et al., that contains the calcification since the calcification locations are known based on geometry. The ground truth masses of these calcifications are known ahead of time. The small diameter calcifications (1 mm) and low-density calcifications (25, 50, and 100 mg/cc) demonstrate the limitations of Agatston scoring compared to the proposed calcium quantification techniques, as they often produce false-negative (CAC=0) scores.
2. Fig 4 – please change the scale of y‐axis, it is difficult to see values. Is it possible to analyze densities separately? What the small, medium, and large insert mean? Is it about phantom size or about calcification size? Please, specify. I assume it is a patients’ size.
   1. Thank you, the scale has been changed to make it less cluttered.
   2. Two different sets of calcifications were analyzed (1) the low-density set (25, 50, and 100 mg/cc) and (2) the normal-density set (200, 400, and 800 mg/cc). Figure 4 shows the low-density inserts separate from the normal-density inserts. This is the area with the most room for improvement when quantifying calcium mass via Agatston scoring.
   3. The small, medium, and large inserts correspond to the diameter of the inserts, 1, 3, and 5 mm diameters, respectively. This has been added to every relevant figure caption.

Fig. 4 Shows the linear regression analysis comparing measured calcium to the known calcium for the low-density (25, 50, 100 mgHAcm-3) stationary phantoms. Every tube voltage (80, 100, 120, 135 kV) and size (small, medium, large) is included in the analysis. (A) shows the results of integrated calcium mass. (B) shows the results of the volume fraction method. (C) shows the results of Agatston mass scoring. The best fit line, along with the root mean squared error (RMSE) and root mean squared deviation (RMSD) values are shown in each plot. The small, medium, and large inserts label corresponds to the diameter of the insert (1, 3, and 5 mm, respectively).

1. I assume there is an error in tables, as it shows 0/216 false negatives for Agatston score
   1. Thank you for that correction, this has been fixed.
2. Why the false‐negatives were excluded from reproducibility analysis? I do not see a rational reason for this, it should be included.
   1. Thanks for this question. We decided to exclude the false negatives because including the false negative scores would reward the techniques with the most false-negative scores regarding reproducibility. A calcium scoring method that produces 100% false negative scores would be perfectly reproducible, but this wouldn’t provide too much insight into the effective reproducibility of the technique in our view.
3. May you display results in graphs (i.e. Fig 5) separately, by CAC density, radiation dose, and patient’s size?
   1. We can include these figures in the supplemental material?

# Reviewer #2 Comments

*Overview:*

Thank you for the opportunity to review this manuscript entitled 'Coronary artery calcium mass measurement based on integrated intensity and volume fraction techniques. The paper describes an interesting and relevant topic.   
Please consider the following general comments on this paper:

*General Comments:*

1. With the recommended Agatston threshold of 130 HU, together with the recommended noise level of about 24 HU, wouldn't simply lowering the Agatston scoring threshold to a lower level increase CAC sensitivity? This is also shown in Table 3. Please comment on this option, also in view of the proposed method in the current study and it's implementation in the clinic.
   1. Thank you for this question. This is correct, but that approach would also increase the false-positive ratio. The spatially weighted calcium score is a sophisticated way of attempting to improve the sensitivity of Agatston scoring. It successfully does that without sacrificing specificity, as shown in this study. However, it lacks in other ways (reproducibility, physical interpretation).
2. Please elaborate on the impact of missing Compton-scattering in the used simulation.
   1. Compton scattering is one of the leading causes of scattered radiation in a material. It occurs due to the interaction of free electrons in the material with X-rays. Compton scattering is directly proportional to the physical density of a material and only weakly related to the X-ray energy. The methods section has been updated to explain this better.

The simulation did not include Compton scatter, the dominant attenuation mechanism in CT imaging due to the interaction of free electrons with the incoming X-ray, but beam hardening was included.

1. For the simulated phantom, which material was used in conjunction with the pure HA to obtain the different densities? Was it a combination of solid water or with muscle?
   1. Thank you for clarifying this. The second material was heart tissue and the methods section has been updated to make this more clear.

Three calcification inserts of different diameters (1, 3, and 5 mm), each with a length of 1.5 mm, and different hydroxyapatite (HA) densities were placed within each phantom. A combination of HA and myocardium was used to vary the calcification densities.

1. Please provide an overview of CT number measurements between the simulated and real phantom, for (at least) the fat ring, material surrounding the calcification, all calcification densities, calibration disks.
   1. Thank you for this suggestion. A table that shows those regions' CT numbers in the methods section has been added.
2. Paragraph 2.3: heading title contains a typo (Agaston)
   1. Thank you for catching this. This has been fixed.
3. For both techniques, were scores calculated based 3 mm reconstructions? Or calculated based on other slice thicknesses, and subsequently recalculated to 3 mm equivalent scores?
   1. Thank you for this insightful question. Hou et al. showed with a slice thickness of 0.5 mm, the 130-HU threshold is still appropriate for 120 kV scans. Those authors also demonstrated the importance of kV specific thresholds for 120, 100, and 80 kV scans. Therefore, for our study we incorporated kV specific Agatston scores and kept the rest of the algorithm identical to the original algorithm defined at 3 mm slice thickness. The discussion section has been updated to make this more clear.

Slice thickness plays an important role in calcium scoring, and traditional Agatston scoring is only defined at a slice thickness of 3 mm. Recent studies have shown that the accuracy and sensitivity of Agatston scoring are improved when slice thickness is decreased 29. Hou et al. demonstrated good correlation between Agatston scoring on a 3 mm slice thickness image compared to Agatston scoring on a 0.5 mm slice thickness image 30. Our simulation was limited to 0.5 mm slice thickness which is expected to provide more accurate and sensitive comparisons for Agatston scoring. Nonetheless, future studies might provide insights by varying the slice thickness, and this study is limited without a direct comparison to Agatston scoring at the gold standard slice thickness of 3 mm.

# Editor's Comments

*Overview:*

I concur with two expert reviewers. The topic is of significant interest and there is a need for a new CAC scoring standard. However, there are some concerns raised by both reviewers. Please consider these comments/suggestions carefully and address them thoroughly.

* + - * Thank you for considering this manuscript. We believe all of the reviewer comments have been sufficiently addressed and the original manuscript has been updated to reflect this. Including updated figures with consistent significant figures throughout the manuscript and a simpler figure layout.