Response to referee
We thank the reviewer for his thoughtful comments. Please find below the reviewer comments verbatim and our responses in italic.

Reviewer: 1

Comments:
The authors have generally clarified most of my previous concerns.

However, I still remain unclear regarding the inclusion of the prediction model in this paper. It is a distraction from the main aim of the paper, which is about diagnostic accuracy of CT angiography, MR angiography etc. The authors are reporting a lot of information which deserves to be split into two papers so that the message of the two aims do not become diluted.

*We thank the reviewer for his constructive advice. We have removed the prediction score from the current version of the manuscript.*

I fully appreciate that prospectively collecting data to externally validate is not a trivial exercise and will take time. But suggesting that the model should be used is problematic without demonstrating it ‘works’ - particular when the sample size is moderate. Was any attempt made to identify existing datasets that the model could be evaluated on?

Developing a prediction model is a separate study and deserves to be seen as such so that all aspects in deriving and internally validating the model can be explored in more detail than is currently described here (again adhering to the TRIPOD Statement for prediction models; www.bmj.com/content/350/bmj.g7594 and annals.org/article.aspx?articleid=2088542 - which the authors appear to have overlooked despite being previously pointed towards TRIPOD). The actual development of the prediction model looks well done and thus just confirms my opinion that this should be a separate paper.

As an aside presenting ROC curves without labelling particular cut-offs on the curve is uninformative, labelling the curve can then allow sens/spec to be read off at particular cut-offs. The calibration plot is ok, but it is widely recommended to also superimpose a lowess calibration curve in addition to the observed and mean predicted probabilities at fifths of predicted risk. Quintiles is incorrectly used, a quintile is a cut--point to create equal sized groups, the correct term if fifths (See the BMJ Stats notes, www.ncbi.nlm.nih.gov/pubmed/7950724). Hosmer-Lemeshow test is an uninformative measure and does not assess calibration, widely influence by sample size and grouping, but importantly doesn’t assess direction of magnitude of (mis)calibration). (see TRIPOD Explanation & Elaboration paper for more details; annals.org/article.aspx?articleid=2088542).

*All the above comments concern the prediction model, which is now no longer part of this manuscript. We thank the reviewer for his valuable feedback, which we will take into account in the preparation of a separate manuscript describing the prediction score.*

The sample size for the main diagnostic accuracy study has now been reported but included in a supplementary box. Why? This should be in the main body of the paper and not tucked
away in supplementary materials.

*As suggested by the reviewer, we have moved the sample size calculation to the main body of the manuscript.*