Reviewer #1: “Comparing sensitivities at cutoffs corresponding to (very) different specificities is not very meaningful… To avoid these issues, it is an accepted statistical practice, to compare sensitivity between methods when their specificity is kept as similar as possible (and vice versa) if multiple tradeoffs between sensitivity and specificity can be easily generated (as shown in a ROC curve). “

Reviewer #2: “It seem odd to discuss differing sensitivities of different tests that have different specificities. Could the authors choose one specificity and tie all of their tests to that specificity?”

We have made two changes to address these concerns:

1.) We changed the cutoff for the MMT, by choosing the optimal cutoff for detection of cancers rather than all lesions. This has resulted in a stricter cutoff that gives the MMT a higher specificity (90.1% compared to 83.7% before) and lower sensitivities. While still lower than FIT (97.1%), the disparity in specificity between the two tests is now half of what it was.

2.) For the direct comparison of sensitivities (barplot, Figure 4), we have set both tests to the same specificity as the MMT (90.1%) so that they can be compared more fairly. We also changed the statistical test for this figure (as Reviewer #1 suggested) to the test for comparing sensitivities at a given specificity implemented in the pROC R package.

Reviewer #1: “About the OOB vs cross validation estimates of generalization error, the references cited in my understanding do NOT support the notion that the "OOB error estimate has been repeatedly shown to more accurately reflect the generalization error than using cross-validation" (although the OOB gets close to the generalization error). And as an aside: cross validation is so popular because it makes very efficient use of data (with 10-fold cross validation the training set is still 90% of its original size; only at the cost of computational overhead, which is not a serious issue for data sets of the size presented here). The Methods sections should in any case clearly state that OOB error estimates on all 490 samples were used to estimate generalization error.”

Reviewer #2: “Could the authors take one of the other CRC/adenomas 16s datasets (from another lab without FIT) and use that as a training set building a model to classify case from control. Then on their dataset, they could combine the predictions of that model applied to their sequences with the FIT to see if the model still improves on FIT alone. This would test how well another lab's sequencing pipeline on an independent cohort was able to make predictions in their current pipeline. I appreciate that this is likely outside the scope of the paper (and that based on the authors responses to reviewers) an independent validation cohort is already currently being recruited. But this would be a more reassuring way to directly determine that they have not overfit their data…”

Both reviewers have concerns about the models being overfit and whether we can rely on the out-of-bag (OOB) error rates for estimating the random forest models’ accuracy. Unfortunately it is not possible to use the approach suggested by Reviewer #2 because features cannot be retroactively added to a random forest model. In other words it is not possible to train a random forest model using only OTU data from another lab, and then add in FIT and apply it to our cohort. However, to address the reviewers’ concerns about relying on OOB error, we have performed additional leave-one-out cross validation as well as 100 iterations of 10-fold cross validation for each of the random forest models. New Figures S2A, S2B, and S5 show that the ROC curves for the cross-validations closely match the ROC curves generated from OOB estimates. We believe these data support the use of OOB error rates and give greater confidence that the models are not overfit.

The authors state in their response that "the number of samples is larger than the number of features", which is a little bit surprising to me; I would have expected more than 490 OTUs constructed at 97% identity. Perhaps the authors could mention the input size of their OTU feature table, this would also allow the reader to better appreciate the feature reduction in the RF models (for which the number of OTUs is reported).

The Reviewer is correct that the initial size of the OTU feature table was much larger than the number of samples. To clarify, only OTUs present in at least 5% of samples were used for feature selection for the models (335 OTUs). This important detail has been added to the methods section. However the optimal MMT model only uses 23 OTUs, which is much less than the number of samples (490). This is what we were referring to when we said "the number of samples is larger than the number of features" in support of using OOB estimates.

Reviewer #1: The authors may also wish to remove a remnant of the serial combination of MMT and FIT from the Conclusions.

Reviewer #2: The authors have indicated that they accept reviewer #1's argument that serial testing does not make statistical sense ("The reviewer's comments regarding serial testing are well taken."). Yet in the discussion, they still state that "One way to approach screening would be to use FIT and the MMT in series, thereby preserving the advantages of the higher specificity of FIT and the superior sensitivity of MMT.". If the authors accept that serial testing is not advisable (since in the words of reviewer #1, there is no free lunch in statistics), they should remove this section from the discussion.

We have removed the remnants of the serial testing results from the discussion and conclusions sections.

line 229: "increased number of false positives (and subsequent colonoscopies) "; could the authors indicate here how many false positives and additional colonscopies would be undergone if their test was used for the "recommended screening population in the United States"

We have added that the MMT would result in approximately 6 million false positives compared to 1.8 million false positives for FIT. We also point out in the discussion that the increased sensitivity of MMT might allow for less frequent testing, thereby reducing the number of extra colonoscopies per year. The screening frequency suggested for the Cologuard test, which has very similar sensitivity and specificity to our model, is once every 3 years. Using the MMT at the same interval would result in roughly the same number of false positives per year as FIT.