**We would like to thank each of the reviewers for taking the time to read our manuscript and for providing valuable insights and comments that helped us improve the manuscript and better explain our approaches and methods. Below you will find our responses to the Reviewer’s questions, provided in bold:**

Reviewers' comments:

Reviewer #1: An excellent experimental design and well written manuscript. I congratulate authors for the explicit details of each and every step.

I have a few minor observations. I recommend if these are incorporated into the manuscript it will further enhance the clarity.

1. Did the authors considered any cases where there is combination of ICH and IVH? If so what was the outcome of the segmentation?

**Thank you for your enthusiasm about the manuscript. We are currently in the process of applying the model to a set of patients with predominant IVH from the CLEAR studies. We plan to write up these results in a separate manuscript in the future.**

2. Though there is a detailed description of the data used, it is not clear which subset of data were selected for training in terms of scanner, size of ICH, pixel intensity range, etc. Inclusion of this will add value to the manuscript.

**We have clarified by adding information on the randomly selected 10 patients used for training:**

**“These scans were randomly selected. These selected patients had a mean (SD) hemorrhage volume of 37.8 (6) mL, had an average HU intensity of 58.8 HU and were scanned on the following scanners: GE (N = 6), Philips (N = 2), Siemens (N = 1), and Toshiba (N = 1). We used the 102 remaining scans as test data to evaluate the performance of the proposed approaches.”**

3. The sentence in Discussion is incomplete "The approach failed on a very "

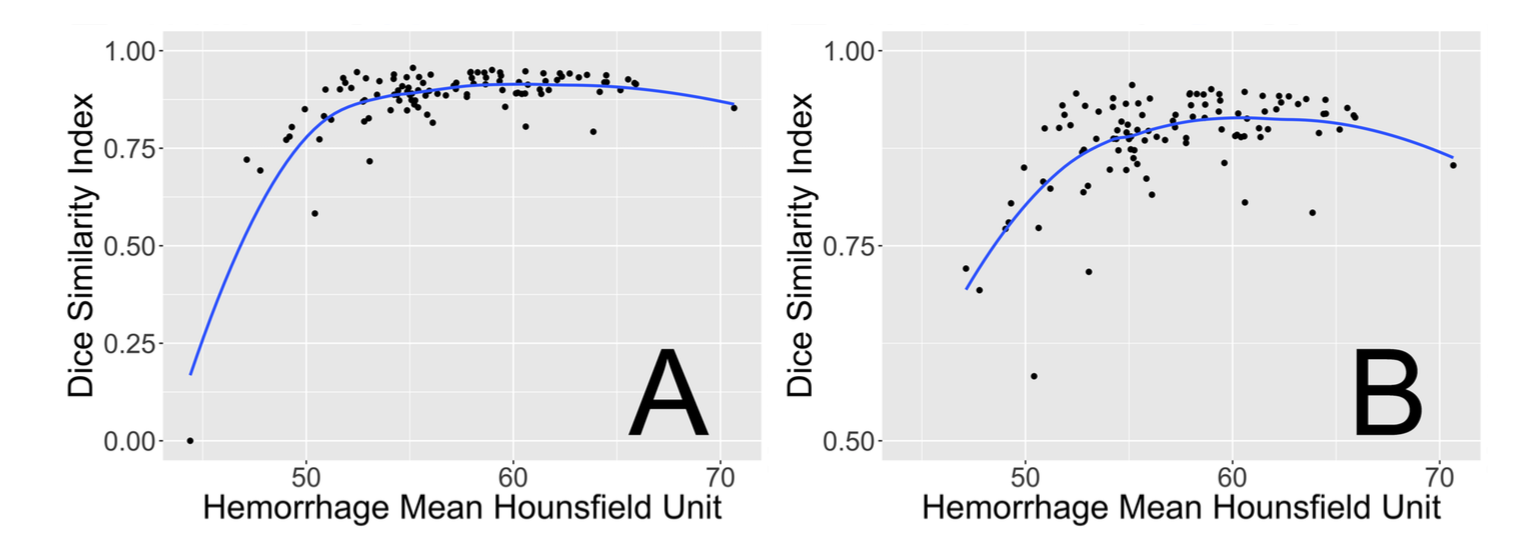
**Thank you for pointing out this error. We have changed the sentence to “The approach failed on a very small subset of the validation set”**

4. In the results section add

a) DSI with respect to HU

b) DSI stats for different volume of ICH (e.g. small, medium and high volume ICH)

**We have added analyses of the DSI compared to mean HU and have found that over an average of 50 HU, there was not a large effect on DSI, but below 50 HU, there was degraded DSI Moreover, the case of failed segmentation had a low average HU (44.4 HU). We have also created plots and summaries for the average DSI for the categories of hemorrhage volume.**

**Specifically we have added the following paragraphs and figure:**

**“In Figure 6, we see the effect of average HU over the hemorrhage versus DSI. We note the failed scan had a much lower average HU in the hemorrhage compared to the rest of the images (panel A). Overall, however, above 50 HU, there does not seem to be a strong effect on the average HU and DSI (panel B).**

**After categorization of the hemorrhage volume, 34 (33.3%) patients had a volume of 0 − 30mL (small), 46 (45.1%) had volumes > 30 to 60mL (medium), and 22 (21.6%) had > 60mL (large). The Kruskal-Wallis test indicated a difference in medians (χ2(2) = 6.5, p = 0.04, see Supplemental Figure S7). After Bonferroni correction, there was no statistically significant difference in median DSI, but the strongest comparison was for the small versus medium hemorrhage sizes (W = 537, p = 0.0502, corrected).”**

5. Add some notes on "Applicability of the proposed methods for SAH, IVH and other types"

**Thank you for this suggestion. We have now added 2 paragraphs about this in the discussion:**

**Although we have shown good performance in patients with ICH, intraventricular hemorrhage (IVH) and subarachnoid hemorrhages (SAH) are two other forms of hemorrhagic stroke that may occur clinically. As obstructive IVH requiring external ventricular drainage was in the exclusion criteria in MISTIE II, no cases with primary intraventricular hemorrhage were included. The 3 cases with the largest overall volume had some considerable IVH extensions, the largest being around 20mL. This leads to an overall underestimation of the hemorrhage volume in our models. We believe the model may perform reasonably well with patients with IVH, yet may require a separate model, but not necessarily SAH.**

**Intraventricular hemorrhages have similar HU ranges for large portions of the hemorrhage, but may have lower HU values of the hemorrhage which are less dense and surrounded by cerebrospinal fluid. Moreover, larger IVHs may cause larger deformations than seen in cases with ICH only, which may negatively affect any registration process. In contrast, for SAH, the process of mask erosion may considerably remove voxels from the hemorrhage, which occurs largely on the cortical surface. This step is important to remove false positives, which may occur from partial volume effects with areas of the skull and voxels towards the cortical surface. If the methods for registration and erosion perform similarly to the cases with ICH, we may also need to re-fit the model using additional cases of IVH and/or SAH.**

6. How did you handle the calcification and its partial volume voxels which is the same region of HU

**We did not perform any steps to directly handle the calcifications or voxels neighboring the calcification. The brain extraction step generally only retains voxels with HU < 100, which removes the calcification voxels, which are typically 100-150 HU. Therefore, after the mask erosion, the voxels neighboring the calcification are removed.**

**If calcifications are not removed, then they will certainly not be in the thresholded range from 40 to 80 HU. These voxels and the neighbors will have much lower percent threshold values than compared to hemorrhages. Also, the neighborhood means and neighborhood standard deviations will be much higher for the partially-volumed voxels compared to hemorrhage values.**

**Moreover, if they have a non-zero probability of ICH from the model, the smoothing operation will dampen the probability and lower the likelihood of those voxels being segmented as hemorrhage.**

7. Fig.5 -- Why there was an underestimation for the ICH of 150cc (manual seg based vol)?

**Thank you for this comment. In response, we have added some explanation for this poor performance in the manuscript. That case had a particularly large extension of IVH, which did not segment well and have added this to the discussion:**

**Intraventricular hemorrhages have similar HU ranges for large portions of the hemorrhage, but may have lower HU values of the hemorrhage which are less dense and surrounded by cerebrospinal fluid. Moreover, cases with larger IVH volumes may cause larger deformations than seen in cases with ICH only, which may negatively affect any registration process. In contrast, for SAH, the process of mask erosion may considerably remove voxels from the hemorrhage, which occurs largely on the cortical surface. This step is important to remove false positives, which may occur from partial volume effects with areas of the skull and voxels towards the cortical surface. If the methods for registration and erosion perform similarly to the cases with ICH, we can perform the same procedure and re-fit the model using additional cases of IVH and/or SAH.**

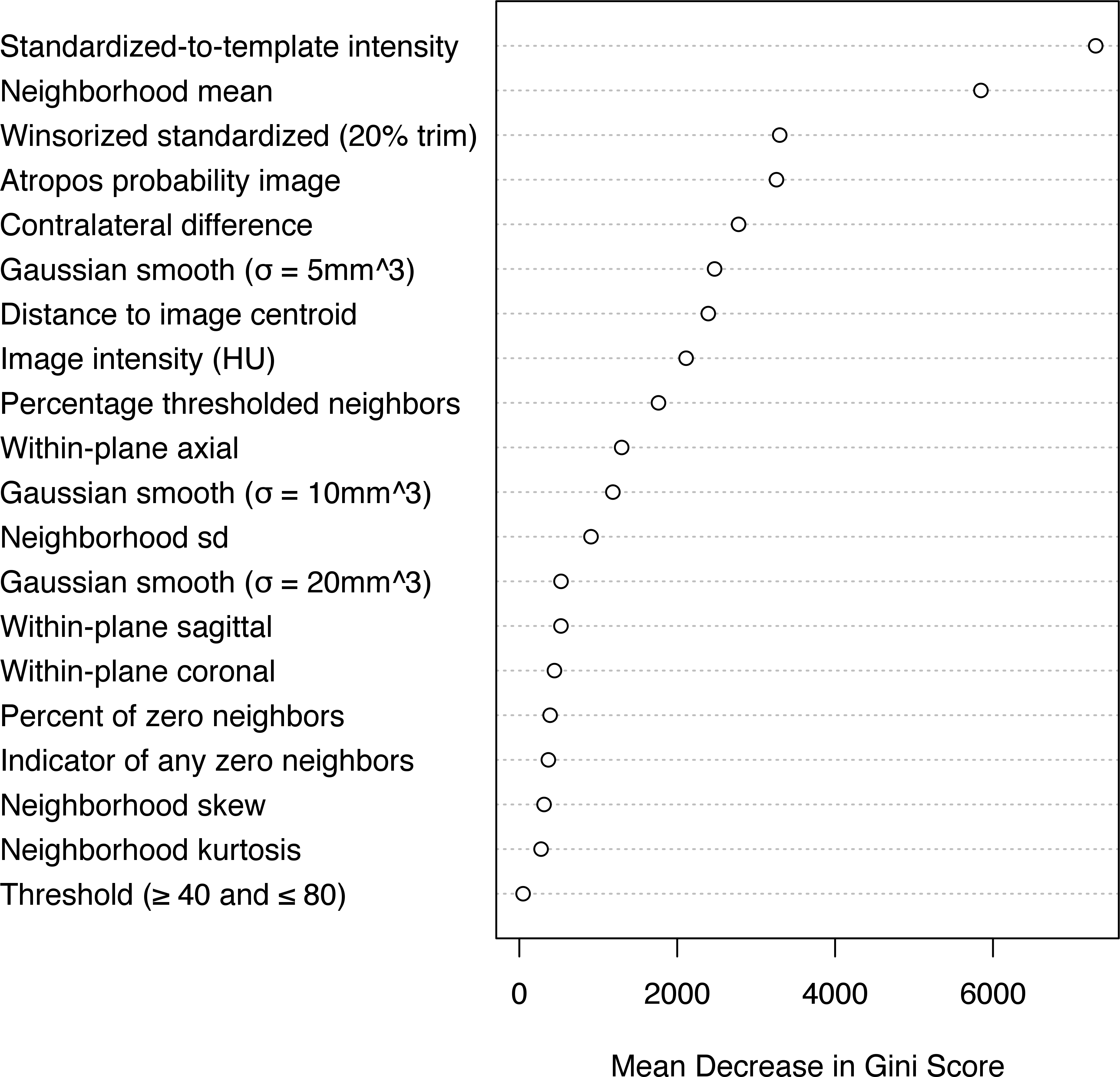
8. Comment on underestimation by automatic algorithm in some of the large volume ICHs

**The underestimation of larger volume hemorrhages happened in patients with had similar extensions of IVH to the case with the largest hemorrhage, discussed above. We discuss that in order to segment other hemorrhages such as IVH or SAH, we need to refit of the model with additional cases from other studies, such as the CLEAR study. We have added the paragraph referenced in #7.**

Reviewer #2: Muschelli et al present a random forest based tool for segmenting intracranial hemorrhage in CT images. This is a meritorious manuscript that should be of interest to the field. Another merit of the work is the provide software, especially the web-based tool.

**Thank you for your enthusiasm about the manuscript and the software. We have address each of your comments in a point-by-point response below:**

I didn't see any discussion on what features were the most important in the main text, only in the SI. I think this should be in the main paper with a little bit of discussion. I actually missed all of this my 1st time through the paper. For example, could some features be dropped? Or were there features that worked better in some cases than others?

**Thank you. We agree that the features are what drive the performance of the model and are therefore one of the most important parts of the manuscript. In response to your comment we have added discussion, a table of coefficients, and a variable importance plot to provide insights into which features drive the performance of our model. **

**We believe some features can be dropped, but as the random forest algorithm is robust in that it does not choose covariates that do not decrease prediction error in each decision tree, we have left them in the current form.**

I'm a little curious why linear interpolation followed by thresholding at 0.5 was used for transforming the mask. Wouldn't nearest neighbor interpolation do the same thing in one step?

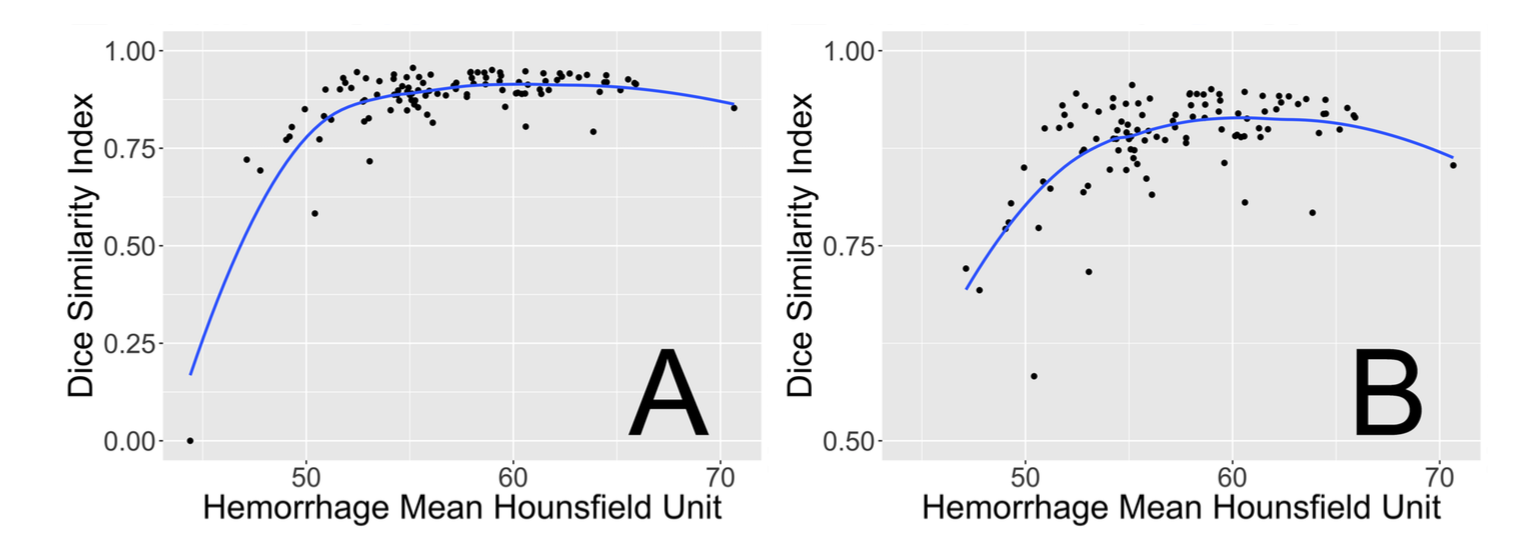
**We agree that these should be relatively equivalent and have added the following sentence:**

**“Using a nearest-neighbor interpolation for these binary images after registration could also be used as a simpler approach and should be relatively equivalent to the re-thresholding.”**

Was there anything interesting about the scan the failed?

**Thank you for this comment. There were many interesting reasons that the scans failed and these failures provide insights into the segmentation algorithm.**

**We have added analyses of the DSI compared to mean HU and have found that over an average of 50 HU, there was not a large effect on DSI, but below 50 HU, there was degraded DSI Moreover, the case of failed segmentation had a low average HU (44.4 HU). We have also created plots and summaries for the average DSI for the categories of hemorrhage volume.**

**Specifically we have added the following paragraphs and figure:**

**“In Figure 6, we see the effect of average HU over the hemorrhage versus DSI. We note the failed scan had a much lower average HU in the hemorrhage compared to the rest of the images (panel A). Overall, however, above 50 HU, there does not seem to be a strong effect on the average HU and DSI (panel B).**

**After categorization of the hemorrhage volume, 34 (33.3%) patients had a volume of 0 − 30mL (small), 46 (45.1%) had volumes > 30 to 60mL (medium), and 22 (21.6%) had > 60mL (large). The Kruskal-Wallis test indicated a difference in medians (χ2(2) = 6.5, p = 0.04, see Supplemental Figure S7). After Bonferroni correction, there was no statistically significant difference in median DSI, but the strongest comparison was for the small versus medium hemorrhage sizes (W = 537, p = 0.0502, corrected).”**

I think the authors could "advertise" their online tool more. Web tools like this are potentially very valuable in clinical setting/hospitals where one may not be able to install outside tools.

**Thank you – we plan on presenting this at the future international stroke conference, where we have presented previous versions of the work, but without the online tool. We have added links to the package and the tool in the abstract text as well.**

Reviewer #3: In this study, the intracranial hemorrhage probability estimation was obtained using four models. Of all the models, random forest classifier provided the best classification as assessed by dice similarity index. Data from multi-center trial was included of which 10 scans were treated as training set and 102 as test set. In general, manual segmentation is considered as gold standard but automated segmentation is preferred due to relatively large data sets involved in clinical trials. This article provides such technique to address the classification with no human intervention.

**Thank you for taking time to review the manuscript. Below we provide a point-by-point response to each of the comments:**

Comments:

1. While converting the images into NIfTI format, thickness was determined from the first converted slice which could vary from scan to scan. No interpolation was sued for variable slice thickness. Also, it is stated that slice thickness was not same throughout the volume. So, how does variable slice thickness been taken care of?

**Thank you for pointing this out. Variable slice thickness has not been taken care of in the sense that in the original image, different voxels within the same scan are of different dimension. Although this is the case, after registration and interpolation to the template, the voxel sizes are isotropic. We have discussed that future versions of dcm2nii will perform interpolation automatically on these images, which was added as a result of our discussions of this exact problem with Dr. Rorden, the creator and maintainer of the dcm2nii software.**

**Registration may be affected by this thickness as well. We did not see an overall difference in the DSI measurements for the scans with variable slice thickness compared to those without: mean (SD) DSI of 0.879 (0.11) without variable slice thickness, and DSI of 0.873 (0.064) with variable slice thickness.**

2. Rigid registration was used to co-register images to the template. Rigid registration may not work for scans that have bigger or smaller brain compared to the template brain. How does brain size affect the registration? Provide discussion on using rigid registration for variable brain sizes.

**Thank you for this comment. As our model is not fit using voxel-level spatial information our registration to the template need only be approximate (and therefore we found that a rigid registration was appropriate). The registration was done so that we could make the contralateral feature and the inter-scan standardization feature, which required the image to be in approximately template space. We have added the following paragraph for clarification:**

**“This rigid registration does not ensure brains are the same size or that voxels match across subjects. Having voxels registered across subjects is not necessary for our model, as we do not incorporate voxel-level spatial information into the model. Although brains with different shapes and sizes may map to different areas in the template space, the goals of this registration are to reorient the image, ensure isotropic voxel sizes for smoothing and other operations described below, and preserve the relative volume of the ICH. All image preprocessing and analysis are done in MNI space, described as template space, unless otherwise specified.”**

3. Was brain mask erosion done before or after the application of rigid registration? It is not clear from processing pipeline Figure 1.However, the steps in section 2 suggests that the masking was done after registration. Please elaborate.

**Thank you for this question. This has been clarified in the text to indicate erosion was performed after registration and we have updated this in the workflow figure. The rationale is that after registration, all voxels from all scans are the same size, and a single erosion procedure can be used, whereas care may need to be taken if eroding the brain mask in native space.**

4. Provide description on selecting HU intensity threshold value between 40 and 80. Include the discussion on the application of same range of threshold on all training scans. All scans may not have same intensity range.

**Although all scans may not have the same intensity range, it is assumed they do by the calibration of the CT scanner and how the Hounsfield unit is calculated. Moreover, our collaborators have indicated that these ranges are appropriate for the majority of the scans they read and segment, and these ranges have been used in previous work (Prakash et al 2012. We have added this citation to these justifications to the section for the threshold predictor.**

5. Within-plane standard scores are used as calculated. But it is not clear whether intra-subject and/or inter-subject standardized images were used to obtain predictors. Please explain with reason of including or excluding intra-subject and inter-subject image standardization.

**Both intra-subject standardized images were used as predictors, such as the within-plane standardized images and the Winsorized/trimmed z-score images. Although CTs are measured in HU, which are standardized, the intensities still have variability scan to scan. These standardizations will normalize the intensities so that shifts in intensities will be offset, such as those potentially due to shifts due to scanner. Inter-subject image standardization using the Standardized-to-template Intensity allows the voxel intensities to be comparable across patients. Gillebert et al. 2014 also have shown that this predictor can detect voxels outside of a standard range such as the hemorrhage.**

6. Section "Initial segmentation" appears to explain that the model applied in the study could be used with Atropos, but it does not provide details on initial segmentation.

**We have added information with regards to the Atropos segmentation.**

**The section now reads:**

**Consider, for example, Atropos, a previously published, open source, general segmentation tool based on Markov random fields for image segmentation. Atropos combines an initial segmentation based on k-means with an expectation-maximization algorithm for a finite mixture model along with a Markov random field prior. We used Atropos to conduct a 4-tissue class segmentation which provides the probability for each class. We combined the top 2 probability classes into one class as Atropos orders the classes by the mean intensity and hemorrhages have higher HU values. This combined probability was then used as a predictor, denoted by Atropos(v). Although Atropos has been shown to perform well in other studies for tissue-class segmentation, the Atropos segmentation did not perform adequately in our ICH CT data. However, using the Atropos segmentation probabilities as predictors can be done seamlessly in our approach. Similarly, the results of any other segmentation approach can be incorporated in our approach and the relative performance of methods can be compared.**

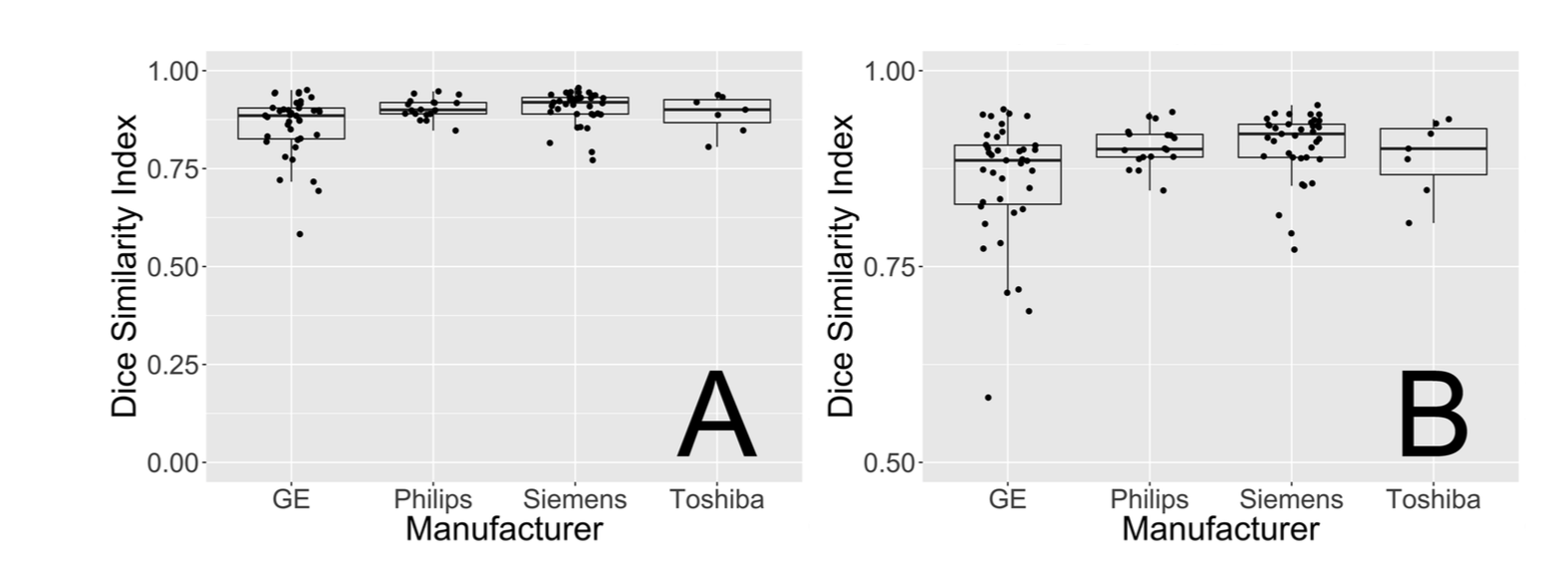
7. Contralateral difference images are obtained with right-left flip. This can be done only if the images are centrally located. Was this done before or after registration? Provide justification.

**Thank you for bringing this to our attention. We have now explained that the contralateral images are created with the registered images. Just as you suggested, it is only appropriate to do this right-left flip in the template space (which is centrally located) as opposed to in the subject’s native space.**

8. Describe the effect of different scanners on the calculation of predictors.

**Thank you for this question. We have provided DSI for each scanner across the population of test sets and a discussion in the following text and supplemental figure:**

**“In the following section, we will use the DSI from the random forest model. In the 102 test scans, the median DSI was 0.89 for patients scanned in a GE scanner, 0.9 for Philips, 0.9 for Toshiba, and 0.92 for Siemens (Supplemental Figure S6). The patient with a failed segmentation was scanned using a GE scanner. The Kruskal-Wallis test indicated a difference in medians (χ2(3) = 11.6, p = 0.009). After Bonferroni correction, the difference of DSI of patients scanned with Siemens versus GE scanners was the only statistically significant comparison (W = 428, p = 0.008, corrected).”**

****

9. Provide justification for using standardized-to-template intensity as predictor. In general, the standardization is applied prior to the calculation of other predictors.

**Although in most other imaging modalities (e.g. MRI) intensity standardization must be performed on the intensities from the scanner before feature creation or aggregation across a population, that is not generally the case with CT scans because of calibration of the CT scanner and how the Hounsfield unit (HU) is calculated. HU standardizes the attenuation of the area compared to the mean attenuation value of water and the mean value of air for that scanner. Thus, units are scanner-standardized. The standardized-to-template intensity can be used to detect voxels outside of a standard range such as the hemorrhage. This predictor attempts to use the intensities of non-stroke patients that takes into account the spatial pattern of the mean and variability of these intensities. The within-scan standardizations can offset shifts in the distribution, but only work on the non-spatial distribution of the intensities and no spatial information is directly in those.**

10. What were the criteria used in selecting 10 scans from 112 scans? Are 10 scans selected from single scanner?

**The scans selected were selected randomly. We have added the following text:**

**“These scans were randomly selected. These selected patients had a mean (SD) hemorrhage volume of 37.8 (6) mL, had an average HU intensity of 58.8 HU and were scanned on the following scanners: GE (N = 6), Philips (N = 2), Siemens (N = 1), and Toshiba (N = 1). We used the 102 remaining scans as test data to evaluate the performance of the proposed approaches.”**