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?

**We are currently applying the model to a set of patients with predominant IVH from the CLEAR studies.**

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 added information as to the randomly selected 10 patients used for training.**

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

**This sentence has changed 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 HU. Moreover, the case of failed segmentation had a low average HU (44.3 HU). We have also created plots and summaries for the average DSI for the categories of hemorrhage volume.**

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

**We have added 2 paragraphs about this in the discussion:**

**As the obstructive intraventricular hemorrhage (IVH) requiring external ventricular drainage was in the exclusion criteria in MISTIE II}, no cases with large intraventricular hemorrhage were included. We believe the model would perform reasonably well for IVH, but not necessarily subarachnoid hemorrhages (SAH), two other forms of hemorrhagic stroke.**

**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. The process of mask erosion may considerably remove voxels from the SAH, 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. 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)?

**We have added some notation on this. That case had a particularly large extension of IVH, which did not sequent well in the method.**

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

**The larger volume hemorrhages had similar extensions. We discuss the potential refitting of the model with IVH and that is a future goal using the CLEAR study.**

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.

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?

**We have added some discussion, a table of coefficients, and a variable importance plot to show this. 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. For simplicity and computational speed, we plan on re-fitting the random forest with only the top 10 predictors and see if performance suffers.**

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?

**We have added sections about DSI and mean HU in the hemorrhage and note that the failed case had a low average of HU. As the neighborhood mean is a strong predictor in the model, this likely caused a failed segmentation.**

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.**

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.

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?

**Variable slice thickness has not been taken care of in the sense that in the original image, different voxels are different sizes in real world dimensions. 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.**

**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.**

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.

**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. Although brains of 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.

**This has been clarified in the text to indicate erosion was performed after registration. 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), but we agree that this range likely does not perform well across all scans.**

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.

**Within-plane scores are done at an intra-scan level, and we have added text to indicate that. We believe the Standardized-to-template Intensity is indeed an inter-subject image standardization. These intensities can be interpreted as the number of voxel-wise standard deviations away from the voxel-wise mean over the population of control patients used to create the voxel-wise mean/sd.**

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.**

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.

**We have noted that these are done on the registered images. The template is centrally located so it is appropriate only after registration.**

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

**We have provided DSI for each scanner across the population of test sets and a discussion.**

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 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.**

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 provided information as to which scanners and ICH were used.**