Reviewer #1: I have enjoyed reading this revised paper "Validated Automatic Brain Extraction of Head CT Images".

I suppose that it is every diligent scientists' worst day when you submit a paper and find out that some substantial work was published very recently that alters some of your claims. In this case, there are a suite of companion papers that have appeared dealing with brain extraction using CT, brain vs CSF segmentation following extraction, validation against hand segmentation, lobar volumes, hydrocephalus, etc:

<http://thejns.org/doi/pdf/10.3171/2014.9.PEDS12426>

<http://thejns.org/doi/pdf/10.3171/2014.9.PEDS12427>

<http://thejns.org/doi/pdf/10.3171/2014.9.PEDS12428>

In terms of the present work, these works affect priority claims. They are not the first papers to do validate brain extraction using CT (Mandell et al 2014a), nor the first to do so using diseased rather than normal brains (Mandell et al 2014b, c).

**We agree, this can be one of the worst days and thank the reviewer for discussing these papers.**

**We agree that that Mandel et al 2014a introduces a brain and fluid extraction (denoted as the Mandell method) and Mandell et al 2014b and c perform brain extraction using scans of brains with disease. The Mandell method works on brain extraction slice-wise and requires seed regions from users. Moreover, the proposed method required calibration on the 5 subjects analyzed. There is no indication if these calibrated parameters perform well in a large (> 1000) scans.**

Second, the present work can be misconstrued, and I fear that even the title is misleading. They do not perform brain extraction. They do skull stripping and calculate intracranial volume. Although they mention brain volumes as being enabled from this work, they skirted the issue of scans being done with multiple slice thicknesses, letting a focus on voxels be sufficient here (I think they take one slice thickness measurement from the DICOM header). This seems to imply that the present approach will not work for brain volume determination if it uses 1 voxel volume for scans with multiple voxel volumes.

**Response – As the NIfTI standard requires one voxel size per dimension, when converting from DICOM to NIfTI, only one voxel size is taken for the z-direction (as these images are acquired axially).**

**This loss of information can affect 3D operations on the image based on real-life millimeters (mm) rather than voxels. For volume estimation, which relies on accurate pixel dimensions in mm, this is crucial. Although the NIfTI image stores the data with only one z pixel dimension, we use the ImagePositionPatient DICOM field to accurately add up each voxel with the correct z pixel dimension. To reiterate, the images are stored with only one voxel dimension, but we correctly calculate intracranial volume with the DICOM header information, which takes into account the variable slice thickness.**

**We now define intracranial volume as the volume inside the skull, including subarachnoid spaces as well as those filled with cerebrospinal fluid to be more explicit.**

More important for 'brain extraction', is that after skull stripping, there are tremendous difficulties that are encountered to suitably separate brain from CSF on CT imagery, none of which is addressed here.

Mandell et al 2014a dealt with smoothing (another priority claim here), as well as a detailed process of erosion and dilation to accurately segment extracted brain from CSF.

**Response – We did not previously discuss CSF extraction in the text, but have added the reference of Volkau et al (2010), which discusses deriving cutoffs to extract CSF within CT images.**

**With respect to dilation and erosion, we use FSL’s hole filling operation to close any holes within the brain mask. Although Mandell et al 2014a used dilation and erosion, it was on the brain mask. We perform smoothing on the original data, which Mandell 2104a does not mention.**

**We have added discussion on how to adapt the method to potentially achieve brain extraction in the subset of scans that failed, including adding steps of 1) neck removal using registration to a CT template, 2) using a higher smoothness constraint in BET, but these steps are not implemented in the original work presented here, and 3) dilation and erosion to fill any holes that are not filled by FSL’s hole filling operation, especially those caused by air after extraction of ICH.**

**Volkau, Ihar, Fiftarina Puspitasari, and Wieslaw L. Nowinski. "Ventricle boundary in CT: partial volume effect and local thresholds." *Journal of Biomedical Imaging* 2010 (2010): 15.**

Also in the present paper, no study of scan noise or gradients were studied. CT can be full of streaks and artifacts that render brain extraction complex.

**Response – We agree with this point. We have added a discussion of the fact that we have not analyzed this. In many cases, if there are large artifacts, a scan will be repeated. This is not relevant if the scan to be analyzed has artifacts, however. Artifacts can definitely cause problems for this method, as artifacts of hyperintensities will be discarded due to thresholding. We believe that smoothing can dampen other effects such as streaks and metal artifacts, but these cases do remain difficult for extracting accurate measurements or structures.**

This is a very software-specific paper. If you are not intimately familiar and experienced with FSL, then this paper is not transparent. Specific software expertise should not be a requirement to read and understand a paper in a journal such as NI. I would recommend that they expand and explain what amounts to 'jargon' in the paper where features specific to FSL are mentioned.

**We have added a paragraph in the methods section describing the fractional intensity parameter, and why we chose the FI measures.**

**The fractional intensity value is used to distinguish between brain and non-brain tissues and determines where the edge of the final segmented brain is located.**

**G\_f, the locally estimated upper threshold of CSF, is a fraction that lies on the way between G\_min and the local maximum intensity B\_max. It is defined as:**

**G\_f = f \* B\_max + (1-f) G\_min,**

**and as G\_min is 0 for areas with extracranial tissue, then f \* B\_Max is the threshold. Therefore, using 0.1 \* B\_max (approximately 40-100) gives a lower threshold of 4, but 0.01 \* B\_max gives a threshold of 0.4, which is below 1 and as CT are integers, there should be not much different on the surface of the brain where there is extra cranial tissue for FI < 0.01.**

The gold standard for segmentation is not manual segmentation but ground truth scans. Although there are marvelous ground truth scans for MRI, we do not have them for CT. So we are able to show consistency with manual segmentation, but that is still not validating our CT volumes. Validation for CT remains, unfortunately, at a lesser standard than for MRI where true ground truth semi-synthetic images are available. One can do a clinical trial with volunteers who would accept both MRI and CT images, validate a tool on MR, and then compare with CT derived brain extractions. This study has, to my knowledge, not been completed (although I am aware of efforts in this direction).

**Response – We agree with your sentiments exactly about the lack of ground truth for CT scans. Although phantoms exist, these do not reflect the variability or shape of human brains. Moreover, large publicly available databases of MRI scans are available whereas open CT images are sparse if available at all. We hope to provide an online module to test CT brain extraction on a user’s data.**

Extracting brain from skull becomes problematic in infants where brain apposes dura and scalp directly. This is a bit far removed from adult stroke patient data, but it becomes of critical importance in an ongoing NIH RCT (ClinicalTrials.gov registration number NCT01936272) where CT is being used to extract brain brain volumes in young infants. Will the present methods only work on older children and adults?

**Gousias et al. (2008) analyzed 33 2-year-old children who had been born prematurely and found that after preprocessing of co-registration and neck removal, and dilation, BET was adequate for brain extraction, and then brain labeling. Moreover, Feng et al. (2012) demonstrated BET performed well in automated brain extraction for pediatric MR images of neonates < 2 months (N = 90), 1-2 year infants (N = 141) and 5-18 year old children (N = 60) with a Jaccard index above 0.9 for all groups. Although CT scans may have differences, we believe this method should be robust at least to children, but would like to validate our method on infants and neonates.**

**Shi, Feng, et al. "LABEL: Pediatric brain extraction using learning-based meta-algorithm." *Neuroimage* 62.3 (2012): 1975-1986.**

**Gousias, Ioannis S., et al. "Automatic segmentation of brain MRIs of 2-year-olds into 83 regions of interest." *Neuroimage* 40.2 (2008): 672-684.**

I hope that the above comments are helpful.

Reviewer #2: The authors have adequately addressed all the points I raised in the initial review of the paper.

**Thank you for again reviewing the paper and for the helpful comments.**

Reviewer #3: The manuscript, entitled "Validated Automatic Brain Extraction of Head CT Image", presents a validation of the FSL BET tool for CT images. This work is the first to validate this tool for CT and optimizes a method parameter and evaluates smoothing for this purpose. The validation is performed using two datasets: 1) 36 scans with manual segmentations, 2) 1095 longitudinal scans from 129 patients.

The paper was reviewed previously by two reviewers and has accordingly been improved by the authors. In comparison to the previous version, the authors added an analysis of longitudinal data without ground truth to estimate the failure rate and intra class correlation of the method. I think this is a strong improvement of the work.

I agree with the authors that proper validation of a tool for a new application can be novel research. I think the topic of accurate CT brain segmentation tool is important and valuable, and used the data sets suit the purpose well. However, in my opinion at least two major points in this paper should be improved to make it a proper validation paper that can possibly have a great impact on the field.

1)      I suggest that the authors extend the discussion of the validation study they perform. Currently a proper discussion is lacking in my opinion. It would be very useful to discuss: a) the strengths and limitations of the performed validation study, b) other literature on brain segmentation in CT, and c) other points mentioned in the minor comments below.  In addition, it would be good to end the discussion/conclusion with a recommendation for future research. What would the authors recommend researchers to use for analyzing CT brain data?

**We have added a discussion of the literature, the limitations of our study due to inclusion criteria, and added the minor comments from below, and future research.**

2)      Because it is important for the validation to know the quality of the ground truth, I would recommend the authors to provide some insight in the performance of the manual segmentation if that is possible. I would suggest reporting inter-observer variability on the performance measures in Figure 2. I understand that this would require another observer to manually segment 36 CT brain scans and that this is time-consuming, but this would give the article a lot more value.

**Response – we have added the performance measures of Figure 2 for rater 2, including the union and intersection of rater 1 and rater 2’s segmentation (Supplemental Figure 1). We have also added the measure of inter-rater reliability, represented by the correlation of the ICV estimate, and Bland-Altman plot of the ICV, of rater 1 compared to rater 2 (Supplemental Figure 2).**

Minor comments:

1)      Given the CT-images in Figure 3, undersegmentation is the problem for brain extraction in CT. It seems to me that oversegmentation would not happen because of the high bone intensities. Could the authors please comment on this? Would a simpler method using thresholding only not be sufficient?

**Undersegmentation is commonly the problem with this method, as oversegmentation is limited by the initial thresholding of 0-100 Hounsfield units (HU). We do think this thresholding gives a sample way of doing a rough brain extraction, but extracranial tissue resides in the same HU range as brain tissue. Therefore, we believe one must do some combination of thresholding and region growing. BET performs additional operations to this, but we believe simple thresholding is a good first step, but not sufficient.**

2)      A short description of the BET methodology should be given in this paper. What does fractional intensity mean and why is it important to optimize this parameter?

**This has been added to the methods.**

3)      The FI parameter for BET was validation and three values (0.01, 0.1, 0.35) were evaluated. Fig. 3 shows that without smoothing FI=0.01 gives the best resulting segmentation. Why did the authors choose these evaluated values for FI? Why was FI not decreased below 0.01, as Fig. 3 indicates that this improves performance?

**The measures given by 0.01 had high values in all performance measures. We could choose smaller values for FI, but we believe this may result in more unstable results and that these results were sufficient for our analyses and visual inspection. We also present a case where lack of smoothing does not perform well, even with an FI of 0.01.**

4)      Could the author please provide some intuition on why the smoothing works that well?

**We believe that the thresholding gives a good rough segmentation, but images may have different levels of noise, especially those due to artifacts. We believe this smoothing reduces these artifacts and reduces the noise in these images. Furthermore, areas close to the outside of the body, such as the extracranial tissue, have many non-thresholded values (i.e. zeros), which downweight their Hounsfield units to those lower than in the brain tissue, which are surrounded by other tissues. BET does not remove those on the surface of the brain largely due to the connectedness to other brain voxels as well as these values not getting reduced to below the 2% lower threshold BET employs.**

5)      The abstract can be more concise, e.g. 1) The sentence 'The intracranial volume…for that scan' can be removed as this basic procedure does not need to be explained in the abstract, 2) 36 images form 36 different patients -> 36 scans.

**We agree with the reviewer. This style was used because previously we had multiple scans from the same patients at different time. This verbage “36 images form 36 different patients” was in part to make it explicit to the reviewers that we had removed multiple scans per person. We can remove this if as it should be clear that there is one scan per patient.**

6)      Introduction: "This step is necessary because CT images often contain non-brain human tissues (e.g. skull, eyes, skin)…" Is this not always the case when scanning humans?

**We wanted to make it more explicit that CT images contain non-human elements such as pillow, medical devices. We agree that non-brain tissues are always captured, and have changed the statement.**

7)      The authors mention the papers by Solomon et al. and Rorden et al. in the introduction. Please extent the discussion of this paper by mentioning briefly what those studies did and pointing out the similarities and differences with the current manuscript.

**This section has been added to the Introduction.**

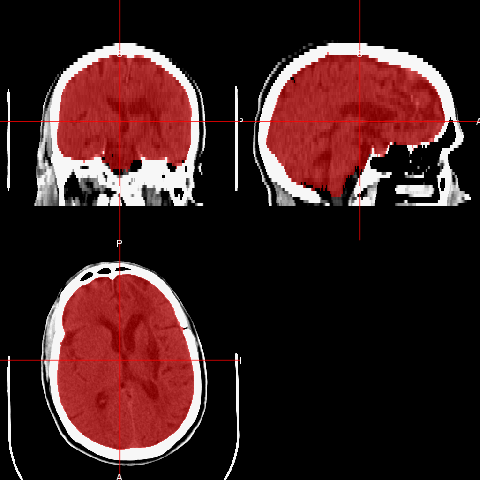
8)      Methods (2.1): I recommend mentioning the most important inclusion criteria as well as patient demographics (age, gender) here as well, instead of only referring to another paper.

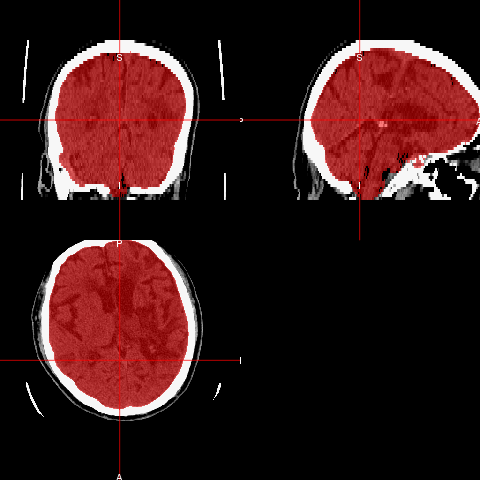
**We agree that this information is important for generalizability and have added these to the Methods section.**

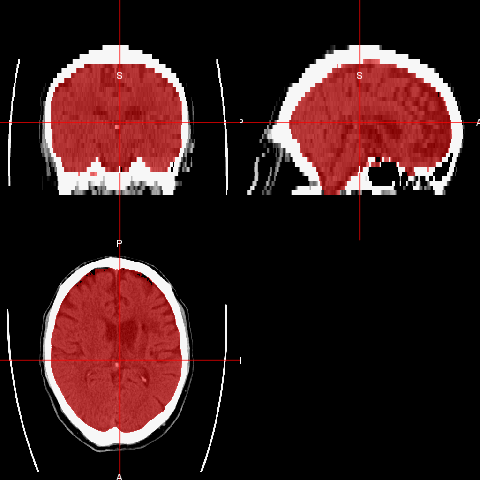
9)      How generalizable would this method be as the analysis is performed on subjects with hemorrhage only?

**We believe this method is applicable to subject without hemorrhage. Unlike images acquired with MRI, no large repositories of CT images are publicly available. This makes validation of methods for a large set of setting difficult. We believe that hemorrhages would likely cause poorer performance due to the fact they change the quantiles of the distribution of the image, which BET uses in segmentation. Moreover, we have at follow-up time points where no hemorrhage is present, which are included in the ICC analysis, where skull stripping works well.**

**We discuss that BET performs well in follow-up scans. Although there may be tissue damage where the hemorrhage occurred, and these patients still only generalize to our inclusion criteria, we believe this demonstrates that our method can apply to healthy adult brains as well.**

****

****

****