We would like to thank the Editor, Associate Editor and reviewers for their comments and recommendations on previous manuscript submission NIMG-14-2266. In particular, we would like to thank the editor for correctly noting that, indeed, this is the first validation paper for CT skull stripping. While Smith (2002) had done validation of BET skull stripping for MRI images, there is currently no validation for CT skull stripping. Our paper addresses this important gap in the literature. In addition to addressing the comments of the reviewers we have further strengthened the paper by:

1) Increasing the number of validation CT images from 20 to 36, removing multiple scans per patient. This is closer to the 45 validation images used by Smith (2002) for MRI. All estimated performance measures remain high.

2) Deploying the skull stripping procedure on a longitudinal study of CT brain images and studying the scan-rescan reliability for intracranial volume estimation. We used 1095 scans on 129 people for a mean (SD) of 8.5 (2.8) longitudinal repetitions per subject.

3) We have estimated the failure rates for each processing pipeline in this large collection of scans.

Please see the attached reviewer comments and questions (bolded) and our response to these concerns.

Reviewers' comments:

**Reviewer #1: In this paper, the authors present validation of brain extraction from CT images using the brain extraction tool (BET) of FSL. With their optimized parameters (FI and smoothing), good agreement between manual extraction and the automated technique is achieved based on the metrics of accuracy, sensitivity, specificity and DSC.**

The authors should address the following points:

**1) Why was the gantry tilt correction necessary? Does this not result in degraded slices given the much higher inter-slice spacing compared with the in-slice voxel size?**

- Gantry tilt was not necessary and we now explain this in the paper. However, we have added the gantry tilt correction for our general pipeline, which involves rigid body registration of the skull-stripped images. This step requires gantry tilt correction, though slice information is not degraded, as the correction is done one a slice-by-slice basis and no interpolation is applied as the image is only shifted with zero padding to offset the tilt. This was especially simple because all scans were acquired axially. As registration is a common next step, we believe estimating the performance on corrected images is useful.

**2) It is indicated that the slice thickness can vary across the volume. However, in section 2.3, it is indicated that the thickness is taken from the first slice and then assumed to be homogenous throughout the image (I presume image volume). What was the slice thickness used for? What effect did the assumption of homogenous slice thickness have in studies, which had varying slice thickness throughout the study? Was a homogenous slice thickness assumed for the tilt correction, which could result in distortion?**

- As the NIfTI standard allows for only one voxel size for each dimension, when converting from DICOM to NIfTI, one voxel dimension must be chosen for the "height", as these are axially acquired, of each voxel. If one is calculating the volume of a mask by multiplying the number of voxels by the voxel dimensions, then the measurement will be incorrect if the voxels are not the same size. Therefore, when images have variable slice thickness, to calculate any volume within the image, one must multiply the voxels by the correct dimensions given by the DICOM header information.

The gantry tilt correction operates on the DICOM images and calculates distance using the ImagePositionPatient attribute; so variable slice thicknesses are taken into account.

**3) Was the Gaussian filter applied in 2D i.e. a slice at a time or in 3D?**

We apologize for this oversight in the original version of the paper. The Gaussian filter was applied in 3D. We have now added this clarification to the text.

**4) Given the very small differences between the FI 0.01 and 0.1 and the spread shown in figure 2, it is a bit surprising that the differences reached the high level of significance shown. But, regardless, as stated in the paper, the differences are inconsequential.**

When we investigated FI we were surprised by the large differences in the quality of skull stripping between the previously FI values (0.35) and the FI value of 0.1. We have also considered FI=0.01 because we wanted to investigate whether the improvement continues for even smaller values of FI.

Though the spread of the data given in Figure 2 appears small, the Wilcoxon signed-rank test is done on the differences between the measures at an image-level. For example, using smoothed data, if an image has a Dice of 0.95 using FI=0.01 and 0.94 using FI=0.1, then the difference would be 0.01. The signed rank test is done on these differences, and the majority of them are positive, therefore given a statistically significant result. We have concluded, as correctly pointed by the reviewer, though these differences are statistically significant, there are no practical differences below FI<0.1 after smoothing. We consider that this is both useful and under appreciated in the literature.

**5) Some comment is perhaps warranted why in these studies the FI is quite small. A small FI gives little weight to local maxima, with the threshold dominated by the threshold determined from the global intensity histogram, based on details given in reference [5]. This seems a bit counter intuitive so some discussion on this is warranted.**

FI is a tuning parameter for BET and we have investigated empirically the choice of this tuning parameter based on objective measures. We have found that these small values of FI provide uniformly good results across all of our images. At the very least, this indicates that other studies should carefully consider smoothing and the choice of FI as tunable parameters. Our intuition is that the t2 estimate is 0, and then the t\_l estimate becomes:

t\_l = (I\_max - t\_2) \* FI + t\_2 = I\_max \* FI

which, if I\_max is approximately 100 (since we thresholded) and using an FI of 0.01, then t\_l is 1. Thus, using a higher FI (say of 0.35), this t\_l would be around 35, and this threshold would remove much of the brain area. Also, Smith (2002) notes with respect to the FI parameter: "for an MR sequence that requires changing [FI], one value normally works for all other images taken with the same sequence". We believe that, as CT units are relatively standardized, this applies to CT data.

**Reviewer #2: The manuscript by Muschelli et al., titled "Validated Automatic Brain Extraction of Head CT Images", presents an evaluation of the FSL Brain Extraction Tool (BET) on the clinical population. It is a concise and well-written manuscript; however, I do not recommend it for publication in NeuroImage because it lacks novelty.**

**The authors state that the novelty of the manuscript is in the comparison of automatic segmentation (carried out by BET) to manual segmentation (carried out by "one expert reader", according to the Methods, and by several "expert CT readers", according to the Abstract). It is true that neither the Rorden et al. (2012) or the Solomon et al. (2007) paper carry out such a validation; however, it has been carried out in the original BET paper by Smith (2002), albeit on healthy subjects.**

Smith (2002) carried out the validation on MRI scans, not CT scans, the modality for which we would like to use BET. We find it very encouraging that the editor, Smith, who wrote the original paper, agrees with the fact that this is a novel contribution. To further address your concerns, which we agree with, we have increased the number of manual segmentations from 20 to 36 and have added a large scan-rescan reliability study for the estimated brain volume after skull stripping.

Moreover, as the reviewer notes, Smith (2002) carried out the validation on healthy subjects, whereas we present brain extraction in a set of diseased patients with significant levels of mass effect.

**The result of the validation presented in the manuscript is that BET works very well on the patients used in the study. This seems somewhat underwhelming to me, because BET is an immensely popular tool that has been heavily used for more than 10 years. It is comforting that the authors did not find any significant problems with it, but, in my opinion, this result is not novel enough to merit a publication, because the authors do not provide any novel method or methodological improvement over the existing BET algorithm.**

We wish to only provide a framework for adapting BET, which has overwhelmingly been used for MRI, for CT images. While we agree that we have added nothing to the original BET algorithm, we consider that adapting it to CT images is extremely important. This is a paper designed to be useful and practical for a wide spectrum of clinical researchers. Thus, we do not propose to "re-invent the wheel", though we proposed a specialized "set of new wheels" for CT research. The editor agreed that our proposal is, indeed, novel. We have also found that: 1) smoothing the images prior to using BET improves the performance of the approach for CT; 2) the value of the FI can dramatically affect the quality of skull stripping; 3) much smaller values of FI than currently recommended in the literature perform better on our CT data; and 4) these brain extraction parameters do well in a population of diseased patients with mass effect.

**A novel finding stated in the manuscript is the result that smoothing the data prior to BET is better than no smoothing. While novel, this is does not seem very relevant to the field, because Rorden et al. (2012) have already used smoothing in their method.**

Rorden et al. (2012) does not indicate that they run BET on smoothed images. From Rorden et al. (2012): "To emulate the proprietary ABLe method, we used BET (Smith, 2002) and AIR 3.08 (Woods et al., 1998) to replicate the ABLe technique (Solomon et al., 2007). We started by setting all voxels with a Hounsfield brightness greater than 100 (centers of bone and choroid plexus) to match the dark intensity of air. Next, we manually adjusted the intensity of each image to provide a clear 8-bit grayscale image of the soft-tissue and then scalp-stripped the images using BET with a fractional intensity threshold of 0.35." Thus, although Rorden et al (2012) smooth the CT images for CT registration to a template, they did not indicate any smoothing prior to applying BET.

**Muschelli et al., 2014 vary the BET threshold of fractional integrity (0.01, 0.1, and 0.35), and suggest that the 0.01 and 0.1 thresholds provide good brain extraction; however, they show that the threshold used by Rorden et al. (2012), which is 0.35, works just as well on the smoothed data.**

We agree that 0.35 works just as well on the smoothed data for those presented in the validation, but we see that the failure rate is much higher in the new replication/reliability analysis included into the current paper.

**By the way, the authors state that the 0.35 threshold was "recommended in Rorden et al., 2012"; this is not entirely correct, because the Rorden et al. paper does not explicitly advocate this value, it is simply the only threshold value they report using.**

We agree and the wording has been changed in the text from “recommended” to "used”.

**The sample used in the study is, in my opinion, too small to provide sufficient power. There are 19 scans, but only 16 patients; were some patients scanned twice, or thrice?**

- These duplicate scans were removed and additional, independent, patients were added. We now have 36 scans with manual segmentations, one per subject.

**Were these patients analyzed as independent data points, which is a bad methodological choice?**

- The multiple scans per patient were removed.

**Also, the scans were obtained using 3 different scanners (Siemens, GE and Philips); the authors do not look into the possible impact of the scanner, presumably because the sample size is so small that the difference across scanners cannot be studied adequately.**

- As Hounsfield units are assumed standardized, we did not expect large between-scanner variability. Indeed, in the larger studies in the current paper we did not find a large scanner effect. In the current paper we present a large (1095 scans: 129 patients from 26 sites, with a mean (SD) of 8.5 (2.8) scans per patient) longitudinal study of CT scans and investigate the within-person reliability of the extracted brain volumes. For each scanner, we present the failure rates. Although we find that the Neurologica scanner had a failure rate of 16.7%, only 30 scans were acquired using that scanner. There was no significant difference in failure rates across scanners (p = 0.10). After taking out the scans from Picker scanners, as there were only 2 scans, there still was no difference (p = 0.074).

**In addition, the p-values from the Wilcoxon signed-rank test are not corrected for multiple comparisons.**

- We agree that multiple comparisons were not taken into account, but we believe the estimates of performance are high and therefore statistical significance is secondary to the estimated effects of performance. We now explain in the text that a simple Bonferonni correction would not change the original message.