We would like to thank both reviewers for their careful review of our manuscript, and reviewer #1 for the detailed and insightful comments. What follows is a point-by-point response to the reviewer's comments. The reviewers' comments are in bolded text, our responses to the comments in regular weight text. To aid in assessing our modifications we have attached a marked up version of our manuscript which highlights every change made.

Reviewer #1

The manuscript by Pipitone describes a technique for automated segmentation of the hippocampus and hippocampal subfields. I reviewed this manuscript previously and this version is improved. However, problems persist, particularly in the abstract, that require attention. The authors insist on using the term accuracy (p. 1) but then within the paper, state that there is not agreement for even manual segmentation of the entire hippocampus (let alone the subfields). Thus, it seems incorrect to state that their method is accurate when the "gold-stand" is not agreed on. While the authors touch on this in some form in the manuscript, I think they need to take out instances of accuracy in the abstract (and elsewhere) and instead state that the automated segmentation generally matches quite well what would be obtained in the same subject using manual tracing. Closely corresponding, consistent, even reliable, seem like much better choices here.

We have updated all uses of the word "accuracy" or related terms. We have used the terms that the reviewer deems to be more appropriate throughout the manuscript.

Along these lines, the manuscript continues to suffer from issues of over use of jargon, making it hard to follow. When the authors employ precision here, it wasn't clear to me that they mean the same numbers came out during repeated testing of the same algorithm? This should probably be made clearer.

In the revised manuscript, line 512-514 now reads:

> *Instead, we must evaluate how well the lower-resolution MAGeT-Brain hippocampal subfield segmentations correspond in form to the segmentation protocol used in the high-resolution images.*

In the discussion, on lines 567-569, we have replaced the use of the word "precision" with "agreement" and specified that we see an improvement "over repeated randomized trials":

> *… we have found that generating a template library reduces the variability in segmentation agreement (i.e. MAGeT-brain more consistently produces segmentations in greater agreement with manual segmentations than does basic-multi-atlas method, over repeated randomized trials).*

Another example of jargon is the sentence: "meaningfully bootstrap a template library..." This sentence is opaque to me.

To avoid confusion, where applicable we have reworded sentences that discuss “bootstrapping” a template library and instead describe the process as “generating” a template library, as was done in our earlier paper on MAGeT-brain (Chakravarty, 2013).

Specifically with respect to the quoted sentence, we have reworded this as (lines 559-561):

> *The core claim the MAGeT-Brain method is based on – that a useful template library can be generated from a small set of labelled atlas images – is validated in…*

Even the use of bootstrap in the title is a little confusing, I suggest a title that can more clearly convey the content to a wider audience.

We have removed the use of “bootstrap” from the title. It now reads:

> *Multi-atlas Segmentation of the Whole Hippocampus and Subfields Using Multiple Automatically Generated Templates*

Finally, it wasn't clear to me that FIRST provided "radically" different definitions of the hippocampus from the plots showed here. I think there is tendency in the manuscript to overstate MAGeTs accomplishments relative to other methods. The plots support that MAGeT brain is doing better overall, but I don't see the whopping advantage to justify these kinds of statements.

In the revised manuscript, lines 654-657, we have been more specific (and less editorial) in comparing the labels from each method:

> *With the exception of FSL FIRST all methods correlate well with the semi-automated SNT volumes provided in the ADNI database. However, the FreeSurfer and FSL FIRST hippocampal segmentations are on average about twice the volume of those from all other methods.*

Regarding the inclusion of hippocampal subfields, given the resampling necessary to do this, I wasn't really convinced that much was gained here other than that MAGeT could do this, given fairly noisy input. I suggest reframing this part, if the authors still feel strongly about keeping this section in, to talk about the subfield segmentations as a proof of concept.

We do feel the subfield experiment is necessary as it is both novel and highlights the relevance of our method in situations where manual labels are extremely expensive (time/effort-wise). Therefore, throughout the paper we have qualified this experiment as a “proof-of-concept”.

Overall, I think the manuscript needs to be more evenly toned with its conclusions and how it compares MAGeT to other work, and greater consideration is still required with how they treat subfield segmentation.

We hope this concern is addressed by the modifications made above, e.g. by reframing the subfield results as proof-of-concept, simplifying language and editorializing with more concrete description.