We thank the reviewers for their careful review of our manuscript and their detailed and insightful comments. We have addressed their concerns below and believe that we have greatly improved the scientific quality of the our manuscript as a result.

Before addressing each reviewers' comments individually, in summary we made the following major revisions to the paper:

1. We re-ran the comprehensive cross-validation experiment (Experiment 1) using fully manually segmented whole hippocampus labels (following the Pruessner protocol) for training and validation on ADNI subjects. The original analysis using the SNT semi-automated labels has been moved to Supplementary Materials. This should address the concerns around the validity of using the semi-automated SNT labels as "ground truth" in this experiment.
2. We ran a similar cross-validation experiment on the First Episode Psychosis dataset to ensure that the parameter settings found in Experiment 1 also apply to younger patients and subjects.
3. We add hippocampal volumes of the First Episode psychosis dataset as computed by FSL FIRST and FreeSurfer to serve as comparisons for MAGeT-brain volumes when assessing agreement and biases.
4. Simplified the subfield cross validation experiment so as to maximise performance and make our results more easily comparable to Yushkevich et al. 2010. These experiments now include Dice Similarity Coefficients for the automated segmentations and bias estimates with respect to a ground truth.
5. We carefully rewrote and reworded parts of the paper to make it clear that the experiments (in particular, the subfield segmentation experiment) evaluate precision/reliability of segmentation with respect to an atlas segmentation, not accuracy with respect to ground truth (like a histological ground truth).
6. Rather than default to the standard organization of many manuscripts (Introduction, Materials and Methods, Results, Discussion), given the large amount of data that is now included, we have chosen the following reorganization to improve readability:

    - Introduction
    - Experiment 1 - 10-fold cross-validation using manual segmentations of ADNI data
    - Experiment 2 - 5-fold cross-validation using manual segmentations of first episode psychosis subjects, and comparison to other standard techniques
    - Experiment 3 - Comparison of our method against other standard methods
    - Experiment 4 - Validation of subfield segmentations
    - Discussion
    - Supplementary Materials
        i. Experiment 5 - 10-fold cross-validation using SNT segmentations of ADNI

    Note that the Sections for Experiments 1-4 include a targeted Materials and Methods and Results section so as to be self-contained. Many of the experiments proposed depend on the results of Experiments 1 and 2, and so this organization should improve the logical

> flow of the manuscript. The original 10-fold cross validation using the semi-automatically generated SNT labels has now been labelled Experiment 5 and is moved to the Supplementary Materials. We have chosen to include this experiment so that others who have used these labels for evaluation (of which there are several) can directly compare their results to ours.

What follows is a point-by-point response to the reviewers' comments. The reviewers' comments are in bolded text, our responses to the comments in regular weight text, and additions to the manuscript indented and italicized.

Reviewer #1

The manuscript by Pipitone et al. presents a novel method for performing automatic segmentation of the hippocampus and hippocampal subfields. Using relatively few traced hippocampi, the authors derive fairly good reliability ratings for hippocampus and subfields compared to manually traced values. The authors also show (with the exception of one subfield) that automated segmentations exceed what one might expect from shuffling of voxels (an estimate, to my read, of chance performance).

We thank the reviewer for the thorough and complimentary appraisal of our work.

Overall, the method presented is important as it will likely save time compared to other methods with overall segmentation of the hippocampus. I was less convinced of its utility for hippocampal subfields though and to me this part of the manuscript requires some rethinking and revision. The authors may even consider taking out this part of the manuscript all together because to my read, at least, this part seemed to stand on somewhat shakier ground. Thus, the majority of my comments below are focused on segmentation of subfields with the exception of some of the minor comments toward the end. One last consideration is that the authors need to include statistical results within the Results rather than referring to figures in which statistical differences are visually evident but not numerically provided.

We have attempted to address these comments based on the reviewer's criticisms in the newly revised version of the manuscript. More detailed responses are provided below.

MAJOR

1. In many places, I noted some confusion between the ideas of validity and actual ground truth hippocampal subfield neuroanatomy. The authors have shown convincingly that the reliability values are good in the hippocampal subfields, with the exception of the CA4/DG subfield. Because the tracing method used though is based on the Winterburn et al., a method which, unlike the Yuskevich method, is not based directly on actual histological data, it seems difficult to make strong statements about

"ground" truth.

We agree, and also feel this confusion between the ideas of validity and ground truth was present in the other experiments on whole hippocampal volumetry. To address this confusion we have thoroughly changed language throughout the paper so as to make clear that in the experimental evaluations we are not assessing ground truth validity but are instead assessing the reliability/precision of our segmentations with respect to the manual segmentation protocol used to segment the atlases and target images (see Minor issues below for specific instances in the paper). We have completely removed the use of the phrase “ground truth” (since, as we in the paper, and reviewer #1 points out, ground truth in hippocampal segmentation is ill defined even in histological studies).

With respect to the question of the basis for the tracing method used in Winterburn et al. we note that while the contrast differences were used to guide the manual labels in Winterburn atlases that these manual definitions are, in fact, completely informed by the Durvenoy (2005) and the Mai et al. (2008) histological atlases. Yushkevich et al. (2009) conduct a study of post-mortem acquisitions and base their manual segmentation protocol on the Duvernoy (2005) and the Fatterpekar (2002) MRI/histological atlases. The segmentation protocol used in Yushkevich et al. (2010) is based on an 4T acquisition MRI-protocol (Mueller and Wiener, 2009). Thus, the Winterburn atlases have as much grounding in histology as either Yushkevich study. We’ve added a more detailed summary of these works to the introduction in order to put our use of the Winterburn atlases in context (see Page 4, lines 122-135).

It is also worthwhile noting that since the initial submission of our manuscript, there have been recent advances from the Yushkevich group that revise their initial labelling based on serial histological acquisitions (Adler et al., 2014). While this refined reconstruction and labelling is impressive, to the best of our knowledge this work has yet to be integrated and validated in the context of a large scale fully automated pipeline. We have also added this note to the Discussion.

Also, in order to improve the rigour of our subfield segmentation validation, we have now completely revised the evaluation section for Experiment 3 (the new section for the evaluation of subfield segmentation). In this current experiment we perform a multi-fold leave-one-cross validation in order to getter a better sense of the reliability of the protocol. This now gives us 20 segmentations with which to evaluate our segmentations. The resulting Kappa distributions are now presented in Fig 5, Bland Altman plots of our segmentations are presented in Fig 6, and qualitative visualization of the subfields are given in Fig 7. We believe that this approaches the evaluation stringency that we have put in place for the other experiments (see below).

Simply put, there is currently no consensus amongst neuroanatomists exactly how subfields such as CA4/DG differ from CA3/2, especially on lower resolution MRIs and in the hippocampal head, differentiate. In fact, many methods lump these subfields together, esp. at 3T. Even histological reconstructions depend primarily on one or two

(postmortem) brains and there has yet to be any evidence that individual variability in these fairly small subfields can actually accurately be identified on relatively low-resolution (.3 mm +) scans (see my later comments on this as well). Thus, it is difficult not to attribute the low reliability of CA4/DG not just to its small volume, as the authors do, but the fact that there is no consensus in the field exactly where this area is, leading to likely variability in tracings.

In addition to the language changes mentioned above, to highlight the issue of definitions we have added the following paragraph to the conclusion of the paper (Page 30, lines 703-715):

> *This points to a larger issue of how to truly validate subfield segmentations, both in high resolution images and in standard T1-weighted images. There are several manual subfield segmentation methodologies, and they do not agree on which regions can be differentiated, even on high-resolution scans. See Table 9 for a comparison of MRI-based manual subfield segmentation methodologies. A further complication is that different researchers have differing operational definitions for the subfields and how they ought to be parcellated. The disagreement in the community has led to an international working group devoted to normalizing the ontology and segmentation rules for the hippocampal subfields (http://www.hippocampalsubfields.com/). In addition, there have been recent advances from the Yushkevich group to revise their MRI subfield segmentation protocol based on anatomy discerned from serial histological acquisitions (Adler et al., 2014). The definitional and operational disagreements suggest that direct comparison across automated methods using "ground truth"-based overlap similarity metrics, such as Dice's Similarity Coefficient, are not possible without carefully taking into account the differences in underlying segmentation protocols and image characteristics.*

To highlight the definitional ambiguities we have included a table contrasting the segmentations of several manual MR subfield segmentation protocols. (Page 29, Table 9):

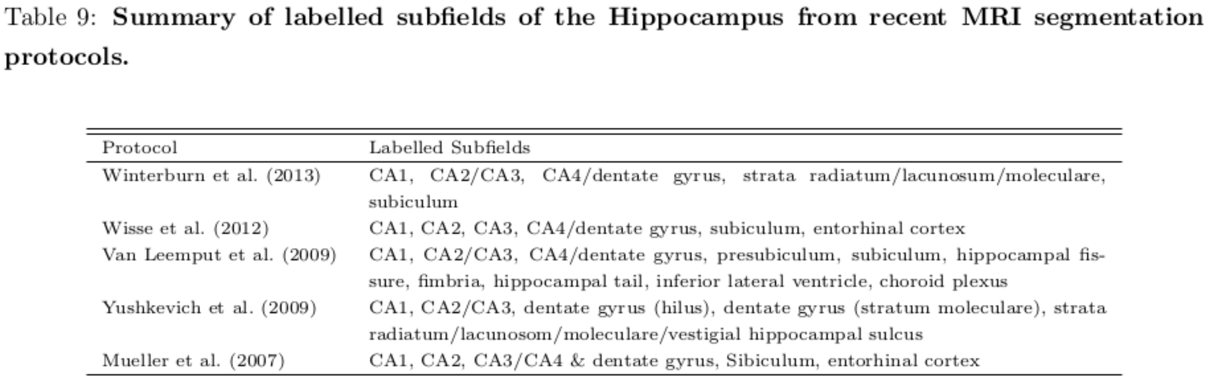

Table 9: **Summary of labelled subfields of the Hippocampus from recent MRI segmentation protocols.**

Protocol	Labelled Subfields
Winterburn et al. (2013)	CA1, CA2/CA3, CA4/dentate gyrus, strata radiatum/lacunosum/moleculare, subiculum
Wisse et al. (2012)	CA1, CA2, CA3, CA4/dentate gyrus, subiculum, entorhinal cortex
Van Leemput et al. (2009)	CA1, CA2/CA3, CA4/dentate gyrus, presubiculum, subiculum, hippocampal fissure, fimbria, hippocampal tail, inferior lateral ventricle, choroid plexus
Yushkevich et al. (2009)	CA1, CA2/CA3, dentate gyrus (hilus), dentate gyrus (stratum moleculare), strata radiatum/lacunosom/moleculare/vestigial hippocampal sulcus
Mueller et al. (2007)	CA1, CA2, CA3/CA4 & dentate gyrus, Sibiculum, entorhinal cortex

Despite definitional disagreements, previous work from our group (Winterburn et al. 2013) show that the subfield tracing protocol is highly reliably (Page 262, Winterburn et al. 2013):

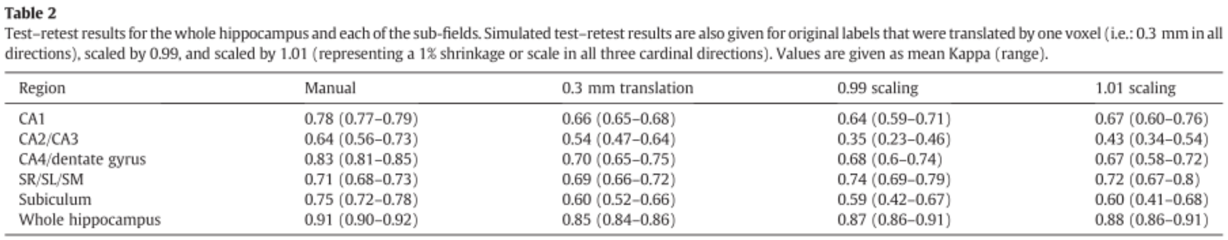

Table 2
Test–retest results for the whole hippocampus and each of the sub-fields. Simulated test–retest results are also given for original labels that were translated by one voxel (i.e.: 0.3 mm in all directions), scaled by 0.99, and scaled by 1.01 (representing a 1% shrinkage or scale in all three cardinal directions). Values are given as mean Kappa (range).

Region	Manual	0.3 mm translation	0.99 scaling	1.01 scaling
CA1	0.78 (0.77–0.79)	0.66 (0.65–0.68)	0.64 (0.59–0.71)	0.67 (0.60–0.76)
CA2/CA3	0.64 (0.56–0.73)	0.54 (0.47–0.64)	0.35 (0.23–0.46)	0.43 (0.34–0.54)
CA4/dentate gyrus	0.83 (0.81–0.85)	0.70 (0.65–0.75)	0.68 (0.6–0.74)	0.67 (0.58–0.72)
SR/SL/SM	0.71 (0.68–0.73)	0.69 (0.66–0.72)	0.74 (0.69–0.79)	0.72 (0.67–0.8)
Subiculum	0.75 (0.72–0.78)	0.60 (0.52–0.66)	0.59 (0.42–0.67)	0.60 (0.41–0.68)
Whole hippocampus	0.91 (0.90–0.92)	0.85 (0.84–0.86)	0.87 (0.86–0.91)	0.88 (0.86–0.91)

I think the authors need to be clearer about the fact that only the Yushkevich et al. method builds subfield segmentation atlases based on post-mortem brains, something that the Winterburn atlas does not do.

This is true of Yushkevich et al. (2009) only whereas Yushkevich et al. (2010) studies manually segmented in vivo focal MR images. As mentioned above, we have included more detail about these studies in the introduction. In addition, we have also discussed the recent work from that group from Adler et al. (2014) in the Discussion to highlight the differences between our work and the Yushkevich work.

My suggestion would be for the authors to consider other segmentation methods which are more widely accepted in the field, such as those of Mueller et al. 2007. Does this approach lead to better reliability? My guess is that it would. The methods proposed here would appear equally compatible with other tracing methods and it would be helpful to know if this methodology would also work [with] more conservative approaches or those approaches better rooted in ultra-high-resolution (post-mortem) templates.

We thank the reviewer for raising this point and we do believe that this is actually one of the great strengths of our methodology. As we have demonstrated in the segmentation work for the whole hippocampus (Experiments 1, 2, 3, 5) we can easily accommodate other methods. However, one of big differences between our approach and the approach of others is how we have chosen to define our subfields. So rather than try to estimate volumes on T2-weighted data that is of high-resolution in the coronal plane and highly anisotropic along the anterior-posterior axis, we asked where it was possible to actually identify subfields in data that approximates that T1-weighted data that are used in standard research practices. In addition, our algorithm relies on whole brain registration. It is unclear at the moment how the registration methodologies would have to be changed to accommodate data that has voxel dimensions of 0.4mm x 0.5mm x 2mm. Further to the best of our knowledge, Mueller's work has not been made publicly available, which would require that we adapt her segmentation protocol to our data, which has higher resolution in the anterior-posterior. It is unclear at the present time how to do this. We do believe, however, that the reviewer's comment is an extremely valuable one, and we are considering how to do this in future work. To this end, we have added the following the Discussion section of our manuscript:

> *Experiments 1, 2, and 5 have demonstrated that our algorithm flexibly accommodates different whole hippocampus manual segmentation methodologies. We have not explicitly evaluated a subfield definition other than the Winterburn protocol, and therefore it is possible that using an alternate subfield definition could improve the reliability of our automated subfield definitions. For example, established definitions such as those from Mueller et al. (2007) could be a prime candidate for further exploration. In addition, the conservative nature of the Mueller definition (labelling of the 5 slices in the hippocampus body only) would likely further aid in reliability measurement. However, there are two main logistical problems that we would have to overcome prior to implementation. The first is that these definitions were developed for data that is highly anisotropic (0.4mm × 0.5mm × 2mm), and it is unclear how our algorithms would deal with such atlases used as input. The second is that, since these atlases are not publicly available, we would be have to re-implement the protocol using our atlases. At the present time it is unclear how we would adapt these protocol to data that we used, where subfield segmentations are defined on 0.3mm^3 voxels. However, the impact of subfield definitions in the context of our work is an important one and should be considered in subsequent studies.*
> (See Page 29, Lines 677-689).

While the Yushkevich et al. method does segment the DG from CA fields, this is based on post-mortem atlases, a fact that is not discussed or mentioned in the current manuscript.

As mentioned above, we compare and contrast subfield segmentation protocols from Yushkevich and others in Discussion, and in the Introduction we now explicitly discuss the post-mortem basis for the Yushkevich atlases (Page 4, Lines 127-132):

> *Yushkevich et al. (2009) manually segment hippocampal subfields on high-resolution (either 0.2mm^3 isotropic or 0.2mm × 0.3mm × 0.2mm resolution voxels) T2-weighted MR images acquired 127 from five post-mortem medial temporal lobe samples. Then, using nonlinear registration guided by shape-based models of the subfield segmentations, and manually derived hippocampus masks of the target images, the authors demonstrate accurate parcellation of hippocampal subfields in clinical 3T T1-weighted MRI volumes.*

2. Although the methods here are interesting and promising, I thought more comparison was needed with the Yushkevich et al. methods. Along the lines of point #1, it might be helpful to directly compare reliability for different subfields reported here with those of Yushkevich et al. I realize the authors did this to some extent but I think they could go even further. Again, given that the "ground" truth is not based on post-mortem data with this methodology (or at least is several steps removed from it), it is important to compare with methods that derive from MRIs collected on scans in which these segmentation can be made with significantly more accuracy (and subsequently validated

with histological staining).

In the discussion section (Page 30, Table 10) we have included a table which includes a direct comparison of the overlap similarity scores (Dice's) for our method, Yushkevich et al. (2010) and Van Leemput et al. (2009):

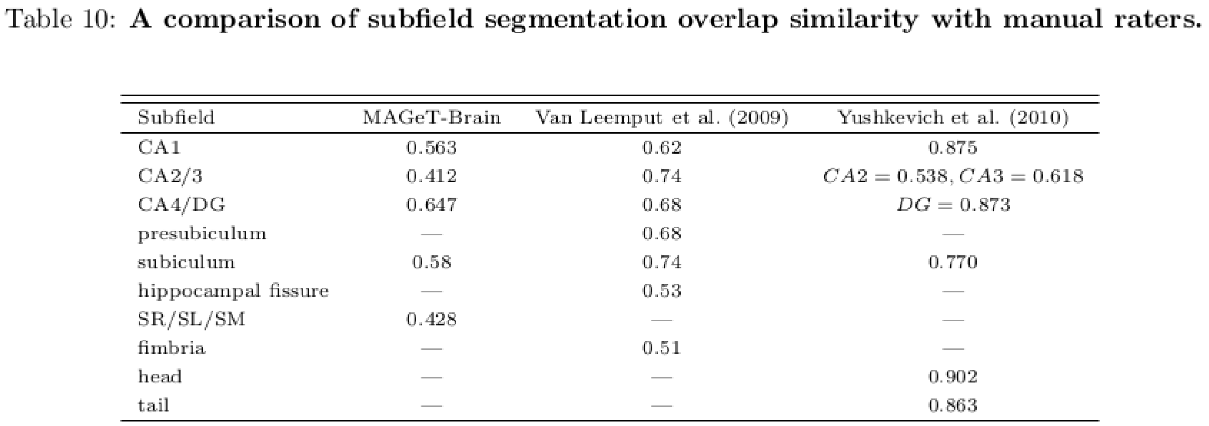

Table 10: **A comparison of subfield segmentation overlap similarity with manual raters.**

Subfield	MAGeT-Brain	Van Leemput et al. (2009)	Yushkevich et al. (2010)
CA1	0.563	0.62	0.875
CA2/3	0.412	0.74	$CA2 = 0.538, CA3 = 0.618$
CA4/DG	0.647	0.68	$DG = 0.873$
presubiculum	—	0.68	—
subiculum	0.58	0.74	0.770
hippocampal fissure	—	0.53	—
SR/SL/SM	0.428	—	—
fimbria	—	0.51	—
head	—	—	0.902
tail	—	—	0.863

As mentioned previously, in addition, we have replaced our original subfield validation experiment with one that is more similar to that of Yushkevich et al, (2010). Our updated experiment is a modified leave-one-out cross-validation with the Winterburn atlas images and segmentations, described as follows:

> *In this experiment, the Winterburn atlases (Experiment 2, section 3.2) are resampled to 0.9mm-isotropic voxel resolution to simulate standard 3T T1-weighted resolution images. Image subsampling is performed using trilinear subsampling techniques. In each round of LOOCV, a single atlas image is selected and treated as a target image to be segmented by MAGeT-Brain. So as to have an odd-sized atlas library, atlas image is segmented once using each possible triple of atlas images, and corresponding manual segmentations, from the remaining four unselected atlases. Thus, for each of the five atlases, a total of 4 segmentations are evaluated, resulting in a combined total of 5 × 4 = 20 segmentations evaluated overall. We chose an atlas library with an odd number of images so as to ensure unbiased label fusion when using majority voting (see Discussion).*
> (Experiment 4, Page 21, Lines 493-501)

MINOR

1. Intro: "The hippocampus is a brain structure situated in the medial temporal lobe, and has long been associated with learning and memory (den Heijer et al., 2012; Scoville and Milner, 2000)."

The authors should also probably include reference to other work more specifically focusing on the hippocampus (the Scoville and Milner work involved the whole medial temporal lobe, not hippocampus exclusively). The authors might consider citing work by Squire or others in this context.

Agreed. We have included citations for the following works:

Jeneson, A., & Squire, L. (2012). Working memory, long-term memory, and medial temporal lobe function. *Learning & Memory*, *19*(1), 15–25. doi:10.1101/lm.024018.111

Wixted, J., & Squire, L. (2011). The medial temporal lobe and the attributes of memory. *Trends in cognitive sciences*, *15*(5), 210–217. doi:10.1016/j.tics.2011.03.005

2. "accurate identification of the hippocampal subfields is indeed possible using this methodology."

Use of accurate here is misleading as there is no consensus regarding exactly where subfields are, particular in areas like the head, using MRI. The gold standard for accuracy is comparing an in vivo MRI with a histological reconstruction in which staining is possible. This current method does not involve this and the authors should probably make this issue clear. Accuracy is of course only as good as the underlying knowledge of exactly where the subfields are. Because in vivo and ex vivo scans from the same participants are currently lacking, which would represent the true "gold standard," statements about "accuracy" should be tempered and considered carefully here.

Agreed. This sentence is now rewritten to read: "... reliable reproduction of hippocampal subfield segmentations in standard 3T T1-weighted images is possible." (Page 30, Line 719-720)

In addition, as noted earlier, we have removed mention of "gold standard" throughout the manuscript and instead refer to reliability and overlap similarity with the atlas segmentation protocol.

3. "We then performed a leave-one-out validation to determine if hippocampal subfields can be accurately identified using our multi-atlas framework."

Again, accuracy is relative here. I think what the authors mean is that the algorithm produced a good match to manual tracing. This should probably be stated as otherwise, accuracy comes across as misleading.

Agreed. This sentence in the Introduction (Page 5, Line 146-147) now omits "accuracy" and defers discussion about evaluation to section 3.4, Experiment:

> *...we investigate hippocampal subfield segmentation by conducting a leave-one-out validation using the Winterburn et al. (2013) manually segmented*

> *high-resolution MR atlases.*

4. The authors should probably clarify upfront (in the Intro) that atlas refers to manually segmented images based on a recent method developed by Winterburn et al. rather than something like the Duvernoy atlas (although that in turn is derived, in part, from this atlas).

We have specifically defined the term *atlas*, as well as other commonly confused terms such as *template* and *label* in Section 2, Page 5. Specifically, on line 151, we define an atlas: ,

> *We use the term atlas to mean a manually segmented image ...*

In addition, we have also been careful to include a subsection for each experiment describing the atlases and manual segmentation methods used in each.

More specifically, in the introduction to Section 3, which summarises the experiments described in our paper (Page 7), we now state (lines 191-192):

> *Experiment 4 assesses hippocampal subfield segmentation quality in a leave-one-out cross-validation on **the five high-resolution manually segmented Winterburn MR atlases (Winterburn et al., 2013)***

(Emphasis not included in the manuscript)

5. T1 resolution: 1: 25mm _ 1: 25mm _ 1: 2mm

This seems somewhat low to make statements about subfields. This should probably be clarified.

We apologize if this was unclear, but the T1-weighted images used for the segmentation of the subfields were 0.9 mm isotropic voxels (resampled from the original 0.3mm isotropic voxel Winterburn atlas images). We suspect that the reviewer is referring to voxel-dimensions of the ADNI data that was used in Experiments 1 and 3. We trust the reorganization of our work has helped in keeping track of the different data that were used.

6. What is the full voxel resolution of the EFGRE-BRAVO sequence?

We have included this in our manuscript, page 21, line 486-488:

> *... FGRE-BRAVO, with the following parameters: TE/TR/TI =3.0ms/6.7ms/650ms, flip angle=8° , FOV = 15.3cm, slice thickness=0.9mm, 170 in-plane steps for an approximate 0.9mm isotropic voxel resolution*.

7. "Hippocampal MAGeT-Brain-based segmentations using both ANIMAL and ANTS

registration algorithms demonstrate good overlap with SNT Gold Standard segmentations (maximum mean DSC of 0.84 when using 9 atlases, 19 templates, ANTS registration, and majority vote label fusion); Figure 3). Qualitatively, both ANIMAL and ANTS-based segmentations demonstrate trend overlap accuracy that increases with the size of atlas library and template library. Improvement in accuracy plateaus with template libraries larger than ten images."

The authors should report actual statistics here

This paragraph now includes mean DSC scores to illustrate our claim (Experiment 1, Page 11, lines 318-322):

> *We find that for MAGeT-Brain segmentations, similarity score increases as atlas and template library size is increased, although with diminishing returns and an eventual trend towards a plateau (Figure 2a). For instance, with 9 atlases and using ANTS for registration and majority vote fusion, the mean DSC scores for 1, 5, 9 and 17 templates are 0.844, 0.865, 0.867, 0.868, respectively. A maximum similarity score of 0.869 is found when using 9 atlases, 19 templates, ANTS registration, and majority vote label fusion.*

8. "The use of MAGeT-Brain with ANTS registration shows a pronounced increase in segmentation accuracy over MAGeT-Brain with ANIMAL registration, across all other variable settings we tested. Additionally, by itself, using a weighted voting strategy did not significantly improve segmentation accuracy, contrary to the findings of Aljabar et al. (2009) in basic multi-atlas segmentation. Given these findings, in the remainder of this section only results using the ANTS registration algorithm and majority vote fusion will be shown." Again, these statements are lacking any statistical data to back them up.

We have included a summary of Pearson correlations between ANTS and ANIMAL-based results (Experiment 1, Page 11, lines 323-330):

> *The ANTS registration method consistently outperforms ANIMAL registration over all variable settings we tested (mean increase in DSC is 0.078). Pearson correlations of DSC scores when using weighted voting and when using non-weighted majority vote label fusion (with ANTS registration) for all combinations of atlases and templates are r > 0.899, p < 0.001, with a mean difference in DSC score of 0.002. This result suggests that using a weighted voting strategy does not significantly improve MAGeT-Brain segmentation agreement, contrary to the findings of Aljabar et al. (2009) for basic multi-atlas segmentation. Thus, in the remainder of our experiments only results using the ANTS registration algorithm and majority vote fusion will be shown.*

9. "With an increasing number of templates, MAGeT-Brain shows improvement over

multi-atlas-based segmentation in overlap accuracy when using the same number of atlases and voting method (Figure 4). The two methods converge in accuracy when using seven atlases. Peak improvement in MAGeT-Brain accuracy (0.02 DSC) is found when one atlas is used with a template library of 19 images."

Once again, some statistical tests are needed to back up these statements. Just showing errorbars on plots is not sufficient.

We have included a Pearson correlation MAGeT and multiatlas scores and a mean increase over all atlas and template library size (Experiment 1, Page 11, Lines 331-334):

> *With at least five templates, MAGeT-Brain consistently shows a higher DSC score than multi-atlas segmentation wth the same number of atlases: r = 0.94, p < 0.001, mean DSC increase = 0.008 (Figure 2b). The magnitude of DSC increase grows with template library size but shows diminishing returns with larger atlas libraries. Peak increase (+0.025 DSC) is found with a single atlas and template library of 19 images.*

10. "In general, across hippocampal subregions the percent error in volume of MAGeT-Brain segmentations (relative to the full-resolution Winterburn atlas segmentation) compares favourably to the error resulting from image resampling (Figure 6). In particular, the CA1, CA4/DG, and SR/SL/SM subregions all show a percent error in volume that is at or lower than resampling error. The Subiculum and CA2/CA3 subregions show distinctly larger percent error in volume than is found through resampling."

Again, statistics are needed to back up statements here.

The updated results section for this experiment now includes (Page 22, Lines 523-531):

> *Figure 6a shows the overlap similarity scores between the MAGeT-Brain segmentations and the resampled Winterburn atlases for each hippocampal subfield across all subjects and folds of the validation. Mean and standard deviation DSC scores of the subfields are shown in Table 7, along with DSC scores for the resampled atlas segmentations when perturbed slightly and compared to the originals. We find that the CA4/DG subfield shows the highest mean DSC score of 0.647±0.051, followed by the Subiculum and CA1 subfields having scores of 0.563±0.046 and 0.58±0.057, respectively. Both the CA4/DG and molecular regions score below 0.5. These scores may seem low but not when taken in context and compared to existing (semi-)automated methods (see Discussion). The whole hippocampus is segmented with a mean DSC score of 0.816±0.023.*
>
> *Figure 6b contains Bland-Altman plots comparing MAGeT-Brain volumes with manual volumes across all validation folds. MAGeT-Brain displays a conservative proportional*

> *bias — small hippocampi are over-estimated in volume, and larger hippocampi are underestimated (a mean maximum difference of approximately 200mm^3 across all subfields). MAGeT-Brain display a slight conservative fixed bias, tending to underestimate all subfields except CA4/DG (mean underestimation: CA1 = 76mm^3, CA2/3 = 56mm^3 , CA4/DG = −16mm^3, Subiculum = 48mm^3, SR/SL/SM = 96mm^3).*

Reviewer #2

This paper describes the MAGeT procedure that maps the hippocampus from an adaptive set of templates based on mapping of pre-existing atlases, and establishes an optimal set of atlases and templates to be used in applications of schizophrenia and AD datasets. This work is clearly important as it creates a framework where existing atlases can be used to create templates in populations under study without the need to create manual segmentations which is time consuming and not achievable by many groups. The second contribution of this work is that it shows large hippocampal subfields can be segmented in 3T scans with acceptable errors.

In general, the validation process and results of experiments 1, 2 and 4 are thoroughly described. Some issues with regard to experiment 3 need clarification (below).

Major concerns:

1. Experiment 1 uses SNT segmentations in 69 scans to cross validate the MAGeT procedure. The issue with using SNT segmentation is two-fold:

a. SNT segmentations come from semi-automated maps of one atlas. This means that all 69 segmentations contain an inherent common characteristic from the same atlas subject's segmentation. Mapping these segmentations back onto each other may in fact create a biased result (i.e., artificially improved accuracy) than if all 69 segmentations were manually delineated.

b. Even though the SNT segmentations have been shown to have ICC of 0.94, they are still not the same as manual segmentations, as shown in Figure 12 some of them have inaccurate labeling.

Therefore, a better approach would be to use the manual segmentations that the SNT segmentations compared to, if it is possible to obtain them. Or use a dataset that has manual segmentations to start with.

The reviewer raises an excellent point about the semi-automated nature of SNT labels potentially leading to biased MAGeT-brain segmentation results either because of inherent bias in automated procedure (leading to higher reported accuracy) or because of poor segmentation consistency (leading to lower reported accuracy).

Therefore, we chose to recreate our original cross-validation experiment using 60 manually segmented ADNI baseline images segmented by co-author Jens Pruessner (Experiment 1). The original SNT cross-validation experiment with 69 subjects has been moved to the Supplementary Materials (Experiment 5), since comparisons to SNT segmentations are widely reported in the literature and so serve as an informal benchmark between algorithms. Note, due to the lack of manual segmentation available for the entire ADNI dataset, we chose to use the SNT labels as our reference labels for this work. We have also discussed the pitfalls of using this segmentation approach on pages 27-28, lines 613-620:

> *On that note, one author (JW), an expert manual rater (Winterburn et al., 2013), identified regular inconsistencies in the SNT segmentations: occurrences of over- and under-estimation, as well as misalignments of the entire segmentation volume (Figure 5). Although the SNT segmentations are used as benchmarks for validation in many other studies (Table 8), these segmentation inconsistencies present the possibility that a more accurate and consistent benchmark segmentation protocol ought to be used in order to truly understand the results of such validations. Indeed, our replication of the 10-fold cross-validation using SNT segmentations (Experiment 5, Supplementary Materials) shows noticeably poorer mean similarity scores for both MAGeT-Brain and multi-atlas.*

Another option would be to use the FEP dataset of 81 scans all with manual segmentations.

In addition to the redesign of Experiment 1, we also chose to replicate the experiment using the FEP dataset. In this revision to the FEP experiment (Experiment 2), we include a 5-fold Monte Carlo cross-validation using all 81 scans with manual segmentations (see point 2, below).

2. For experiment 3, the optimal parameters (number of atlases, templates) for other experiments were determined on experimenting with AD (with elderly normal) data. How exactly they translate into a younger cohort is not clear. It would be beneficial to see that the same cross validation procedure on the FEP data would yield similar parameters or plots (as figures 3-5).

As noted above, we included a cross-validation in the FEP experiment similar to the cross-validation performed with 69 ADNI subjects, but on the younger first episode psychosis dataset. To simplify things, we used ANTS for registration and majority vote fusion, but varied the number of atlases (1-9) and templates (1-19) and report the change in Dice overlap score.

3. The FEP experiment is not described well to be included in the manuscript in its current form. Further experiments should be conducted. For example:
a. The evaluations should follow those used in experiment 4: gold standard, FSL,

MAPER.
b. The evaluation should also include controls as did for the ADNI data.

In addition to the cross-validation using the FEP dataset described above, we have also expanded upon the analysis of the single MAGeT-brain segmentation of the entire FEP dataset. Specifically, we evaluate our segmentations with those produced by FSL and FreeSurfer (MAPER is unavailable as it is not distributed nor open-source). See Experiment 2 results (Section 3.2.2, Page 18), and Figure 3, page 16.

Minor concerns:

1. For experiment 2, the errors obtained on subfields should be compared with those obtained based on FreeSurfer subfield labels as well.

Since the subfield segmentation protocol differs between the method used by FreeSurfer (van Leemput et al. 2009) and that of the Winterburn atlases, we felt it would not be meaningful to directly compare segmentations of the same images. Instead, as noted above (in response to comments from Reviewer #1), in the discussion section we now include a comparison of subfield segmentation protocols (Table 9, Page 29), and a comparison of reported subfield segmentation accuracy from Van Leemput et al. 2009, Yushkevich et al. 2010, and from our subfield segmentation validation in Experiment 4.

2. In Figure 12, the SNT is labeled as "Manual." SNT segmentation, as mentioned in the manuscript, is derived from semi-automated mapping of an atlas, not manually performed.

Fixed. All references to the SNT segmentations now refer to these as SNT labels, and that they are semi-automated segmentations, not manual.