Response to NeuroImage Reviewers

Reviewer #1:

Content:
The authors presented a novel neuroanatomical preprocessing pipeline for cortical thickness estimation (ANTs). For evaluation public available databases like IXI and OASIS where used. There results were compared to FreeSurfer in a scan-rescan test (reproducibility). Furthermore, the results of both methods where used in a age and gender prediction test based on a simple linear model.

Major comments:
I really like that you evaluated your method on large datasets and that both method and data are public available. Figure 1 provides a nice overview about your method and the relation of the different parts. There are a lot thickness studies, but relatively less method paper for a method paper. I think for a motivation of cortical thickness analysis some important examples would be enough. On the other hand I would expect more methods for thickness estimation especially volumes based (in example Hutton2008, ...), or full pipelines like BrainVisa, BrainVoyager, BrainSuite, etc.

We agree with the author regarding evaluation of publicly available methods on publicly available data sets as they encourage good scientific practices such as reproducibility. While many of the individual pipeline components which comprise antsCorticalThickness have been evaluated elsewhere in prior publications, the contribution of the work in this manuscript includes the tuning, coordination, and packaging of all these components into a single, intuitive, automated pipeline coupled with a comparative evaluation with the most widely used cortical thickness pipeline, FreeSurfer, on a large, publicly available cohort. While this pipeline may lack front-end visualization such as that offered by BrainVisa, BrainVoyager, etc., we believe that its open source availability and documented performance on such a large data set warrants attention by the NeuroImage readership. Furthermore, this is perhaps the largest-scale evaluation yet attempted on control subjects from diverse datasets.

Another important part of the preprocessing is the correction of noise (like FSL Susan, MRF-Filter in the segmentation, or NLM-Filter) and I am surprised that you did not mention it at all, although I expect that noise-correction is part of our segmentation.

We added the following to the end of the first paragraph in the Methods section:

> Although other preprocessing components are possible (e.g., noise reduction as in Smith (1996)), the major steps constituting the ANTs pipeline are limited to those enumerated above.

We also note that the Atropos segmentation procedure incorporates Markov random field (MRF) regularization and also benefits from additional smoothing related to the smoothness of the probabilistic priors.

The section and figure about the brain extraction is a little bit to short, because there is no full reference paper (the reference is just a poster). The examples in figure 2 does not proof that your method works better than FreeSurfer or another method. Please use the BWP (Brain Web Phantom,http://brainweb.bic.mni.mcgill.ca/brainweb/) and the SVE (Segmentation Validation Engine,http://sve.bmap.ucla.edu/) for validation.

We agree that the qualitative results in Figure 2 do not prove that our method works better than FreeSurfer. They were simply meant to convey representative qualitative differences from a randomly selected sample (sample figures were selected using the random number generation capabilities in R). Unfortunately, a full paper detailing the brain extraction method has not been prepared nor are there immediate plans to do so. Assessment of our method using the SVE has been on the first author's to-do list since September 28, 2013 but, due to the recent transition of LONI from UCLA to USC, the SVE has yet to be put back online. As is written on the website (http://sve.bmap.ucla.edu), "We are currently transitioning SVE to a new webserver, during which time only the Results Archive is available." When SVE is ready, we will be sure to post the results of our brain extraction. We did, however, use the Brain Web Phantom data to evaluate the performance of our Atropos segmentation tool (and reported the performance in http://www.ncbi.nlm.nih.gov/pubmed/21373993) which factors heavily into our brain extraction algorithm.

The multivariate processing sounds really good, but it plays no further role. It would be nice if you can show an example, an application or a figure to motivate this introduction or remove it otherwise.

Done. We have removed the description of the multivariate capabilities of the script antsMultivariateTemplateConstruction.sh.

Figure 5 show to many regions to get useful overview about the average result of both methods. Please add an average error of all regions/voxel/vertices. Furthermore, an average brain surface that codes the error of each regions would be nice. Maybe you can combine this figure with table 1.

Done. We added Table 2 with a reference in the text.

You mentioned correctly the advantages and limits of the repeatability test for the accuracy of a method. But for validation/evaluation there are multiple ways to simulate artificial thickness phantoms like spheres, cubes etc., where you can test methods under well known conditions.

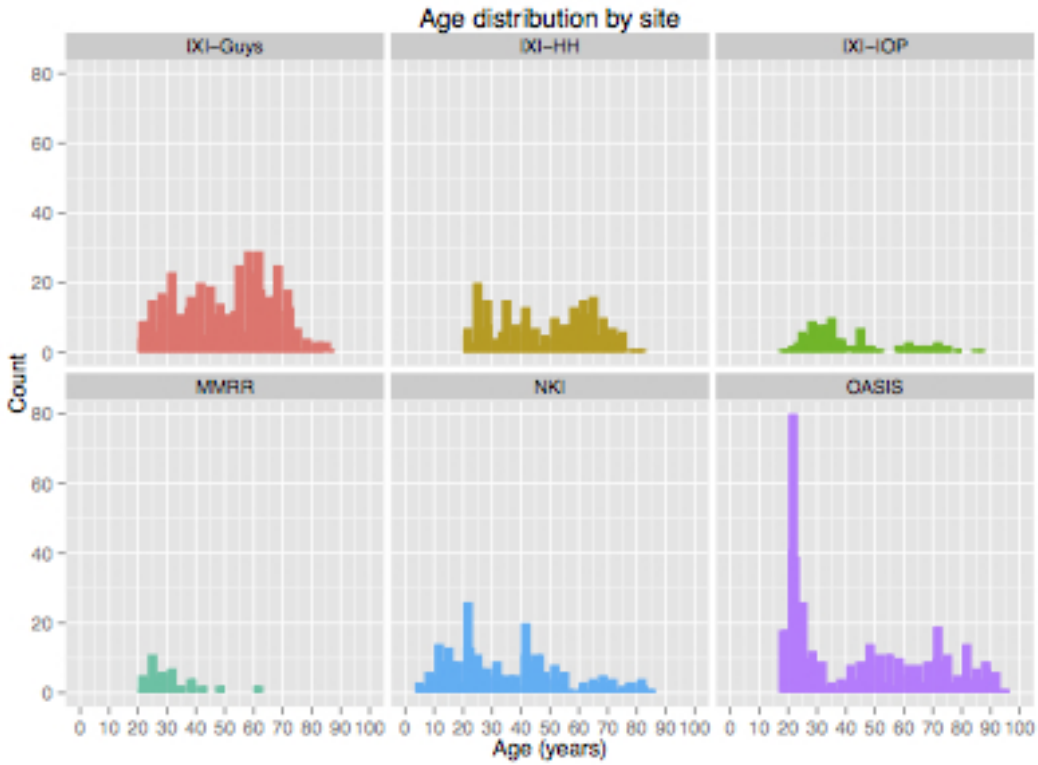
Indeed, there are numerous simulations that can be carried out to specifically validate thickness methods (apart from other preprocessing steps like segmentation). In fact, in the original DiReCT paper, such simulations were performed and reported (http://www.ncbi.nlm.nih.gov/pubmed/19150502). Instead of evaluating individual components (which has been performed elsewhere), we are more interested in an holistic evaluation of the ANTs cortical thickness pipeline where assessment is performed on the output of the entire workflow. This has not been done before using measurements of clinical relevance on such a large scale against a well-known package such as FreeSurfer.

Furthermore, I think that real data can not proof the accuracy of a method, because ground truth is unknown. Of course we have our expectations and models, but if a measure fits better for a models, it does not proof that the measure is more exact or correct than another. The large number of thickness studies shows that nearly everything influence cortical thickness. On one side, this makes thickness very interesting, but due to that high variation can be always some side effects. A linear model for cortical changes in thickness or volume does not fit very well for this large age range (I think there should be some Fjell papers).
Therefore, the age and gender predication is a nice example of application of your tools, but its no proof that your approach works better than FreeSurfer.
Please write your observations in a more conservative style.

We agree with the reviewer---certainly the outcome of our comparative prediction modeling is not absolute proof that ANTs cortical thickness measurements are more accurate than those of FreeSurfer. However, we believe we are quite measured in our claims, specifically in the introduction, where we write:

> Although FreeSurfer validation has included histological [75] and image-drawn [51] comparisons, such manual assessments were extremely limited in terms of number of subjects and the number of cortical regions. In addition, there was no mention in these studies of the number of human observers making these measurements nor discussion of quality assurance. Alternatively, without ground truth, other forms of evidence can be adduced [e.g., 10] in making comparative inferences. In this work we use demographic-based assessments (based on well-studied relationships between cortical thickness and age/gender) to show that ANTs outperforms FreeSurfer-based thickness estimation for these data in terms of prediction.

The reviewer also has a point concerning the age range for the prediction modeling particularly with the addition of per-site comparative prediction modeling (see Table 3). Consistent with the age ranges across all sites shown in the figure to the left, we restricted predictive modeling for both ANTs and FreeSurfer pipelines to the age range [20,80] (except for MMRR which was



excluded based on the small cohort size and limited age range). This is also similar to the age range used in Hogstrom et al. "The Structure of the Cerebral Cortex Across Adult Life: Age-Related Patterns of Surface Area, Thickness, and Gyrification", Cerebral Cortex, 2013.

Minor comments:
The article could be a little bit shorter and more direct.

In addition to the edits previously mentioned, we also removed non-essential paragraphs originally in the Introduction.

2. Method:
I would be nice to have a separate subsection for the usability of your method. Furthermore, I think I would be enough to say that your method is a set of shell-scripts similar like FreeSurfer, that it works therefore on different platforms on a parallel level. Detailed shell commandos should may be mention in the appendix or on a website or in the help file.

We agree and have removed the scattered set of shell scripts listed as footnotes in the original manuscript. They are currently listed in Table 4 in the Appendix.

2.2 Public data resources

What is the important relation between the MICCAI 2012 Grand Challenge and this work? The way how you create your templates does not fit in the data section very well.

We clarified by rewriting this sections as follows:

> To generate the priors for each T1 template, we used the multi-atlas label fusion (MALF) algorithm of [96] in conjunction with a labeled subset of the OASIS data set.

and added the following footnote to clarify which subset of OASIS data set:

> These data were originally acquired by the first and last authors to aid in the MICCAI 2012 Grand Challenge and Workshop on Multi-Atlas Labeling. The data was released under the Creative Commons Attribution- NonCommercial license. Labelings were provided by Neuromorphometrics, Inc. (http://Neuromorphometrics.com/) under academic subscription.

3.1 Repeatability
The error measures belongs in the method part.

We prefer to keep the repeatability measures in the Results section similar to other works (e.g., [45]).

3.2 Age prediction assessment
The section "Previous research ... in the literature." belongs more to the discussion part, whereas the description of the prediction belongs to the method part.

Done. We moved the paragraph "Previous research…" to the Discussion section.

What is VOLUME in equation (3) and (5)? TIV?

Yes, we added this clarification in the text.

2.4 Computation time
Well, I think Figure 8 is dispensable and that the average time are enough.

Done. We removed Figure 8.

4. Discussion
Your check is good to find strong outliers, but checking of some slices says nothing about the total brain-extraction quality. I think you need not to impress the reader with our heroic deed. Simply writing that an expert checkt for major problems would be sufficient.

Done. These two sentences were removed in the Discussion. We also modified a similar statement at the end of the Methods section to read

> The brain extraction and segmentation results from both pipelines were visually inspected to screen for major problems.

4.1 Repeatability of thickness measures
I miss the goal of the section "A crucial component ... OASIS scan-rescan data."

We agree and removed the paragraph.

There is a "." missing after "... (mean Dice > 0.58) Despite ...".

Fixed.

4.3 Computation time
I think one sentences is enough to say that you focus on quality rather that fast calculations, especially because the speed of ANTs were not worse that FreeSurfer.I like the sentence about "garbage-in/garbage-out", but I think that it does not fit in the calculation time section.

We added the suggested clarification:

> Computation time for the registration and segmentation components of the ANTs pipeline are substantial but are not significantly worse than those of FreeSurfer.

and removed the "garbage-in/garbage-out" sentence.

Reviewer #2:

The goal of this paper is to introduce a new pipeline ANTs (Advanced Normalization Tools) to calculate cortical thickness with their own tool DiReCT (from Das et al, Neuroimage 2009) and then to compare its performance to the commonly used Freesurfer. It is applied to 1205 T1-weighted scans from four public image data bases that have quite variable acquisition protocols. The evaluation of each method depends on the repeatability metric within the same individuals scanned multiple times and their value in predicting changes of cortical thickness with either age or gender (after including a training set). Overall, automated pipelines that are robust and freely available are an important community resource. It is certainly useful to have more options.

My main comments are:
1. Three main results are provided for both pipelines: percent error variability of the 62 cortical regions (Fig 5), relative mean square error values of predicting age (Fig 6), and the area under the curve for an ROC analysis of gender prediction (Fig 7). ANTS and Freesurfer are

similar for the first result, but ANTs seems to outperform Freesurfer in the latter two results. Despite this, there is actually no raw data in the paper. There are no absolute cortical thickness maps over the full brain to evaluate between the two methods. Do they both show similar regional differences? Are there differences in absolute cortical thickness derived by the two methods in certain regions and what are the biases (and potentially why)? What about some cortical thickness versus age curves for the whole 1205 sample (whole brain or certain regions) for each method where each site has a different symbol and perhaps male and female are colored differently so one can peruse the changes with
age, the potential biases, and the degree of scatter in the actual data? Otherwise, I don't really get an impression on the similarities or differences between the two methods.

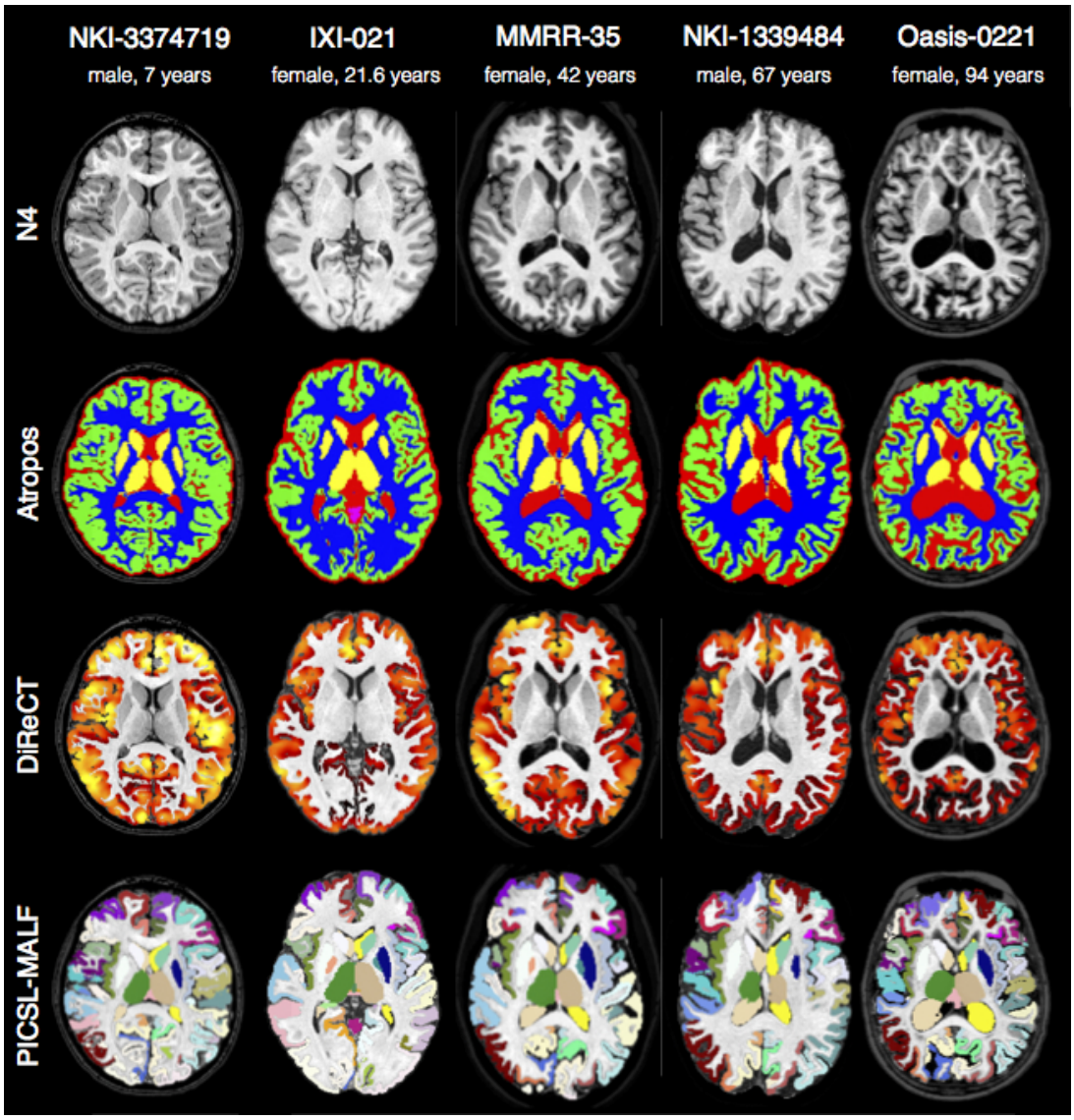
*We completely agree and actually did have such plots available (although not distinguished by site) before submission but they were ultimately not included. This has been remedied in the current submission (see Figure 8). The difficulty is determining which of the (62 regions / pipeline * 2 pipelines =) 124 plots to actually put in the paper so as to present an unbiased assessment of both methods and not double the size of the manuscript. We decided to average the random forest model variable importance measures from both pipelines and show the left/right counterparts of the most "important" regions based on that average. We also noted in the manuscript that all 124 plots can be found online with the address given in the Appendix. We also include, in online material, visualization of cortical thickness mapped to the surface in all processed t1-weighted MRI.*

2. One area of interest is the notion of whether the two software packages work well over diverse data sets acquired at different field strengths, resolutions, contrasts, vendors, etc. However, there is no data presented separately for each of the four data sets as they are all amalgamated as far as I can tell. Therefore, there is really no indication on how well the software works with each data set and whether a given acquisition protocol works much better than another.

As described in response to the previous point, we present this information in the plots of Figure 8 (and online) by including site as the shape descriptor for each point. We also performed a prediction modeling comparison per site and provide those results in Table 3. We also release the full summary csv files for these datasets which will allow interested readers to investigate the subsampling of the datasets that might serve their own interests. However, our main focus is indeed the amalgam of these datasets. Identifying optimal scanners or sequences is an area of work that has its own issues, including the quality control standards employed for individual scanners, and we have no access to such details for these public datasets. See Han's "Reliability of MRI-derived measurements of human cerebral cortical thickness: The effects of field strength, scanner upgrade and manufacturer" for additional details as well as related publications:
http://www.ncbi.nlm.nih.gov/pubmed?linkname=pubmed_pubmed&from_uid=16651008

3. It appears as though a different template was created for each data base (two for the NKI). While this might work when the mean cortical thickness over a larger region of interest is assessed (such as the 31 per hemisphere in this paper), this does not permit a vertex-based analysis of the entire combined group which would be of interest at this analysis procedure is commonly used in the literature.

One of the outputs of the pipeline outputs are subject-specific DiReCT cortical thickness maps (see figure below) which make the described pipeline perfectly compatible with voxel-based



analyses. Although not described at length in the original manuscript, the user can use the '-t' option to specify the atlas to which the subject is registered for analysis. For example, referring to listing 1, suppose we wanted to use a template built from, let's say, the ADNI control data ("adniControlTemplate.nii.gz") to perform a voxel-based analysis in that template space (while

using the IXI template for brain extraction and segmentation priors). We would simply specify this as follows:

```
antsCorticalThickness.sh \
      -a IXI/T1/IXI002-Guys-0828-T1.nii.gz \
      -e IXI/template/T_template0.nii.gz \
      -m IXI/template/T_template0ProbabilityMask.nii.gz \
      -f IXI/template/T_template0ExtractionMask.nii.gz \
      -p IXI/template/Priors/priors%d.nii.gz \
      -o IXI/ANTsResults/IXI002-Guys-0828-
      -t adniControlTemplate.nii.gz
```

This would cause the following additional files to be created

- `IXI/ANTsResults/IXI002-Guys-0828-CorticalThicknessNormalizedToTemplate.nii.gz`
- `IXI/ANTsResults/IXI002-Guys-0828-SubjectToTemplate0GenericAffine.mat`
- `IXI/ANTsResults/IXI002-Guys-0828-SubjectToTemplate1Warp.nii.gz`
- `IXI/ANTsResults/IXI002-Guys-0828-SubjectToTemplateLogJacobian.nii.gz`
- `IXI/ANTsResults/IXI002-Guys-0828-TemplateToSubject0Warp.nii.gz`
- `IXI/ANTsResults/IXI002-Guys-0828-TemplateToSubject1GenericAffine.mat`

where "`Template`" refers to the template specified by the -t option. For this evaluation paper, we limited comparison to an ROI analysis as this is the most straightforward and limits possible bias between converting surface to volume data or vice versa (cf. [46]).

We added the following paragraph to the discussion section:

> 4.2. Voxel/vertex-based analysis
>
> One of the limitations of our evaluation was the limitation of comparative analysis to mean ROI thickness values defined by the 62 cortical regions of the DKT atlas. Quite common in the literature, however, are point-wise (vertex- or voxel-based) analyses [e.g. 15]. The ANTs pipeline described in this work is equally applicable to such studies. The only additional requirement is the specification of the normalization template. For this work we opted for the ROI analysis to avoid potential bias issues when navigating between surface and volume representations [46]. Future work will certainly explore such analyses.

4. The paper is not written as a typical scientific paper focused on experimental results. Granted this is partly due to a description of the methods but it is filled with many anecdotal (non-scientific) comments, has 28 footnotes, and seems to rely on many external links.

We have removed most of these footnotes and links and relocated them to a table in the appendix. We also reduced discussion of implementation-specific details.

Other comments:

5. The first paragraph is unnecessarily overdoing the justification as there are 46 references (!) in the first paragraph alone on cortical thickness papers in a myriad of conditions. Just focus on a few key examples for your paper.

We prefer to keep the first paragraphs as they underscore the importance of having publicly available and mature cortical thickness pipelines for the myriad applications to which they are applied.

Similarly a 6 page Introduction is on the long side - get to the point of the paper more quickly on what is needed in the literature. Furthermore, there is a lot of background and justification in the evaluation section which should only be reporting actual results.

We removed two rather large, non-essential paragraphs and trimmed down much of the background and discussion of implementation details (which were relegated to Table 4 of the Appendix).

6. Figure 3 is referenced on page 12 before Figure 2 - Figures usually need to be referred to in order of appearance in the text.

This was the result of a latex issue which has since been corrected.

7. Page 13, Figure 2. It is confusing to show the templates, etc from the different data bases before the databases are described themselves.

We agree and have completely reorganized the methods section. In the current version of the manuscript, the data description was placed before the description of the individual pipeline components. Additionally, the template construction description is placed ahead of the other components as that figures in prominently to the brain extraction and brain segmentation steps.

8. Would it be useful to have a table with the key acquisition parameters for the four different data bases?

In the interest of brevity, we leave it to the interested reader to look up such parameters. These data are publicly available and the websites are listed in the Appendix.

9. Page 16. Not sure I find the names of the programs or footnote 7 appropriate or illuminating in a scientific paper.

Although these moments of partial levity may burden the reviewer, we do believe that scientific writing benefits from the presence of perhaps a few cultural references that serve to keep the more lighthearted reader interested.

10. Page 18. T2 and FA should not be mentioned as they have nothing to do with the current analysis.

Done.

11. Figure 3. The templates have quite different intensities and contrast. Is this partly due to differences in scaling in the presentation or is it all due to the acquisition?

There are distinct qualities in imaging data sets which lend to actual differences in intensities and contrast in the corresponding templates although presentation probably contributes to perceived differences. When creating the figure, the automatic scaling feature in ITK SNAP was used. We invite the interested reader to download these templates and look for themselves (http://figshare.com/articles/ANTs ANTsR Brain Templates/915436, listed in the Appendix). Also, see our earlier response regarding effects of scanner etcetera on T1-derived measurements.

12. Results, Figure 5, page 28. The repeatability in the 62 regions appears to be ~1-3%. Are there problematic areas and why would that be the case?

Indeed, Freesurfer and ANTs cortical thickness reliability measures are correlated (Pearson correlation = 0.44). We make a few clarifying statements in the paper regarding the possible sources of these reliability patterns:

> ANTs and FreeSurfer cortical thickness mean reliability are correlated across all regions (Pearson correlation = 0.44). Although our thickness reliability measurements represent the compound effect of registration, segmentation, anatomical labeling, and the thickness computation algorithm, this correlation suggests that these effects are non-random. That is, reliability measurements are influenced by characteristics intrinsic to the underlying neuroanatomy as represented in approximately one millimeter resolution volumetric T1-weighted MRI. Perhaps the least reliable region is entorhinal cortex (Region 4 in Figure 5) which has relatively small volume, is challenging to distinguish from surrounding structures [71] and is also relatively thin. Spatial variation in segmentation accuracy is known to relate to a structure's volume and tissue characteristics and this has led to a body of research on both segmentation and acquisition protocols that are optimized for specific regions. Perhaps the most substantial work in MRI has focused on temporal lobe structures including the hippocampus. Both FreeSurfer and our own group have optimized protocols to address such concerns (http://www.hippocampalsubfields.com/). Given caveats associated with cost vs. benefit, our current results suggest that optimized protocols may be relevant for additional cortical regions.

Reviewer #3:

In the manuscript, "Large-Scale Evaluation of ANTs and FreeSurfer Cortical Thickness Measurements", the authors examine the reliability and predictive ability of automated estimates of cortical thickness using these two packages. ANTs is developed by a number of the authors of the manuscript. Both are popular open-source efforts and I have experience with both packages. The authors report that both packages have similar test-retest reliability. This, of course, is a necessary component of a useful tool, but is insufficient (as simply giving the same answer 100% of the time would be reliable, but hardly useful). Lacking a "gold standard" of histology, they go on to show how thickness measures from ANTs can be used as better predictors of age and of gender using several freely available datasets. This was done using an exceptionally large population (with a total of 1200 brains) to ensure that results would be readily generalized.

The methods are well-described and scripts are provided in a git repository for ANTs (verified). In addition, several demonstrative examples are included in a separate git repository. The results are straightforward, clear, and of importance. Thus, my concerns are few and small:

1) All results have been presented as test-retest reliability or as how well thickness measures can predict age or gender. Can the authors provide some indication of the actual thickness estimates from each, how the two tools differ in this, and how these measures relate to histological "gold standards"? For example, how large is the effect of age physically on thickness where it is known histologically and as estimated here?

*Figure 8 was added to the manuscript which shows key regional age vs. thickness scatter plots where site is indicated by point shape. Since it would have been difficult to include all 124 such plots (62 regions / pipeline * 2 pipelines) in the manuscript, we only included a few of the most discriminative (determined in terms of both random forest models variable importance) and pointed the reader to the Appendix where we provide the online location of all such plots.*

2) The results of the modeling show we can better predict age or gender via ANTs, but no indication is given as to what regions are being most heavily loaded in the models. There are published reports of where age and gender show the largest effects on thickness. Are these regions being picked up as the main components in the models? Are the models consistent across ANTs and FreeSurfer or are they converging on different areas as being most relevant?

In addition to the regional age vs. thickness plots given in Figure 8, we also provide variable importance plots for the aggregated ANTs and FreeSurfer random forest age prediction models in Figure 7 so that the interested reader can view how consistent across pipelines are the various regional importances.

3) On p 19, the authors note that to generate priors for each T1 template, the multi-atlas label fusion (MALF) algorithm was used. It's not clear exactly whether this would be needed for general use or whether it is only there for the purposes of the modeling of age or gender based on thickness. In Figure 1, it suggests Atropos is used for 6-tissue segmentation, but here it appears MALF is used. This was unclear.

We added the following clarification in the caption of Figure 1:

> All template-based prior probability maps are generated prior to pipeline processing of each individual subject.

We also edited the description mentioned (formerly on page 19)

> To generate the priors for each T1 template, we used the multi-atlas label fusion (MALF) algorithm of [96] in conjunction with a labeled subset of the OASIS data set. First, we normalized the labeled OASIS subset to the template. We then performed MALF on the template using the normalized labeled data as input. This resulted in a labeled, parcellated template consisting of 100+ labels defining the different brain regions. We then condensed this template-specific labeling to the six needed for our analysis, viz., cerebrospinal fluid (CSF), gray matter (GM), white matter (WM), deep gray matter, brain stem, and cerebellum. For example, all cortical regions were assigned asinglelabelrepresentingthegraymatter. These binary masks were then smoothed using Gaussian convolution with a one voxel-width kernel. Since the labelings did not describe the extracerebral CSF, we augmented the CSF prior image with the CSF posterior output from running each template through the segmentation component of the above-described pipeline. This new CSF prior was then subtracted from each of the other five prior probability images and limited to the probability range of [0, 1]. The prior probabilities for the five templates used in this evaluation are given in Figure 4.