Response to the reviewers

Thank you for submitting your manuscript to NeuroImage. Your paper, referenced above, has been reviewed by experts in the field. Based on the comments of these reviewers, we feel a major revision would be appropriate

Handling Editor: The paper received mixed reviews. Reviewer 2 has raised very significant issues that definitely need to be carefully addressed. the most important issues pertain to the new results in Tables 2-3 that are not discussed in the text, but they indicate that the longitudinal method actually might be performing worse than the standard method. Addressing this issue is obviously critical, as the paper's significance hinges on it. Other comments were also made and need to be carefully addressed.

We are grateful for the continued participation on the part of the reviewers and editors in improving the submitted manuscript. We are also very appreciative of the opportunity to address the additional concerns of the reviewers and hope that the rejoinder below adequately addresses the substantive issues raised.

Regarding the critical issues mentioned by the Handling Editor as well as similar concerns voiced by the reviewers, we apologize for a serious oversight on our part. The construction of Tables 2 and 3 was incorrect as the ordering of the five pipelines was specified erroneously. In the interest of transparency, this is the corresponding GitHub commit where this issue was fixed:

https://github.com/ntustison/CrossLong/commit/a15f74c50b420b15e2563682244427fb6fc0177df

Again, we apologize for this error and the additional work that this error imposed on the reviewers and editors.

Reviewer 1

Most of the review comments has been addressed.

We appreciate the Reviewer's suggestions which certainly improved the work.

One minor concern:
1. Although the significance of the thickness slope in group comparison has been shown, it is important to also show the thickness values and changes ratio in each region by the proposed method.

The purpose of the paper is to introduce the ANTs longitudinal cortical thickness pipeline and motivate its use by demonstrating utility which is most directly achieved by a

performance comparison with the widely used FreeSurfer streams. Tables 2 and 3, alluded to by the reviewer, contribute to the purpose of the paper by combining visualization of statistical significance with confidence intervals. In contrast, similar tables of thickness and change ratios, per se, are limited in their ability to provide such a direct and interpretable visualization of relative performance. E.g., just because thickness changes are larger (or smaller) in ANTs relative to FreeSurfer, that would not imply that ANTs is necessarily a better (or worse) measurement tool. This consideration is separate from the potential of "clutter" introduced by including the additional one or more densely constructed tables necessary to provide such information for all 62 DKT regions for all five pipelines. Therefore, we do not believe it optimal to include such tables in the manuscript proper. However, we recognize the importance of having access to such supplementary information so, towards this end, we have organized the corresponding GitHub repository for easy navigation for the interested reader and specified this in the manuscript:

> After processing the image data through the various pipelines, we tabulated the regional thickness values and made them available as .csv files online in the corresponding GitHub repository [36]. We also provide the Perl scripts used to run the pipelines on the UCI cluster and the R scripts used to do the analysis below. Additional figures and plots have also been created which were not included in this work. For example, spaghetti plots showing absolute thickness and longitudinal thickness change are contained in the subdirectory CrossLong/Data/RegionalThicknessSpaghettiPlots/.

Reviewer 2

This is highly problematic revision of this paper, for which the core issue in the first submission was the lack of demonstration that the proposed longitudinal pipeline is better at quantifying atrophy rates (temporal change) compared to when all available time points are simply analyzed cross-sectionally:

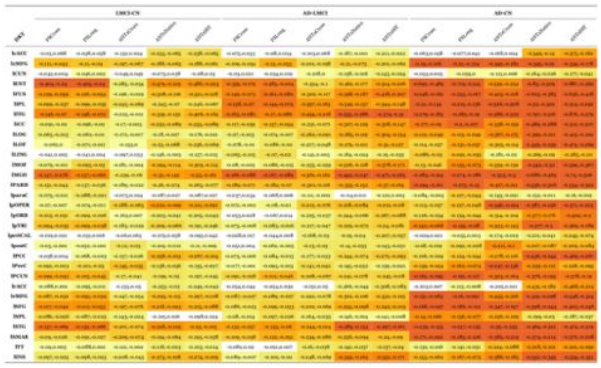
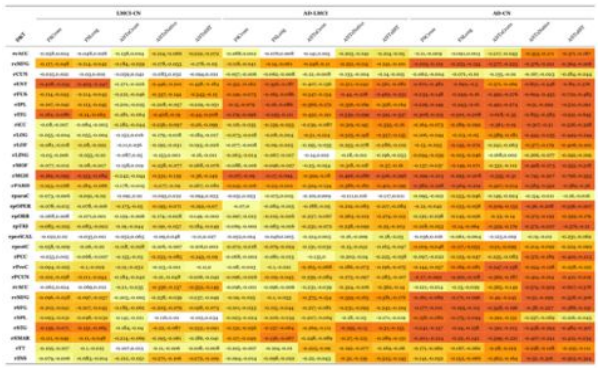
We appreciate the work on the part of the Reviewer for helping us to improve the original draft of the manuscript. Unfortunately, as mentioned earlier, we committed a significant error in the initial construction of Tables 2 and 3 as the ordering of the five pipelines was specified incorrectly. We apologize for this error and the additional work that this error imposes on the Reviewers.

Related, we reconfigured the Tables in two ways: 1) coloring is now based on log-scaled p-values and 2) columns are first ordered by pipelines with groupings based on diagnosis. We believe that this improves visualization of discernible differences in pipeline performance.

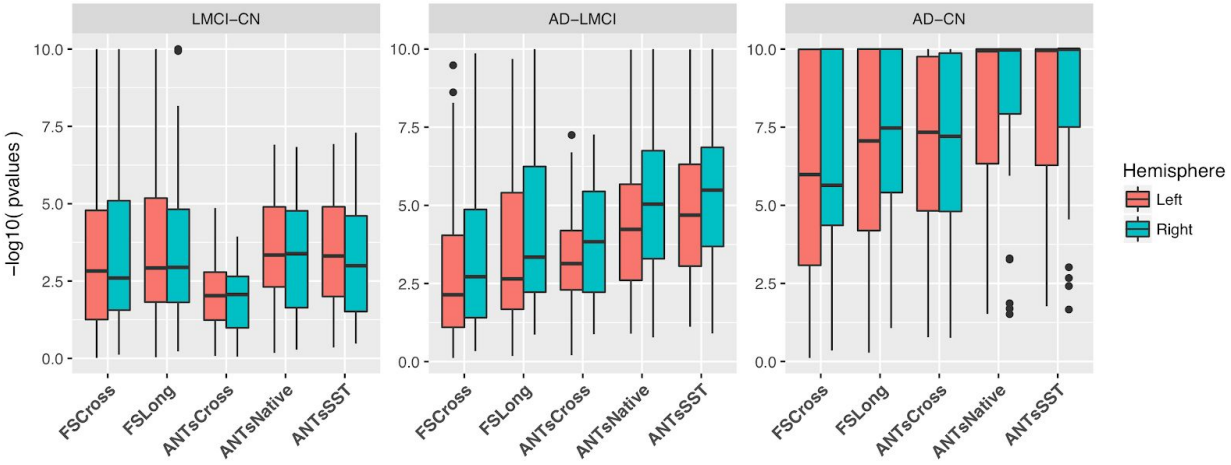
* The authors have now included two tables (table 2 and 3) with patient group differences in estimated atrophy rates in the ADNI data for several methods. Although the results shown in

these tables are not analyzed in the paper, and unsupported claims are made in the discussion section ("We demonstrate that our approach leads to competitive or superior estimates of annualized atrophy"), a closer inspection reveals that the proposed longitudinal method is actually *worse* at differentiating atrophy rates between e.g., controls and AD subjects than the cross-sectional method (or any of the other methods under comparison). What is the purpose of a longitudinal version of a method if it performs worse than the original method?

Again, we apologize for the incorrect ordering in Tables 2 and 3 in the previous manuscript. As one can see in the corrected tables below, the proposed longitudinal methods (Native and SST) do perform better than the original cross-sectional method. We appreciate the astute observation on the part of the Reviewer in identifying this issue and the subsequent motivation for enhancing visual differentiation through the reordering of the Table columns and log-based color scaling.

Table 2	Table 3
	

We have also summarized the Tables in the following set of box plots (Figure 9) where we have plotted the log scaled p-values for each pipeline and for each hemisphere.



* Although the authors insist that the ratio between between-subject variability and residual variability is a meaningful measure of performance of longitudinal methods, the new results in table 2 and 3 cast further doubts on this claim. In particular, the new results indicate that the proposed method over-regularizes across time, in effect making it insensitive to temporal changes. In this light, its low residual variability show in Figure 6 (left) may simply mean that the method simply reports the same thickness estimates across all time points, which is undesirable. Similarly, the high between-subject variability of the method (which really measures the variability in thickness estimates *at baseline time* across subjects, shown in figure 6 middle) may simply reflect that the method averages measurements over time -- effectively using all subsequent time points as mere repeat acquisitions to reduce measurement errors *at baseline*. Although this reflects sensitive estimates of cortical thickness *at baseline*, it does not measure the *rate of change* of these estimates, which is what longitudinal data is all about.

We invite the Reviewer to reassess the criticisms outlined above based on the corrected Tables. In particular, we hope this correction addresses the Reviewer's speculations concerning methodology and possible explanations for the residual and between-subject variabilities. It should be clear that we are not "simply report[ing] the same thickness estimates across all time points" nor are we simply "averag[ing] measurements over time." This seems rather obvious from both Tables where all ANTs-based pipelines demonstrate significance in differentiating diagnosis in the majority of regions with confidence intervals in the expected non-negative range. In addition, we note that the steps constituting the longitudinal cortical thickness pipeline outlined in Subsection 2.2.2 can be confirmed by the pipeline itself:

https://github.com/ANTsX/ANTs/blob/master/Scripts/antsLongitudinalCorticalThickness.sh

and the scripts used to do the analysis:

https://github.com/ntustison/CrossLong/blob/master/Scripts/Analysis/stan_plotResults.R
https://github.com/ntustison/CrossLong/blob/master/Scripts/Analysis/diagnosticSeparation.R

* I have a very hard time understanding/interpreting the two new regression/prediction experiments (equations 3 and 4). The results of the latter experiment (shown in figure 9) are hardly conclusive; whereas the results of the former (shown in figure 8) are puzzling: within a given patient group (e.g., controls), the authors effectively try to predict age from atrophy rate. But being able to do this would imply that the rate of atrophy is non-constant across age (thickness decreases non-linearly with age), and that the non-linearity of this effect can effectively be exploited even within the narrow age window of ADNI subjects (who are all older). Is this really true and/or what is expected?

The Reviewer is absolutely correct. Upon receipt of this round of reviews, the co-authors performed a re-assessment of the set of included experiments and determined that:

1) *The regression/prediction experiments described by Equations (3) and (4) were not very interesting nor were the findings compelling. Not only did they not elucidate performance capabilities across pipelines but, as mentioned by the reviewer, model assumptions were problematic. This took us away from the primary motivation for the manuscript which is to introduce the ANTs longitudinal cortical thickness pipeline and motivate its use with the ADNI data.*
2) *Perhaps motivated by the rescaling and and reconfiguring of Tables 2 & 3, as well as the comments from the reviewers, we felt that we gave short shrift to the experiments involving discernment of diagnostic groups. Using packages, such as FreeSurfer and ANTs, for identifying imaging biomarkers based on brain atrophy quantification, is of significant research interest.*

This reassessment led us to drop the prediction/regression experiments and opt for an expansion of illustrating performance differences based on regional diagnostic contrasts (see Section 3.2).

One minor comment: in equations 3 and 4, presumably regression coefficients are missing?

Although we removed Equations (3) and (4), we did not explicitly mention that the formulae were given in the notation of Wilkinson and Rogers (1973)---popularized with the use of the R language. This was an oversight on our part. This has since been corrected in the current manuscript in introducing a new Equation (3) with the inclusion of the following reference:

> *Wilkinson, G. N. and Rogers, C. E. (1973) Symbolic descriptions of factorial models for analysis of variance. Applied Statistics, 22(3):392–9.*

Reviewer 3

The manuscript under consideration is a revision of a previous submission. The primary changes include:

a. re-processing the data at a single site and with exactly the same software version
b. addition of details on the methods, in particular the registration and statistical evaluation
c. addition of a new subsection in the results part: evaluation of regional cortical atrophy rates obtained from a longitudinal analysis
d. addition of a new subsection in the results part: prediction of age and MMSE scores by estimates from the different pipelines
e. removal of a part of the results section which the authors "felt was not sufficiently general" (why is that, actually?)

"General" in the sense of being limited to only the entorhinal cortex whereas the added Results subsections include performance evaluation over the entire cortex.

f. extension of the discussion with two passages on further methods/performance description and on advantages/limitations

My overall impression is that these changes are appropriate and constitute an improvement of the manuscript.

> *We are grateful for the Reviewer's efforts in improving the manuscript. We are also appreciative of the additional issues raised by the Reviewer which we attempt to address below.*

Several comments and requests of the previous reviewers have not made it into the revised manuscript, but have been addressed by the authors' response to the reviewers. Below I will detail to what extent these remarks have been taken into account and what I perceive to be remaining issues that still need to be addressed.

There are two comments that were originally raised independently by more than one reviewer:

1. absence of a longitudinal analysis

The authors now include a longitudinal analysis and report differences in slopes between cognitively normal, MCI, and AD patients.

Results are presented by means of two color-coded tables that contain approximately 900 single estimates. As this format is not very accessible, graphical presentation, or a selection of relevant regions, is suggested to allow for more detailed evaluation.

> *We empathize with the Reviewer's concerns and so, as mentioned previously, we re-evaluated the presentation format and opted to make two major changes: 1) the color scaling is now based on the log-scaled p-values for better visual differentiation and 2) re-ordered the columns so that performance can be more easily compared between pipelines. We believe that this facilitates evaluation of comparative performance across regions for the reader who is interested in methodological considerations while providing confidence intervals for the clinically-oriented reader who is interested in AD-specific findings. These tables are reproduced below in our reporting of the expansion of Section 3.2.*

The evaluation of the results, then, should not be limited to the simple statement that "All pipelines demonstrate differentiating capabilities for the majority of cortical regions". This leaves it to the reader to draw any further conclusions. A more thorough evaluation, also with regard to the question whether or not performance improvements are also reflected in this analysis, is required.

We agree and have added Section 2.6 and Section 3.2. Further details below.

2. indirect evaluation of the algorithms performance

The authors now give an extensive justification for their primary evaluation criterion, the ratio of between-subjects and residual variances. This helps the reader to understand the motivation and meaning of criterion.

As I understand it, the argument goes as follows: whenever observations that necessarily include measurement error (such as cortical thickness) are used as explanatory variables in a linear model, one has to expect biased regression coefficients as well as increased residual variance. Both are inversely related to the proposed variance ratio, which can hence be interpreted as a measure of reliability ('reliability ratio').

An interpretation of this measure in terms of reliability, as the authors suggest, seems to contradict, however, their previous statement that "none of these precision-type measurements, per se, indicate the utility of a pipeline-specific cortical thickness value as a potential biomarker". I would agree with that statement, but add that reliability/precision is still a necessary, albeit not sufficient, requirement for a useful biomarker.

*We agree with the Reviewer's synopsis of the variance ratio. We would like to clarify that our criticism of *specific* instances of precision-type measurements proposed in the literature should not be generalized to include the variance ratio regardless of its possible interpretation as a type of reliability measurement. The variance ratio takes into account both between-subject and residual variance *as a ratio* which is optimal when minimizing the latter while simultaneously maximizing the former. Optimization of the precision-type measurements in [66,67] only minimize some measure of residual variance.*

Another issue with the variance ratio, in my eyes, is that its present description only covers the scenario of measurement error in covariates, i.e. the use of thickness estimates as explanatory variables. This is of course appropriate for biomarker studies at which the authors aim. However, the reverse scenario with thickness estimates as outcome measures is equally frequent, and the relevance of the proposed criterion for such a situation is not immediately evident from the manuscript.

For these reasons, my suggestion would be to rather not employ the variance ratio as the primary and general criterion for evaluation. Instead, the three criteria of 1. reliability, 2. sensitivity to longitudinal change, and 3. predictive/discriminative power should be given the same importance, and should be separately demonstrated for the proposed method.

In fact, further evaluation of the methods is already offered by using cortical atrophy rates to predict age and MMSE scores. These are certainly interesting variables in this application context, but why don't the authors go for the main criterion, which is to distinguish between

patient (sub-) groups, instead of using these two proxies? This is somewhat puzzling, since it is expressly and repeatedly stated that this would be the intended use of the proposed method.

We agree with the Reviewer. Upon receipt of this round of reviews, the co-authors performed a re-assessment of the set of included experiments and determined that:

1) *The regression/prediction experiments described by Equations (3) and (4) were not very interesting nor were the findings compelling. Not only did they not elucidate performance capabilities across pipelines but, as mentioned by the reviewer, model assumptions were problematic. This took us away from the primary motivation for the manuscript which is to introduce the ANTs longitudinal cortical thickness pipeline and motivate its use with the ADNI data.*
2) *Perhaps motivated by the rescaling and and reconfiguring of Tables 2 & 3, as well as the comments from the reviewers, we felt that we gave short shrift to the experiments involving discernment of diagnostic groups. Using packages, such as FreeSurfer and ANTs, for identifying imaging biomarkers based on brain atrophy quantification, is of significant research interest.*

This reassessment led us to drop the prediction/regression experiments and opt for an expansion of illustrating performance differences based on regional diagnostic contrasts (see Sections 2.6 and 3.2).

2.6 Regional diagnostic contrasts based on cortical atrophy

The variance ratio explored in the previous section is a generic desideratum for statistical assessment of performance over the set of possible use cases. In this section, we narrow the focus to the unique demographical characteristics of the ADNI-1 study data and look at performance of the various pipelines in distinguishing between diagnostic groups on a region-by-region basis. Previous work has explored various aspects of Alzheimer's disease with respect to its spatial distribution and the regional onset of cerebral atrophy. For example, although much work points to the entorhinal cortex as the site for initial deposition of amyloid and tau [77], other evidence points to the basal forebrain as preceding cortical spread [78]. Other considerations include the use of hippocampal atrophy rates as an image-based biomarker of cognitive decline [79], differentiation from other dementia manifestations (e.g., posterior cortical atrophy [80]), and the use of FreeSurfer for monitoring disease progression [81]. Thus, longitudinal measurements have immediate application in Alzheimer's disease research. To showcase the utility of the ANTs framework, we compare the generated longitudinal measurements and their ability to differentiate diagnostic groups (i.e., CN vs. LMCI vs. AD).

Pipeline-specific LME models were constructed for each DKT region relating the change in cortical thickness to diagnosis and other pertinent covariates. In the notation of [82], these regional LME models are defined as:

$$\Delta Y^k \sim Y_{bl}^{\ k} + AGE_{bl} + ICV_{bl} + APOE + GENDER$$

+ DIAGNOSIS + VISIT : DIAGNOSIS + (1|ID) + (1|SITE) (3)

where ΔY k is the change in thickness of the kth DKT region from baseline (bl) thickness measurement Y k . We also include random intercepts for both the individual subject (ID) and the acquisition site. Modeling was performed in R using the lme4 package [83] followed by Tukey post-hoc analyses to test the significance of the "LMCI−CN", "AD−LMCI", and "AD−CN" diagnostic contrasts. Tables 2 and 3 provide the 95% confidence intervals where the cell color denotes the log-scaled adjusted region-specific *p*-values.

and it's Results counterpart:

3.2 Regional diagnostic contrasts based on cortical atrophy

Table 2: 95% confidence intervals for the diagnostic contrasts (LMCI-CN, AD-LMCI, AD-CN) of the ADNI-1 data set for each DKT region of the left hemisphere. Each cell is color-coded based on the adjusted log-scaled p-value significance from dark orange (p < 1e-10) to yellow (p = 0.1). Absence of color denotes nonsignificance.

	LMCI−CN					AD−LMCI					AD−CN				
DKT	FSCross	FSLong	ANTsCross	ANTsNative	ANTsSST	FSCross	FSLong	ANTsCross	ANTsNative	ANTsSST	FSCross	FSLong	ANTsCross	ANTsNative	ANTsSST
lcACC	-0.03,0.068	-0.038,0.058	-0.132,0.024	-0.235,-0.065	-0.238,-0.065	-0.075,0.033	-0.08,0.024	-0.103,0.068	-0.187,-0.001	-0.211,-0.022	-0.063,0.058	-0.077,0.041	-0.168,0.024	-0.349,-0.14	-0.375,-0.162
lcMFG	-0.111,-0.043	-0.11,-0.04	-0.197,-0.067	-0.188,-0.063	-0.188,-0.061	-0.109,-0.034	-0.13,-0.053	-0.201,-0.058	-0.21,-0.073	-0.201,-0.062	-0.19,-0.106	-0.21,-0.124	-0.342,-0.182	-0.345,-0.19	-0.334,-0.178
lCUN	-0.042,0.004	-0.046,0.002	-0.049,0.049	-0.073,0.038	-0.08,0.03	-0.03,0.021	-0.034,0.019	-0.108,0	-0.138,-0.016	-0.145,-0.024	-0.053,0.005	-0.059,0	-0.115,0.006	-0.164,-0.026	-0.177,-0.041
lENT	-0.404,-0.24	-0.405,-0.24	-0.285,-0.054	-0.479,-0.219	-0.486,-0.223	-0.359,-0.179	-0.385,-0.204	-0.354,-0.1	-0.462,-0.177	-0.514,-0.226	-0.692,-0.489	-0.719,-0.515	-0.539,-0.254	-0.83,-0.509	-0.887,-0.562
lFUS	-0.129,-0.059	-0.126,-0.054	-0.196,-0.029	-0.308,-0.116	-0.321,-0.128	-0.149,-0.073	-0.161,-0.082	-0.302,-0.117	-0.398,-0.187	-0.418,-0.207	-0.248,-0.162	-0.255,-0.167	-0.425,-0.218	-0.623,-0.385	-0.656,-0.418
lIPL	-0.099,-0.037	-0.099,-0.035	-0.245,-0.069	-0.245,-0.07	-0.246,-0.067	-0.138,-0.07	-0.144,-0.074	-0.357,-0.163	-0.349,-0.157	-0.344,-0.148	-0.21,-0.134	-0.215,-0.136	-0.526,-0.308	-0.52,-0.302	-0.514,-0.292
lITG	-0.146,-0.07	-0.148,-0.072	-0.225,-0.012	-0.359,-0.132	-0.406,-0.165	-0.165,-0.082	-0.17,-0.086	-0.454,-0.218	-0.535,-0.286	-0.574,-0.31	-0.279,-0.185	-0.285,-0.19	-0.586,-0.322	-0.797,-0.516	-0.876,-0.579
liCC	-0.091,-0.02	-0.096,-0.02	-0.17,-0.025	-0.252,-0.089	-0.255,-0.089	-0.117,-0.039	-0.137,-0.054	-0.231,-0.071	-0.307,-0.129	-0.328,-0.147	-0.177,-0.09	-0.2,-0.107	-0.338,-0.159	-0.489,-0.288	-0.512,-0.307
lLOG	-0.065,-0.013	-0.062,-0.01	-0.172,-0.017	-0.18,-0.027	-0.176,-0.021	-0.07,-0.013	-0.074,-0.017	-0.262,-0.091	-0.285,-0.119	-0.304,-0.134	-0.112,-0.049	-0.113,-0.049	-0.367,-0.175	-0.399,-0.211	-0.413,-0.222
lLOF	-0.062,0	-0.071,-0.011	-0.153,0	-0.23,-0.068	-0.236,-0.069	-0.078,-0.01	-0.086,-0.02	-0.217,-0.048	-0.279,-0.101	-0.31,-0.127	-0.114,-0.037	-0.131,-0.057	-0.303,-0.114	-0.439,-0.239	-0.474,-0.269
lLING	-0.041,0.005	-0.041,0.004	-0.097,0.033	-0.148,-0.003	-0.157,-0.015	-0.065,-0.015	-0.07,-0.02	-0.141,0.003	-0.184,-0.024	-0.19,-0.035	-0.086,-0.03	-0.091,-0.035	-0.181,-0.02	-0.269,-0.09	-0.285,-0.111
lMOF	-0.079,-0.011	-0.095,-0.03	-0.181,-0.004	-0.294,-0.114	-0.303,-0.114	-0.08,-0.005	-0.086,-0.015	-0.255,-0.059	-0.326,-0.128	-0.378,-0.171	-0.13,-0.046	-0.153,-0.073	-0.359,-0.139	-0.542,-0.32	-0.599,-0.367
lMGH	-0.147,-0.076	-0.137,-0.066	-0.239,-0.06	-0.32,-0.142	-0.35,-0.161	-0.166,-0.088	-0.167,-0.089	-0.361,-0.162	-0.442,-0.247	-0.471,-0.265	-0.283,-0.194	-0.274,-0.186	-0.523,-0.3	-0.686,-0.465	-0.74,-0.506
lPARH	-0.151,-0.044	-0.137,-0.036	-0.189,-0.022	-0.26,-0.074	-0.265,-0.077	-0.189,-0.071	-0.182,-0.07	-0.301,-0.116	-0.355,-0.152	-0.37,-0.164	-0.294,-0.161	-0.275,-0.15	-0.417,-0.211	-0.536,-0.306	-0.554,-0.322
lparaC	-0.075,-0.011	-0.088,-0.021	-0.075,0.024	-0.087,0.027	-0.087,0.027	-0.037,0.034	-0.068,0.006	-0.111,-0.002	-0.114,0.011	-0.122,0.003	-0.084,-0.005	-0.127,-0.044	-0.143,-0.021	-0.152,-0.011	-0.16,-0.019
lpOPER	-0.07,-0.007	-0.074,-0.011	-0.188,-0.065	-0.222,-0.099	-0.211,-0.091	-0.072,-0.002	-0.08,-0.01	-0.212,-0.076	-0.218,-0.084	-0.212,-0.08	-0.115,-0.037	-0.127,-0.048	-0.346,-0.194	-0.387,-0.236	-0.371,-0.223
lpORB	-0.105,-0.031	-0.099,-0.026	-0.163,0.007	-0.203,-0.041	-0.205,-0.042	-0.053,0.028	-0.067,0.014	-0.225,-0.037	-0.244,-0.066	-0.267,-0.088	-0.126,-0.034	-0.134,-0.044	-0.314,-0.104	-0.377,-0.176	-0.402,-0.2
lpTRI	-0.094,-0.035	-0.099,-0.038	-0.183,-0.029	-0.209,-0.066	-0.191,-0.046	-0.073,-0.008	-0.083,-0.016	-0.217,-0.047	-0.229,-0.072	-0.24,-0.081	-0.142,-0.068	-0.155,-0.08	-0.333,-0.143	-0.377,-0.2	-0.369,-0.189
lperiCAL	-0.019,0.021	-0.031,0.016	-0.06,0.065	-0.079,0.058	-0.095,0.042	-0.028,0.017	-0.044,0.008	-0.168,-0.03	-0.19,-0.04	-0.207,-0.057	-0.029,0.021	-0.055,0.003	-0.174,-0.019	-0.211,-0.041	-0.243,-0.074
lpostC	-0.05,-0.001	-0.052,-0.001	-0.12,-0.03	-0.109,-0.012	-0.11,-0.009	-0.051,0.004	-0.062,-0.005	-0.13,-0.03	-0.14,-0.033	-0.143,-0.031	-0.08,-0.019	-0.092,-0.028	-0.211,-0.1	-0.207,-0.087	-0.209,-0.084
lPCC	-0.058,0.004	-0.068,-0.003	-0.157,-0.026	-0.258,-0.103	-0.267,-0.104	-0.075,-0.006	-0.084,-0.013	-0.177,-0.033	-0.244,-0.074	-0.272,-0.093	-0.106,-0.029	-0.124,-0.044	-0.278,-0.116	-0.436,-0.244	-0.469,-0.267
lPreC	-0.092,-0.023	-0.101,-0.03	-0.145,-0.055	-0.138,-0.036	-0.135,-0.027	-0.077,-0.001	-0.093,-0.013	-0.141,-0.042	-0.145,-0.032	-0.139,-0.021	-0.139,-0.054	-0.162,-0.074	-0.247,-0.136	-0.239,-0.112	-0.228,-0.095
lPCUN	-0.099,-0.042	-0.105,-0.045	-0.17,-0.041	-0.199,-0.05	-0.197,-0.045	-0.091,-0.028	-0.112,-0.046	-0.208,-0.067	-0.241,-0.078	-0.245,-0.08	-0.165,-0.095	-0.191,-0.117	-0.323,-0.164	-0.376,-0.193	-0.378,-0.19
lrACC	-0.088,0.001	-0.095,-0.011	-0.133,0.05	-0.253,-0.05	-0.249,-0.042	-0.054,0.044	-0.054,0.039	-0.151,0.05	-0.266,-0.044	-0.308,-0.083	-0.103,0.007	-0.113,-0.008	-0.205,0.021	-0.431,-0.182	-0.468,-0.214
lrMFG	-0.087,-0.032	-0.095,-0.039	-0.247,-0.054	-0.293,-0.115	-0.297,-0.108	-0.087,-0.027	-0.089,-0.027	-0.292,-0.078	-0.302,-0.106	-0.332,-0.125	-0.151,-0.083	-0.16,-0.091	-0.455,-0.216	-0.519,-0.298	-0.548,-0.314
lSFG	-0.107,-0.049	-0.112,-0.053	-0.197,-0.076	-0.218,-0.093	-0.215,-0.088	-0.086,-0.023	-0.099,-0.033	-0.202,-0.069	-0.234,-0.098	-0.242,-0.103	-0.168,-0.097	-0.186,-0.112	-0.347,-0.197	-0.398,-0.244	-0.403,-0.246
lSPL	-0.086,-0.026	-0.087,-0.023	-0.143,-0.024	-0.105,0.016	-0.098,0.024	-0.08,-0.014	-0.097,-0.026	-0.164,-0.033	-0.146,-0.014	-0.141,-0.008	-0.14,-0.066	-0.156,-0.077	-0.256,-0.109	-0.199,-0.05	-0.187,-0.037
lSTG	-0.137,-0.069	-0.132,-0.066	-0.201,-0.074	-0.228,-0.105	-0.23,-0.105	-0.132,-0.057	-0.133,-0.06	-0.244,-0.104	-0.289,-0.153	-0.297,-0.161	-0.239,-0.155	-0.237,-0.155	-0.39,-0.233	-0.464,-0.312	-0.474,-0.319
lSMAR	-0.09,-0.026	-0.091,-0.027	-0.209,-0.074	-0.194,-0.064	-0.195,-0.058	-0.109,-0.038	-0.122,-0.051	-0.234,-0.086	-0.236,-0.094	-0.24,-0.09	-0.171,-0.092	-0.185,-0.106	-0.385,-0.219	-0.374,-0.214	-0.376,-0.208
lTT	-0.09,0.003	-0.088,0.001	-0.112,-0.002	-0.118,-0.023	-0.103,-0.014	-0.082,0.02	-0.091,0.007	-0.16,-0.038	-0.142,-0.037	-0.137,-0.04	-0.132,-0.016	-0.141,-0.031	-0.224,-0.088	-0.218,-0.101	-0.201,-0.092
lINS	-0.097,-0.023	-0.098,-0.023	-0.208,-0.045	-0.275,-0.108	-0.274,-0.109	-0.089,-0.007	-0.102,-0.02	-0.248,-0.069	-0.349,-0.165	-0.351,-0.171	-0.153,-0.062	-0.167,-0.075	-0.386,-0.185	-0.552,-0.345	-0.554,-0.351

Table 3: 95% confidence intervals for the diagnostic contrasts (LMCI-CN, AD-LMCI, AD-CN) of the ADNI-1 data set for each DKT region of the right hemisphere. Each cell is color-coded based on the adjusted log-scaled p-value significance from dark orange (p < 1e-10) to yellow (p = 0.1). Absence of color denotes nonsignificance.

DKT	LMCI−CN					AD−LMCI					AD−CN				
	FSCross	FSLong	ANTsCross	ANTsNative	ANTsSST	FSCross	FSLong	ANTsCross	ANTsNative	ANTsSST	FSCross	FSLong	ANTsCross	ANTsNative	ANTsSST
rcACC	-0.058,0.024	-0.048,0.028	-0.138,0.004	-0.214,-0.066	-0.222,-0.072	-0.088,0.002	-0.076,0.008	-0.141,0.015	-0.203,-0.041	-0.214,-0.05	-0.11,-0.009	-0.091,0.003	-0.217,-0.043	-0.353,-0.171	-0.371,-0.187
rcMFG	-0.117,-0.048	-0.114,-0.042	-0.184,-0.059	-0.178,-0.053	-0.178,-0.05	-0.118,-0.041	-0.14,-0.061	-0.248,-0.11	-0.251,-0.114	-0.241,-0.101	-0.205,-0.119	-0.223,-0.134	-0.377,-0.223	-0.376,-0.221	-0.364,-0.206
rCUN	-0.025,0.022	-0.03,0.019	-0.059,0.042	-0.085,0.032	-0.094,0.021	-0.057,-0.006	-0.062,-0.008	-0.12,-0.008	-0.133,-0.004	-0.14,-0.015	-0.062,-0.004	-0.071,-0.01	-0.135,-0.01	-0.167,-0.023	-0.184,-0.044
rENT	-0.408,-0.254	-0.403,-0.247	-0.271,-0.026	-0.446,-0.201	-0.438,-0.184	-0.331,-0.162	-0.356,-0.187	-0.407,-0.136	-0.511,-0.242	-0.561,-0.282	-0.672,-0.482	-0.692,-0.5	-0.571,-0.269	-0.852,-0.548	-0.89,-0.576
rFUS	-0.114,-0.045	-0.114,-0.043	-0.221,-0.046	-0.337,-0.144	-0.343,-0.15	-0.149,-0.073	-0.165,-0.086	-0.347,-0.154	-0.44,-0.228	-0.462,-0.252	-0.233,-0.148	-0.249,-0.16	-0.492,-0.276	-0.694,-0.455	-0.722,-0.485
rIPL	-0.107,-0.042	-0.113,-0.045	-0.201,-0.025	-0.228,-0.057	-0.229,-0.051	-0.15,-0.079	-0.16,-0.086	-0.366,-0.172	-0.356,-0.169	-0.358,-0.164	-0.229,-0.149	-0.243,-0.16	-0.491,-0.274	-0.51,-0.299	-0.512,-0.291
rITG	-0.162,-0.086	-0.14,-0.065	-0.281,-0.064	-0.408,-0.19	-0.44,-0.208	-0.179,-0.096	-0.193,-0.111	-0.432,-0.191	-0.539,-0.299	-0.591,-0.337	-0.308,-0.215	-0.301,-0.208	-0.618,-0.35	-0.853,-0.583	-0.931,-0.645
riCC	-0.08,-0.007	-0.084,-0.005	-0.182,-0.044	-0.256,-0.097	-0.26,-0.099	-0.116,-0.035	-0.139,-0.053	-0.239,-0.087	-0.319,-0.145	-0.335,-0.16	-0.164,-0.073	-0.189,-0.092	-0.361,-0.19	-0.507,-0.31	-0.526,-0.328
rLOG	-0.055,-0.004	-0.055,-0.004	-0.152,0.016	-0.179,-0.018	-0.184,-0.017	-0.073,-0.018	-0.08,-0.024	-0.31,-0.124	-0.325,-0.148	-0.337,-0.155	-0.106,-0.044	-0.113,-0.05	-0.389,-0.181	-0.434,-0.235	-0.449,-0.244
rLOF	-0.081,-0.018	-0.08,-0.022	-0.11,0.036	-0.192,-0.031	-0.193,-0.026	-0.077,-0.008	-0.09,-0.025	-0.195,-0.035	-0.255,-0.078	-0.286,-0.102	-0.13,-0.053	-0.145,-0.072	-0.241,-0.063	-0.377,-0.179	-0.406,-0.201
rLING	-0.05,-0.006	-0.055,-0.01	-0.087,0.05	-0.153,0.001	-0.16,-0.011	-0.063,-0.014	-0.067,-0.017	-0.14,0.012	-0.18,-0.011	-0.196,-0.033	-0.094,-0.039	-0.103,-0.046	-0.168,0.002	-0.266,-0.077	-0.292,-0.109
rMOF	-0.077,-0.012	-0.08,-0.017	-0.158,0.019	-0.258,-0.077	-0.268,-0.076	-0.088,-0.016	-0.096,-0.027	-0.25,-0.054	-0.318,-0.118	-0.37,-0.16	-0.137,-0.057	-0.149,-0.071	-0.331,-0.112	-0.498,-0.273	-0.555,-0.318
rMGH	-0.162,-0.093	-0.153,-0.084	-0.242,-0.044	-0.331,-0.139	-0.36,-0.149	-0.167,-0.09	-0.17,-0.094	-0.399,-0.18	-0.496,-0.286	-0.526,-0.296	-0.299,-0.213	-0.293,-0.208	-0.555,-0.31	-0.745,-0.507	-0.796,-0.535
rPARH	-0.193,-0.068	-0.184,-0.066	-0.178,-0.005	-0.277,-0.09	-0.267,-0.082	-0.242,-0.106	-0.23,-0.102	-0.324,-0.134	-0.386,-0.182	-0.401,-0.199	-0.382,-0.228	-0.364,-0.219	-0.427,-0.214	-0.583,-0.352	-0.589,-0.36
rparaC	-0.073,-0.006	-0.091,-0.02	-0.091,0.01	-0.093,0.022	-0.094,0.023	-0.051,0.023	-0.075,0.005	-0.101,0.009	-0.111,0.016	-0.117,0.011	-0.095,-0.012	-0.135,-0.046	-0.149,-0.024	-0.154,-0.011	-0.16,-0.016
rpOPER	-0.078,-0.015	-0.078,-0.016	-0.175,-0.05	-0.195,-0.071	-0.192,-0.07	-0.07,0	-0.084,-0.015	-0.188,-0.05	-0.219,-0.083	-0.217,-0.084	-0.12,-0.042	-0.135,-0.058	-0.309,-0.155	-0.36,-0.208	-0.356,-0.207
rpORB	-0.068,0.008	-0.071,0.001	-0.159,-0.006	-0.174,-0.028	-0.149,-0.002	-0.097,-0.013	-0.105,-0.026	-0.237,-0.067	-0.263,-0.103	-0.274,-0.113	-0.132,-0.038	-0.145,-0.056	-0.33,-0.14	-0.373,-0.193	-0.359,-0.179
rpTRI	-0.085,-0.025	-0.083,-0.022	-0.19,-0.044	-0.191,-0.057	-0.184,-0.049	-0.069,-0.002	-0.083,-0.016	-0.232,-0.072	-0.238,-0.092	-0.25,-0.103	-0.128,-0.053	-0.14,-0.064	-0.359,-0.179	-0.372,-0.207	-0.376,-0.21
rperiCAL	-0.022,0.02	-0.035,0.011	-0.052,0.065	-0.09,0.048	-0.11,0.027	-0.032,0.014	-0.046,0.005	-0.124,0.005	-0.16,-0.009	-0.18,-0.03	-0.036,0.016	-0.061,-0.004	-0.125,0.019	-0.19,-0.02	-0.231,-0.062
rpostC	-0.058,-0.009	-0.06,-0.01	-0.118,-0.028	-0.106,-0.007	-0.106,0.002	-0.072,-0.018	-0.079,-0.024	-0.131,-0.032	-0.15,-0.042	-0.165,-0.047	-0.109,-0.048	-0.117,-0.055	-0.21,-0.099	-0.214,-0.092	-0.224,-0.092
rPCC	-0.055,0.005	-0.068,-0.007	-0.153,-0.03	-0.233,-0.085	-0.243,-0.09	-0.068,-0.002	-0.081,-0.013	-0.135,0	-0.202,-0.04	-0.225,-0.058	-0.097,-0.022	-0.123,-0.047	-0.235,-0.083	-0.372,-0.189	-0.402,-0.213
rPreC	-0.094,-0.025	-0.1,-0.029	-0.12,-0.032	-0.113,-0.011	-0.11,0	-0.08,-0.003	-0.1,-0.021	-0.165,-0.068	-0.186,-0.073	-0.196,-0.075	-0.144,-0.057	-0.169,-0.081	-0.247,-0.138	-0.254,-0.128	-0.258,-0.122
rPCUN	-0.102,-0.038	-0.111,-0.044	-0.184,-0.042	-0.21,-0.048	-0.206,-0.042	-0.096,-0.026	-0.119,-0.045	-0.239,-0.084	-0.273,-0.097	-0.285,-0.107	-0.17,-0.092	-0.201,-0.118	-0.362,-0.187	-0.414,-0.214	-0.421,-0.219
rrACC	-0.062,0.024	-0.069,0.012	-0.21,-0.035	-0.336,-0.137	-0.352,-0.149	-0.096,-0.001	-0.096,-0.006	-0.231,-0.039	-0.324,-0.106	-0.362,-0.14	-0.121,-0.014	-0.13,-0.029	-0.365,-0.149	-0.574,-0.329	-0.627,-0.376
rrMFG	-0.096,-0.038	-0.097,-0.037	-0.203,-0.005	-0.238,-0.059	-0.237,-0.049	-0.09,-0.025	-0.1,-0.033	-0.373,-0.154	-0.359,-0.163	-0.381,-0.176	-0.161,-0.089	-0.171,-0.096	-0.49,-0.245	-0.52,-0.299	-0.538,-0.306
rSFG	-0.102,-0.041	-0.107,-0.043	-0.185,-0.062	-0.203,-0.079	-0.198,-0.073	-0.101,-0.033	-0.115,-0.044	-0.217,-0.081	-0.231,-0.095	-0.242,-0.104	-0.177,-0.101	-0.195,-0.115	-0.348,-0.196	-0.38,-0.227	-0.386,-0.231
rSPL	-0.093,-0.031	-0.098,-0.032	-0.141,-0.011	-0.116,0.02	-0.105,0.034	-0.093,-0.024	-0.106,-0.034	-0.207,-0.064	-0.18,-0.03	-0.171,-0.019	-0.158,-0.082	-0.175,-0.094	-0.291,-0.131	-0.237,-0.069	-0.216,-0.045
rSTG	-0.139,-0.071	-0.131,-0.065	-0.184,-0.04	-0.22,-0.087	-0.232,-0.091	-0.132,-0.056	-0.137,-0.064	-0.269,-0.111	-0.295,-0.15	-0.31,-0.155	-0.241,-0.157	-0.24,-0.158	-0.391,-0.213	-0.458,-0.294	-0.482,-0.307
rSMAR	-0.111,-0.049	-0.11,-0.048	-0.214,-0.069	-0.195,-0.061	-0.186,-0.041	-0.117,-0.049	-0.136,-0.067	-0.248,-0.089	-0.27,-0.123	-0.289,-0.131	-0.201,-0.124	-0.22,-0.142	-0.399,-0.221	-0.407,-0.241	-0.412,-0.234
rTT	-0.105,-0.017	-0.1,-0.015	-0.107,0.013	-0.11,-0.006	-0.108,-0.008	-0.105,-0.007	-0.104,-0.01	-0.225,-0.09	-0.192,-0.077	-0.169,-0.06	-0.171,-0.062	-0.167,-0.062	-0.28,-0.129	-0.258,-0.128	-0.235,-0.111
rINS	-0.079,-0.006	-0.083,-0.014	-0.212,-0.051	-0.271,-0.106	-0.272,-0.109	-0.094,-0.014	-0.098,-0.022	-0.22,-0.043	-0.32,-0.139	-0.323,-0.145	-0.141,-0.052	-0.152,-0.066	-0.362,-0.164	-0.52,-0.316	-0.525,-0.324

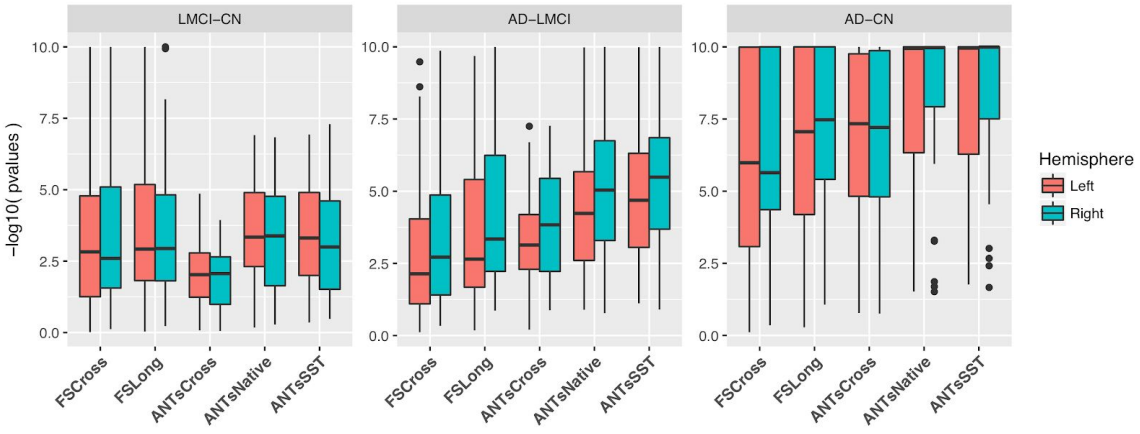


Figure 9: Log-scaled p-values summarizing Tables 2 and 3 demonstrating performance differences across cross-sectional and longitudinal pipelines for the three diagnostic contrasts.

The LME model described in Equation (3) was used to determine region-by-region contrasts for each pairing “LMCI−CN”, “AD−LMCI”, and “AD−CN” using post-hoc Tukey significance testing. It should be noted that no subjects were included that switched diagnostic groups during the acquired study schedule. These findings are provided in Tables 2 and 3. The adjusted p-values were log- scaled for use in specifying the individual color cell for facilitating visual differentiation. Each cell contains the corresponding 95% confidence intervals. Figure 9 provides a side-by-side

comparison of the distribution of log-scaled p-values separated into left and right hemispherical components and grouped according to contrast.

Consideration of performance over all three diagnostic pairings illustrates the superiority of the longitudinal ANTs methodologies over their ANTs cross-sectional counterpart. Several regions demonstrating statistically significant non-zero atrophy in ANTsNative and ANTsSST do not manifest similar trends in ANTsCross (e.g., LMCI-CN: lateral occipital gyri). Pronounced differences between the ANTs longitudinal vs. cross-sectional methodologies can be seen in both the LMCI−CN and AD−CN contrasts. Although ANTsCross demonstrates discriminative capabilities throughout the cortex and, specifically, in certain AD salient regions, such as the entorhinal and parahippocampal cortices, the contrast is not nearly as strong as the other methods including FSCross and FSLong thus motivating the use of longitudinal considerations for processing of AD data.

Differentiation between the longitudinal methods is not as obvious although trends certainly exist. In general, for differentiating CN vs. LMCI, all methods are comparable except for ANTsCross. How- ever, for the other two diagnostic contrasts ”AD−LMCI” and ”AD−CN”, the trend is similar to what we found in the evaluation via the variance ratio, viz., the longitudinal ANTs methods tend towards greater contrast means versus ANTsCross and the two FreeSurfer methods. Looking at specific cortical areas, though, we see that comparable regions (“comparable” in terms of variance ratio) are consistent with previous findings. For example, we noted in the last section that FSLong has a relatively large variance ratio in the entorhinal regions which is consistent with the results seen in Tables 2 and 3.

In addition to these two major points, I currently have two minor issues:

1. Introduction and / or discussion:

Please state what exactly is the motivation of this work,

In practical terms, the ANTs longitudinal cortical thickness pipeline has been publicly available for some time. People (including us) are using it to do research and we would like to have a reference where we have formalized exploration of performance. We added the following sentence to the Discussion:

This framework has been publicly available as open-source in the ANTs GitHub repository for some time. It has been employed in various neuroimaging studies and this work constitutes a formalized exploration of performance for future reference.

and what the innovation of the proposed methods is.

Regarding innovation---we would simply point to the heretofore nonexistence (in the reported literature) of a longitudinal framework employing a registration-based approach

for measuring cortical thickness in longitudinal data. This should be clear from the abstract where we write

> In this work, we introduce the open-source Advanced Normalization Tools (ANTs) registration-based cortical thickness longitudinal processing pipeline

We also write towards the end of the Introduction

> In this work, we introduce the longitudinal version of the ANTs registration-based cortical thickness pipeline and demonstrate its utility on the publicly available ADNI-1 data set.

Advantages of the proposed method are already mentioned, but is it exactly those features that would lead to the observed performance improvements? That is, I'd be curious how one might explain the observed gain in performance.

If the Reviewer is referring to performance improvements over FreeSurfer, then we would have to defer definitively answering such questions to potential future investigations. FreeSurfer and ANTs pipelines are sophisticated software workflows composed of many parts which approach quantifying thickness in fundamentally different ways. We could speculate but that would not be appropriate for the manuscript.

If the Reviewer is referring to the performance improvement of ANTs Native/SST over ANTs Cross, then we would attributed performance differences to the use of the subject-specific template and corresponding construction of segmentation tissue priors as that is the fundamental difference between the longitudinal pipelines and their cross-sectional counterpart.

2. Methods section:

Given that image denoising can have a substantial impact on which features remain in an image, I think that previous reviewer 2's request about further details has not sufficiently been addressed.

It is good that the authors "have examined the output and made certain that using the denoising algorithm did not produce undesirable results", but it would be even better if it is made transparent to the reader how the denoising affected the images.

In particular, the addition to the manuscript: "Based on outcomes involving previously processed data sets (including ADNI-2), we chose to employ the denoising algorithm [45] for all ANTs-based processing" is too general and therefore hard to comprehend for a first-time reader.

We removed the sentence in question and added Figure 5 (reproduced below) with the following text:

An ANTs implementation of the denoising algorithm of [43] is a recent addition to the toolkit and has been added as an option to both the cross-sectional and longitudinal pipelines. This denoising algorithm employs a non-local means filter [58] to account for the spatial varying noise in MR images in addition to consideration of the inherent Rician noise noise inherent to MRI [59]. This preprocessing step has been used in a variety of imaging studies for enhancing segmentation-based protocols including hippocampal and ventricle segmentation [60], voxel-based morphometry in cannabis users [61], and anterior temporal lobe gray matter volume in bilingual adults [62]. An illustration of the transformed data resulting from image preprocessing (bias correction plus denoising) is provided in Figure 5.

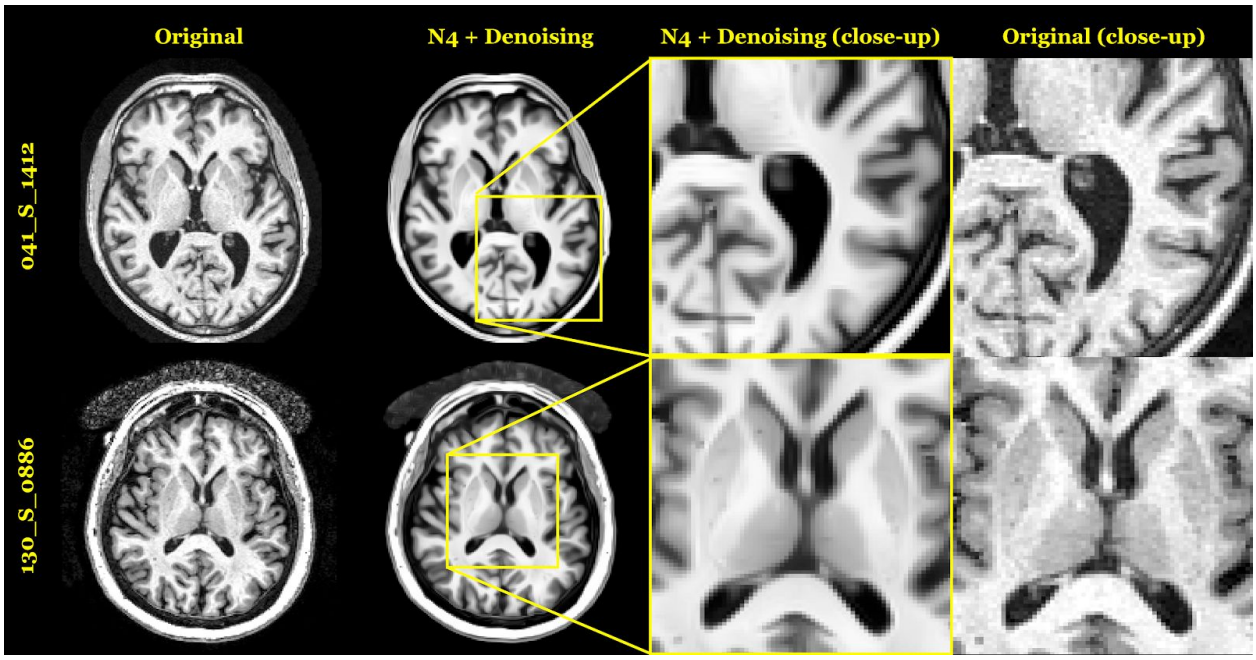


Figure 5: Images from two randomly chosen subjects were chosen to illustrate the effects of bias correction and denoising. The former mitigates artificial low spatial frequency modulation of intensities whereas the latter reduces the high frequency spatially-varying Rician noise characteristic of MRI.

Overall recommendation:

I do not see any major issues with the proposed method itself, but rather with its evaluation, which has already been the basic tenet of the previous reviews. The editor's previous request that the authors "carefully address these concerns, and especially offer additional experiments (including showing group differentiation using these longitudinal estimates)" has been addressed in this revision, but in my eyes not in sufficient scope and detail. This is why I recommend further revision of the manuscript.

We agree with the Reviewer and hope the above responses help alleviate the Reviewer's concerns regarding this revision.