Response to the reviewers

The reviewers saw merit in this work. However they raised strong concerns about the validation of the method, which is important, given that the paper has limited methodological innovation. The authors are encouraged to carefully address these concerns, and especially offer additional experiments (including showing group differentiation using these longitudinal estimates), and resubmit the paper. The authors should be advised that without the suggested in-depth revision, the paper is unlikely to reach acceptance levels.

We appreciate the time spent by the editors and reviewers in assessing our manuscript. Prior to the discussion of the issues raised by the reviewers, we would like to point out that we performed a complete reprocessing of the ANTs data after Round 1 of the review process. This was performed since the different ANTs pipelines were run on two different computational clusters (at the University of Virginia and the University of California, Irvine) over the course of approximately two years during which certain ANTs pipeline components had undergone significant changes. As we note in the revised manuscript, current results were obtained from data processing on a single cluster (UCI) with the most current ANTs software. Note that this does not change the main findings of the original paper but is intended to improve accuracy and reproducibility. Note that we also removed the Results subsection focusing on the entorhinal cortex which we felt was not sufficiently general. The reviewers' suggestions were certainly appropriate but we ultimately decided to go a different direction for supporting the use of the performance criterion described in the manuscript. Based on other recommendations by the reviewers, we opted instead for showcasing the distribution of cortical atrophy rates in the ADNI sub-populations and age/MMSE prediction assessments which provides supporting evidence for the use of the variance ratio as a criterion of performance.

Please see below for a point-by-point response to the issues raised.

Reviewer 1

This paper presents a longitudinal pipeline for processing and analyzing brain MR images and measuring cortical thickness. Overall, the methodological contribution is not large, although it has been validated on the large-scale ADNI dataset. The reviewer has several concerns.

1. For the longitudinal analysis of cortical thickness, it is important to show the trajectories of cortical thickness in individuals and in different groups and across different regions, and then compare with FreeSurfer. These are not provided in the paper.

   We added Section **3.1 Regional cortical atrophy rates based on diagnosis**. *This section explores the use of slope values describing cortical atrophy to separate diagnoses (CN vs. LMCI vs. AD) in the ADNI-1 cohort amongst the different processing streams. In particular, see Tables 2 and 3. Note that we opted for this format given the number of* *pipeline x region x diagnosis combinations. However, as we mention in the manuscript, spaghetti plots and slope distribution plots are provided in the github repository, specifically in the directory to linked to here:*

https://github.com/ntustison/CrossLong/tree/master/Data/

2. "For the ADNI-1 processing described in this work, we created a population-specific template from 52 cognitively normal ADNI-1 subjects." This will bias the analysis, since ADNI includes both normal and abnormal (AD and MCI) subjects.

We use the ADNI template in a similar manner that other studies use other standardized templates. Specifically, as explained in the manuscript, the ADNI template and corresponding prior probability images are mapped to each single-subject template for constructing n-tissue, single-subject priors. Such practice is widely used in the field and not considered to be a biased practice. For example, this paper

> Jamie L. Hanson , Nicole Hair, Dinggang G. Shen, Feng Shi, John H. Gilmore, Barbara L. Wolfe, Seth D. Pollak, “Family Poverty Affects the Rate of Human Infant Brain Growth”, PLoS ONE, 2013.

employs atlases built “from 95 ***normal*** *infants (56 males and 39 females) at neonate, 1-year-old, and 2-year-old” (emphasis added) described in this paper:*

> Feng Shi, Pew-Thian Yap, Guorong Wu, Hongjun Jia, John H. Gilmore, Weili Lin, Dinggang Shen,"Infant Brain Atlases from Neonates to 1- and 2-year-olds", PLoS ONE, 6(4): e18746, 2011.

to study brain differences based on economic factors. These infant atlases built from normal cohorts were also used to in a variety of studies investigating brain differences in non-normal populations including:

- *neonates at high risk for schizophrenia (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4452433/),*
- *in utero alcohol exposure (https://www.ncbi.nlm.nih.gov/pubmed/26616173),*
- *newborns with congenital heart disease (https://www.ncbi.nlm.nih.gov/pubmed/29034164), and*
- *infants with adverse perinatal conditions (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5143343/).*

Even closer to the current study is the FreeSurfer template (https://www.ncbi.nlm.nih.gov/pubmed/10619420) which does not appear to be generated from a number of MCI, AD, and cognitively normal subjects to prevent the *supposed bias claimed by the reviewer. Thus, we believe that our template construction and usage is consistent with standard practice in the field.*

3. Some related references on longitudinal surface reconstruction and cortical thickness measurement are not included and not discussed in the paper.

a. Xue Z, et al. CLASSIC: consistent longitudinal alignment and segmentation for serial image computing. Neuroimage. 2006; 30(2):388-99.

b. Li Y, et al. Consistent 4D cortical thickness measurement for longitudinal neuroimaging study. MICCAI. 2010; 13():133-42.

c. Nakamura K, et al. CLADA: cortical longitudinal atrophy detection algorithm. Neuroimage. 2011; 54(1):278-89.

d. Li G, et al. Consistent reconstruction of cortical surfaces from longitudinal brain MR images. Neuroimage. 2012; 59(4):3805-20.

e. Dai Y, et al. aBEAT: a toolbox for consistent analysis of longitudinal adult brain MRI. PLoS One. 2013; 8(4):e60344.

f. Wang L, et al. 4D segmentation of brain MR images with constrained cortical thickness variation. PLoS One. 2013; 8(7):e64207.

g. Li G, et al. Measuring the dynamic longitudinal cortex development in infants by reconstruction of temporally consistent cortical surfaces. Neuroimage 90, 266-279.

The reviewer is correct in that there are other important methods for characterizing change in cortical morphology. Due to constraints in space and methodological availability, we limited comparison to the longitudinal FreeSurfer stream. However, we did include a representative citation from the list above in the introduction so that the reader is aware that other longitudinal-specific methods exist and merit consideration. Specifically, we added the following to the **Introduction***:*

> Although the FreeSurfer longitudinal processing stream is perhaps one of the most well-known, other important longitudinal-specific methodologies have been proposed for characterizing cortical morphological change. Similar to FreeSurfer, cortical surfaces are generated in [24, 25] permitting vertex-wise quantitation of thickness and thickness change. Application to early infants in [24] further demonstrate the utility of targeted longitudinal considerations.

In addition, we included the following mention of longitudinal assessment strategies at the beginning of the **Statistical evaluation** *section***:**

> Similarly, evaluation strategies for longitudinal studies have been proposed with resemblance to those employed for cross-sectional data such as the use of visual assessment [24], scan-rescan data [12, 25], and 2-D comparisons of post mortem images and corresponding MRI [25]. In addition, longitudinal methods offer potential for other types of assessments such as the use of simulated data (e.g., atrophy [12, 25], infant

development [24]) where ”ground-truth” is known and, regression analysis of longitudinal trajectories of regional cortical thickness [60].

4. Although the proposed pipeline can measure cortical thickness, many important longitudinal cortical properties cannot be simply measured, e.g., surface area, cortical folding, and gyrification. Please discuss.

We added the following discussion of limitations to the **Discussion** *section:*

While we focus on cortical thickness in this work, there are obvious limitations with the ANTs volume-based framework. Without a direct reconstruction of the cortical surfaces, many important cortical properties (e.g., surface area, cortical folding, sulcal depth, and gyrification) [74] cannot be generated in a straightforward manner. Additional work will want to examine these features more closely working towards a more comprehensive idea of how structure changes. This will help determine the relative importance of such cortical features and will undoubtedly guide future methodological development.

5. Please discuss the limitations and applicability of the proposed method and whether it can be useful for brain development studies, e.g.,
a. Lyall AE, et al. Dynamic development of regional cortical thickness and surface area in early childhood. Cerebral cortex 25 (8), 2204-2212.
b. Li G, et al. Spatial patterns, longitudinal development, and hemispheric asymmetries of cortical thickness in infants from birth to 2 years of age. Journal of Neuroscience 35 (24), 9150-9162.

We agree that the proposed method has limitations (see Comment 4) but that it (as well as those methods which are listed above and those not listed) also has utility for brain development studies.

Reviewer 2

This paper describes a new longitudinal processing pipeline for registration based thickness estimation using the ANT's framework. Similar to existing longitudinal pipelines, authors describe the use of a subject specific template which is employed instead of the group template to generate the individual time-point cortical thickness maps. For ROI analysis, also regional labels are propagated using joint label fusion of an atlas set with cortical labels. For validation, a ratio of variances is chosen as the quality measure and compared with different versions of the pipeline and FreeSurfer. Furthermore, a cross-sectional group analysis of AD is performed in ADNI.

The paper is well written and cites the relevant work. I am glad to see that more effort is put into the development of methods that are dedicated to the analysis of longitudinal data. There are a few details that I think should be added into the method description, but my major concerns with this paper is the, in my opinion, incorrect evaluation:

We appreciate the reviewer's assessment and hope that the responses below alleviate the reviewer's concerns.

1. The ratio of variances is often looked at in ANOVAs where it can be very useful, but not here. It is neither good to increase between-subject-variance (as that can easily be achieved by adding noise to all measurements) nor is it beneficial to reduce within-subject-variance (which can be easily achieved by reporting the mean value for all time points or by restricting results to be nearly identical to the subject template estimates). As a consequence, a method that is very noisy when estimating thickness in the template, but very restrictive in allowing time points to vary from this estimate, will perform best. Neither is wanted. More details below.

The reviewer is correct. Estimates of between-subject and within-subject variance are not practically useful, when considered in isolation. *The thought experiments formulated by the reviewer make this quite clear. Fortunately, that is not what we used to evaluate the various longitudinal pipelines. It is the combination of the two measurements in the form of a ratio which provides the actual comparative utility. The hypothetical examples provided where one simply adds measurement error and/or subset selection of results to homogenous populations are clearly not what is intended in assessing the pipeline estimates. For a longitudinal biomarker to be effective at classifying subpopulations, it should have low within-subject variation and high between-subject variation. Without this, subpopulation distinctions would not be possible (eg. if measurements within the subject vary more than those between subjects). Thus we continue to contend that this is a valuable assessment tool.*

A rigorous statistical argument for the proposed approach may be found in 'Linear Regression Analysis' by Seber and Lee. We have added the reference and the following discussion to the Methods section.

> In particular, [66] (Sections 9.6.2 and 9.6.5) demonstrate the role that randomness and measurement error in explanatory variables play in statistical inference. When the explanatory variable is fixed but measured with error (as is plausible for cortical thickness measurements), the within-subject variance divided by the between subject variance is proportional to the bias of the estimated linear coefficient when the outcome of interest is regressed over the explanatory variable (Example 9.2). In short, the larger the rk, the less bias for future statistical analyses based upon the cortical thickness data. When the explanatory variable is considered random and is measured with error (a common assumption in the measurement error literature [67, 68], this bias is expressed as attenuation of regression coefficient estimates to zero by a multiplicative factor rk/(1 + rk) (Example 9.3). Thus, larger rk means less less attenuation bias and hence more discriminative capacity. Note that effect estimator bias is not the only problem— the residual variance is increased by a factor proportional to rk/(1 + rk) ([66], Chapter 3). The same authors refer to the combination of bias and added variance as a 'double whammy'. Indeed, a worse reliability ratio causes greater bias in multiple linear regression in the presence of collinearity and even biases the estimators for other covariates, progression through time included (cf [68],

Section 3.3.1). The same authors state that this bias is typical even in generalized linear models (Section 3.6) and continue to use the ratio as a measure of reliability even in the longitudinal context (Section 11.9).

2. The paper is missing any analysis of longitudinal data. For ADNI only a group analysis with AD is performed based on cross-sectional entorhinal cortex measurements. A longitudinal method needs to be evaluated with respect to its sensitivity to detect longitudinal change and distinguish groups based on atrophy rates not based on average thickness.

Please note the addition of Section **3.1 Regional cortical atrophy rates based on diagnosis**. *This section explores the use of slope values describing cortical atrophy to separate diagnoses (CN vs. LMCI vs. AD) in the ADNI-1 cohort amongst the different processing streams. In particular, see Tables 2 and 3. Note that we opted for this format given the number of pipeline x region x diagnosis combinations. However, as we mention in the manuscript, spaghetti plots and slope distribution plots are provided in the github repository, specifically in the directory to linked to here:*

https://github.com/ntustison/CrossLong/tree/master/Data/

We also added Age/MMSE prediction analyses summarized in Figures 8 and 9 for showing the utility of the variance ratio.

Required Major Changes:

- We recommend performing a reliability analysis based on test-retest data (with or without repositioning in the scanner). Here we assume no anatomical change and thus quantifying the within-subject variance is informative. ICC can be used to evaluate test-retest data, while it is not a suitable measure to evaluate the effect of interest in a longitudinal study (and criteria to evaluate longitudinal studies should not be motivated by referring to their relatedness to ICC). In longitudinal studies, we treat each subject as a "group" of several dependent measurements across time. For a simple longitudinal model with time as the only covariate, the ICC relates the between-subject and the total variance. A high ICC would indicate that measurements across time and within each subject are similar. This is a reasonable measure only in test-retest studies, where we do not expect change, but not in longitudinal studies in general, where we have reason to believe that there is change across time. A test-retest analysis, however, can only indicate reliability of a method, not sensitivity to change (in fact many approaches that incorporate within subject constraints will increase test-retest reliability at the cost of sensitivity to real anatomical changes).

We appreciate the reviewer's comments but contend that the ICC still has relevance in the current evaluation. The reviewer states that "[The ICC] a reasonable measure only in test-retest studies, where we do not expect change, but not in longitudinal studies in general, where we have reason to believe that there is change across time." This would be true if one *does not account for the mean change in time. However, as the reviewer noted, we have adjusted for time in the assessment models, thereby de-trending the data with respect to time. As such, we are considering each subject's residual variation around their fitted time trend and comparing this measure to the between subject variation in measures.*

Regarding the final sentence

- Furthermore, the paper is missing an evaluation of the sensitivity of the method with respect to the main output variable of a longitudinal study: the change of a structure. We recommend demonstrating improvements when distinguishing groups based on within-subject slopes (interaction of time and group). ADNI can be used for this and while large group effects (e.g. atrophy rates of the entorhinal cortex will differ between AD and controls) are a good start, also small group differences (e.g. controls vs MCI stable, or MCI-stable vs MCI-progressors) are also of interest here. This will demonstrate if the improved test-retest reliability was traded by sacrificing sensitivity to longitudinal changes.

> *Again, as mentioned above, please note the addition of Section* **3.1 Regional cortical atrophy rates based on diagnosis**.

Further Details:

- Within-subjects variance needs to be distinguished from residual variance, and a lower within-subjects variance does not necessarily imply a higher precision. What the authors call within-subjects variance should more precisely be called residual or error variance. It is only in simple ANOVA models for independent data that the within-group variance coincides with the residual variance. This is primarily a terminology problem.

> *We have modified the terminology of the paper, utilizing "residual variance" throughout.*

- Between-subjects variance needs to be distinguished from effects-of-interest variance (e.g. between-groups variance), and a greater between-subjects variance does not necessarily imply greater discriminative power to e.g. distinguish between groups. Two groups can more easily be distinguished if their means are far apart, i.e. if the between-groups variance is high. A high between-groups variance implies a high-between subjects variance, but the converse is not true. For example, a high between-subjects variance could also be due to a large age range in two clinical groups that are otherwise the same. We might say that high inter-subject variation is necessary, but not sufficient to detect group differences; and still, this is a cross-sectional effect, and hence not the main effect of interest in a longitudinal study. Longitudinal analyses, in contrast, are within-subjects designs, and the between subjects-variance is not relevant for assessing intra-individual change across time. In fact, we often remove the between subjects-variance, by analysing only the longitudinal changes. In that sense, longitudinal analysis will not benefit from increased inter-subject variance.

We, of course, agree with the reviewer that greater discriminatory power is dependent upon high between-group variation. This will of course depend upon the scientific comparisons being made and is not a function of the method but the biology and environment giving rise to the observations. Further, this statement is looking at one dimension of variation and we have purposely contrasted residual variation with between-subject variation as a general principle. The primary point of using this message is that to afford greatest discriminatory power for a group comparison (conditional upon what the true between-group variation is), one would like measurements to be reproducible (i.e., low residual variation) relative to observed between subject error.

Minor comments:

- Please make clear in the abstract and e.g. page 6 step 4 that the cortical thickness estimation is registration based.

Done. We added the qualifier "registration-based" in a number of places in the paper including the Abstract, Introduction, ANTs cortical Thickness, and Discussion sections.

- Make clear for what results you actually use the patch based denoising algorithm [45]. The problem with those algorithms is that they change the image significantly. While a denoised image looks very pleasing, lots of anatomical information gets destroyed and it is unclear where (locally) how much and if there is a bias with respect to the variable of interest (disease severity, drug level etc).

We have examined the output and made certain that using the denoising algorithm did not produce undesirable results. We added the following in the **Individual time point processing** *paragraph of the* **Unbiased longitudinal processing** *subsection,*

> Based on outcomes involving previously processed data sets (including ADNI-2), we chose to employ the denoising algorithm [45] for all ANTs-based processing.

- The first sentence in 2.3.1 is difficult to understand: "quantify … performance of … pipeline variants and a comparison with their … comparisons."

Done. The sentence was rewritten as follows:

> A summary measure related to the ICC statistic [63] is used to quantify the relative performance of these cross-sectional and longitudinal ANTs pipeline variants along with the cross-sectional and longitudinal FreeSurfer streams.

- Page 8 Fig3: single-single subject template (one single too much)

Done. We removed the redundancy.

- Specify what interpolation methods are used (tri-linear, cubic b-spline?). This is very important as cubic b-spline introduces far less interpolation artifacts.

> *Similar to the two FreeSurfer streams used in this manuscript, choice of interpolation method is an admittedly important but simply one of many parameter/component choices for a given ANTs pipeline. Other image registration choices include the number of scaling pyramid levels, the shrinking factors at each level, the Gaussian smoothing kernel size at each level, the number of gradient steps, the convergence criteria, whether to perform histogram matching and/or clamp voxels with outlier intensity values. And these are only the possible parameters for a single transform and must all be considered for both linear and non-linear registration "stages." A comparable number of component parameter choices exist for preprocessing components (e.g., N4 bias correction) segmentation, cortical thickness estimation, and joint label fusion and it is not obvious that any single parameter (e.g., interpolation method) is more important than any other. Needless to say it is beyond the scope of this manuscript to provide the information for each of these choices nor does it make sense to feature any single parameter choice. However, in the manuscript we explicitly point to the github repository where these pipelines have been implemented. This allows the interested reader to study command calls with corresponding parameter choices.*

- Specify which registrations are linear and which are non-linear (although I think all are non-linear)?

> *Done. Please see the revised* **ANTs cortical thickness** *section. Although the specification of "rigid" had been made in the original manuscript denoting such registrations, we augmented these details with "non-linear" and "deformably" to denote non-linear mappings where appropriate.*

- Equ 1 also assumes same slope for all subjects. In most cases this is not true as the slope changes at different ages and disease severities. It can be assumed constant for a disease group. To a large extend the sigma measures how well the model fits the data. A method that makes all slopes more similar to each other (or to zero) will therefore produce a better model fit. But the underlying model may be to simple to fit the real data.

> *The reviewer makes a good point and so we replaced the previous Stan model with a new one in which the individual slopes are also randomized. Although the results are slightly different, the relative trends between pipeline performance remain the same. Please see revised Figures 5, 6, and 7.*

Reviewer 3

This paper presents a longitudinal version of the cortical thickness measurement pipeline implemented in the open-source ANTS software package. Two slightly different versions are presented, both using a subject-specific template in a manner similar to the longitudinal stream implemented in FreeSurfer; the two versions differ in the way interpolation is handled (in subject space vs. atlas space). Quantitative results are presented on longitudinal data of ~600 ADNI subjects taken from four different disease groups. In particular, a (very simple) linear mixed-effects model is fitted to longitudinal cortical thickness results obtained with both proposed versions and also with FreeSurfer, and the results are analyzed in terms of the estimated parameters of this model (ratio of two parameters, as well as confidence intervals of another parameter).

Although the proposed pipeline is nicely described and seems perfectly reasonable, I am puzzled by the very convoluted and indirect way in which the authors aim to demonstrate the benefits of the proposed longitudinal method:

* Since the contribution of this paper is to introduce a novel longitudinal segmentation pipeline, a very direct validation is to simply demonstrate enhanced differentiation in estimated atrophy rate between the four ADNI diagnostic groups compared to when a cross-sectional segmentation method is used (similar to the cited Reuter/FreeSurfer paper the authors aim to compare themselves to). Confusingly, this straightforward and clarifying experiment is never actually performed, although enhanced differentiation between patients sub-populations is repeatedly mentioned as the ultimate goal of the method. Even more confusing for a paper introducing a longitudinal method is that the validation includes an experiment (Eq. 3) that does not even use longitudinal data.

* Instead of simply estimating (differences in) cortical thickness atrophy from the available longitudinal data in different diagnostic groups, the authors propose to use a linear mixed effects model and concentrate on properties of the estimated parameters of this model only. However, there are several issues with this approach:

> - The model is overly simplistic, ignoring not only the known diagnostic grouping of the available data (lumping together all ~600 ADNI subjects regardless of diagnosis, and imposing the same atrophy rate across all diagnostic groups), but also other commonly modeled fixed effects (such as gender and age) and potentially meaningful random effects beyond merely subject-specific offsets. It is not clear to me what can be learned by fitting such an oversimplified model to longitudinal data.

The reviewer makes a good point and so we replaced the previous Stan model with a new one in which the individual slopes are also randomized. Although the results are slightly different, the relative trends between pipeline performance remain the same. Please see revised Figures 5, 6, and 7.

> - Although the proposed ratio between the between-subject and within-subject standard deviations can perhaps be intuitively understood as measuring how well the model can explain the data, the authors make several very strong claims on how this translates into

greater discriminative capacity without providing any empirical and/or theoretical justification or references to the literature to back this up.

A rigorous statistical argument for the proposed approach may be found in 'Linear Regression Analysis' by Seber and Lee. We have added the reference and the following discussion to the Methods section.

> In particular, [71] (Sections 9.6.2 and 9.6.5) demonstrate the role that randomness and measurement error in explanatory variables play in statistical inference. When the explanatory variable is fixed but measured with error (as is plausible for cortical thickness measurements), the within-subject variance divided by the between subject variance is proportional to the bias of the estimated linear coefficient when the outcome of interest is regressed over the explanatory variable (Example 9.2). In short, the larger the rk, the less bias for future statistical analyses based upon the cortical thickness data. When the explanatory variable is considered random and is measured with error (a common assumption in the measurement error literature [72, 73], this bias is expressed as attenuation of regression coefficient estimates to zero by a multiplicative factor rk/(1 + rk) (Example 9.3). Thus, larger rk means less less attenuation bias and hence more discriminative capacity. Note that effect estimator bias is not the only problem— the residual variance is increased by a factor proportional to rk/(1 + rk) ([71] Chapter 3). The same authors refer to the combination of bias and added variance as a 'double whammy'. Indeed, a worse reliability ratio causes greater bias in multiple linear regression in the presence of collinearity and even biases the estimators for other covariates, progression through time included ([73], Section 3.3.1). The same authors state that this bias is typical even in generalized linear models (Section 3.6) and continue to use the ratio as a measure of reliability even in the longitudinal context (Section 11.9).

- The posterior distribution of the slope parameter is also analyzed, but here only the width of the distribution is considered, with the authors claiming that tighter distributions allow more "definite scientific conclusions to be reached". However, in my opinion equally important is the *mode* of this distribution: if a method yields slope estimates that are always centered around zero (no atrophy can be detected), it is a useless longitudinal method no matter how confident its predictions.

Please note the addition of Section **3.1 Regional cortical atrophy rates based on diagnosis**. *This section explores the use of slope values describing cortical atrophy to separate diagnoses (CN vs. LMCI vs. AD) in the ADNI-1 cohort amongst the different processing streams. In particular, see Tables 2 and 3. Note that we opted for this format given the number of pipeline x region x diagnosis combinations. However, as we mention in the manuscript, spaghetti plots and slope distribution plots are provided in the github repository, specifically in the directory to linked to here:*

https://github.com/ntustison/CrossLong/tree/master/Data/

- The model relies on several hard-coded hyperpriors (values of "5" and "10"), the effect of which is not clear.

These hyperpriors are considered the current best-practice (Gelman 2006). The following text is included in the **Statistical evaluation** *section (with new additions emphasized):*

> [S]pecification of variance priors to half-Cauchy distributions reflects commonly accepted best practice in the context of hierarchical models [64]. **They concentrate mass near zero but have heavy tails, meaning small variance values are expected but large variance values are not prohibited. Even so, results demonstrated robustness to parameter selection.**

Other things:

* Abstract: the order of the experiments is claimed to be the opposite of what is in the paper

This has been corrected in the current manuscript.

* In general the description of the method is very terse; it is sometimes difficult to grasp even the gist of certain processing steps. Some more details regarding e.g, how "voxelwise regional thickness" is computed; how interpolation in both atlas space and subject space is implemented; why CSF is treated so much differently from the other five classes; and how inference is performed with the Stan environment, would be good.

Regarding voxelwise regional thickness, we added the following text in the section **Unbiased longitudinal processing***:*

> **Registration-based cortical thickness.** The underlying registration-based estimation of cortical thickness, Diffeomorphic Registration-based Estimation of Cortical Thickness (DiReCT), was introduced in [8]. Given a probabilistic estimate of the cortical gray and white matters, diffeomorphic-based image registration is used to register the white matter probability map to the combined gray/white matter probability map. The resulting mapping defines the path between a point on the gray/white matter interface and the gray matter boundary. Cortical thickness values can then be assigned at each spatial location within the cortex by integrating along the diffeomorphic path starting at each gray/white matter interface point and ending at the gray matter/CSF boundary. A more detailed explanation is provided in [8] with the actual implementation provided in the class itk::DiReCTImageFilter available as part of the ANTs library.

Regarding implementation of interpolation in both atlas space and subject space---if the reviewer is asking about the actual implementation, or code, of the interpolation methods, note that all interpolation methods are ITK classes as ANTs uses ITK as a code foundation. If the reviewer is asking about interpolation choice, then similar to the two FreeSurfer streams used in this manuscript, choice of interpolation method is an admittedly important but simply one of many parameter/component choices for a given ANTs pipeline. Other image registration choices include the number of scaling pyramid levels, the shrinking factors at each level, the Gaussian smoothing kernel size at each level, the number of gradient steps, the convergence criteria, whether to perform histogram matching and/or *clamp voxels with outlier intensity values. And these are only the possible parameters for a single transform and must all be considered for both linear and non-linear registration "stages." A comparable number of component parameter choices exist for preprocessing components (e.g., N4 bias correction) segmentation, cortical thickness estimation, and joint label fusion and it is not obvious that any single parameter (e.g., interpolation method) is more important than any other. Needless to say it is beyond the scope of this manuscript to provide the information for each of these choices nor does it make sense to feature any single parameter choice. However, in the manuscript we explicitly point to the github repository where these pipelines have been implemented. This allows the interested reader to study command calls with corresponding parameter choices.*

Regarding the unique treatment of CSF, we added the following text to the subsection **Single-subject template, brain mask, and tissue priors:**

> Note that the unique treatment of the CSF stems from the fact that the 20 expertly annotated atlases only label the ventricular CSF. Since cortical segmentation accuracy depends on consideration of the external CSF, the above protocol permits such inclusion in the CSF prior probability map.

Regarding inference with the Stan environment, we added the following text (in bold) in the subsection **Regional within-subject and between-subject variance**:

> The posterior distribution of rk was summarized via the posterior median where the posterior distributions were obtained using the Stan probabilistic programming language [65]. **The R interface to Stan was used to calculate the point estimates of the LME model (1) for cortical thickness across the different pipelines using the default parameters. The csv files containing the regional cortical thickness values for all five pipelines, the Stan model file, and the R script to run the analysis and produce the plots are all located in the github repository created for this work [35].**

* The literature overview on validation (page 13) talks about cross-sectional validation, whereas validation of longitudinal methods (which this paper is about) would be more appropriate.

We included the following mention of longitudinal assessment strategies at the beginning of the **Statistical evaluation** *section***:**

> Similarly, evaluation strategies for longitudinal studies have been proposed with resemblance to those employed for cross-sectional data such as the use of visual assessment [24], scan-rescan data [12, 25], and 2-D comparisons of post mortem images and corresponding MRI [25]. In addition, longitudinal methods offer potential for other types of assessments such as the use of simulated data (e.g., atrophy [12, 25], infant development [24]) where "ground-truth" is known and, regression analysis of longitudinal trajectories of regional cortical thickness [60].

* Discussion: unless I missed it, no "diagnostic prediction using extreme gradient boosting" has been performed in the paper.

> *This was an oversight on our part during the final review of the manuscript. The section has been removed.*

From Brian (not to include but here for transparency).

overall: it's getting there but i think a couple key message points could be improved. first, why technical users might find it interesting. second, why somone like cliff jack would find it interesting.

computationally oriented researchers want to know the technical benefits (full invertible maps, unbiased approach, etc) whereas clinical researchers just want to know how it will improve detection power and atrophy rate estimates. for instance, they want annualized percent change values for each DX.

my opinion of why ppl might use LACT and my score for how well the paper gets this across

* sets up multiple modality studies
* allows faster JLF (see jeff's longitudinal JLF script in ANTs)
* provides smooth parameter estimates for registration and segmentation -based measurements (in addition to and beyond the thickness values)
* provides consistent/invertible mapping through the SST to subject and to template space
* collects / resolves longitudinal segmentation / registration in one place
* high reliability/low failure rate
* detection power is better
* can be customized relatively easily (i think this is covered)

readers should get answers to these questions:

Q: when is LACT "better" than alternative options?

LACT inherits the performance benefits of ACT and therefore provides high reliability for large studies and robust registration and segmentation in human lifespan data as well as in datasets that exhibit large shape variation eg what we see in AD, FTLD, etc.

Q: when would one choose to use this approach over X,Y,Z?

if you have non-human data; if you want to perform TBM; if you want longitudinal JLF; if your data has lots of anatomical variation (need to capture large differences between subjects and still want accurate registration)

Q: what is the relevance to AD?

LACT uses unbiased diffeomorphic registration to provide robust mapping of individual brains to group template space and, simultaneously, high-resolution sensitivity to subtle longitudinal changes over time. Both advantages are relevant to AD. High baseline atrophy levels in AD lead to the need for robustness to large deformations. Sensitivity to subtle longitudinal change over time is particularly relevant to early or preclinical AD studies due to the relatively reduced atrophy rates and smaller difference from control populations. This paper demonstrates that the LACT approach leads to competitive or superior estimates of annualized atrophy that are biologically plausible in AD populations and that may, in the future, support the use of T1 neuroimaging to detect treatment effects in clinical trials. The most sensitive regions in this study are X,Y,Z. Furthermore, in ADNI-1, we reported zero percent failure rate (or whatever) with no subject-specific tuning required.

other notes/questions:

1. in introduction: should make clear some of the other possible applications, just 1 sentence

Although we limit exploration in this work to ROI- based analysis for simplifying comparison with FreeSurfer, there are several additional applications permitted by the ANTs framework such as longitudinal tensor-based morphometry, Eigenanatomy [33], and extension to non-human data.

2. in abstract: make clear the concrete findings. e.g. The LACT pipeline provides unbiased structural neuroimage processing and competitive to superior power for longitudinal structural change detection in ADNI.

Done. Added that sentence to the end of the abstract.

3. in discussion: highlight other uses of pipeline. highlight the reliability of the pipeline: low failure rate is critical in large studies where the need for manual intervention should be minimized. Question: what is the failure rate?

In the **Discussion** *section, I added the following sentence in the first paragraph:*

> All ANTs components are built from the Insight Toolkit which leverages the open-source developer community from academic and industrial institutions leading to a robust **(e.g., low failure rate)** software platform which can run on a variety of platforms.

and at the end of the second paragraph:

> Furthermore, in ADNI-1, we report a zero percent failure rate with no subject-specific tuning required.

4. discussion: might mention future work --- current paper focuses on thickness but many other aspects of the LACT are useful. jacobian, geometric measurements (surface area, geometric thickness, curvature SurfaceCurvature) and the subsequent "neurobattery" applications for more longitudinally sensitive processing of rsfmri, dti, etc.

In the **Discussion** *section, I added the following paragraph:*

> However despite these deficiencies, being inherently voxel-based, the ANTs framework does have advantages not explored in this work but certainly to be utilized in future research. Specifically, the voxel-based input/output processing is conducive to voxel-based analysis strategies (e.g., Eigenanatomy [79]) and straightforward application to non-human research domains. Also, tensor- based morphometric data are directly extracted from the output of the longitudinal processing. And while mesh-based geometric measures are unavailable, digital analogs (e.g., surface area from the digitized Crofton formula [50] and surface curvature [80]) provide a convenient data format for integrated data analysis. Finally, given the importance of structural data, such as T1-weighted images, for other types of neuroimaging studies (e.g., resting state fMRI and diffusion tensor imaging), the longitudinal processing stream provides convenient output for facilitating these other types of analyses.

5. did we verify "lack of bias"? if so , perhaps mention in abstract and conclusions as a takeaway point.

*I don't think we *verified* lack of bias. However, in the* **Discussion** *I did mention how the pipeline addresses the various forms of addressing bias based on the points you raise in item 10 below.*

6. might add an "easy to read" table that summarizes most important statistical findings? existing tables are dense/hard to read

Replaced old figure showing slope distributions with Tables 2 and 3.

7. fix " evaluation strategy is evaluated " in abstract

Done.

8. why 'interpolation' a key word?

Honestly, I can't remember. Removed it.

9. question: regarding prior creation - is JLF quantitatively better for creating the priors vs just mapping from the template? did you do a study on this?

Correct me if I'm wrong but I'm assuming you are referring to the use of the `-a` option in `antsLongCT.sh` vs. not using it. We never did a study regarding these two options. However, the whole reason why we favored using the JLF atlas-based approach (and put it as default) even though it takes longer is because, if you don't, you frequently get "label creep" where strong deep gray matter prior values spread to nearby cortical gray matter e.g., insula from the putamen.

10. in beginning of discussion,, you mention "bias issues" ... might note here (or elsewhere) that LACT addresses bias via:

a. denoiseImage - different noise levels across time, scanners
b. N4 - different bias field '...'
c. SST - reference space, consistency of registration solution from group to subject space
d. syn - symmetric pairwise mapping
e. jlf - not sure ....
f. more?

I added the following blurb to the beginning of the Discussion section (bold denotes additions):

> Herein we detailed the ANTs registration-based longitudinal cortical thickness framework which is designed to take advantage of longitudinal data acquisition protocols. **It inherits the performance capabilities of the ANTs cross-sectional pipeline providing high reliability for large studies, robust registration and segmentation in human lifespan data, and accurate processing in data (human and animal) which exhibit large shape variation. In addition, the ANTs longitudinal pipeline accounts for the various bias issues that have been associated with processing such data. For example, denoising and N4 bias correction mitigate the effects of noise and intensity artifacts across scanners and visits. The use of the single-subject template provides an unbiased subject-specific reference space and a consistent intermediate space for normalization between the group template and individual time points. Undergirding all normalization components is the well-performing SyN registration algorithm which has demonstrated superior performance for a variety of neuroimaging applications (e.g., [57, 73]) and provides accurate correspondence estimation even in the presence of large anatomical variation. Also, given that the entire pipeline is image-based, conversion issues between surface- and voxel-based representations [74] are non-existent which enhances inclusion of other imaging data and employment of other image-specific tools for multi-modal studies (e.g., longitudinal cortical labeling using joint label fusion and the composition of transformations). All ANTs components are built from the Insight Toolkit which leverages the open-source developer community from academic and industrial institutions leading to a robust (e.g., low failure rate) software platform which can run on a variety of platforms.**
>
> Over 600 subjects from the well-known longitudinal ADNI-1 data set with diagnoses distributed between cognitively normal, LMCI, and AD were processed through the original ANTs cross-sectional framework [26] and two longitudinal variants **with no processing failures**.

lastly. in your prediction study, one would expect APOE, gender and brain volume to play a role. also you would expect something more like this:

See if Equations (3) and (4) are sufficient.

~~~~
```{r basicADNIpower}
library( ADNIMERGE )
library( MBESS )
library( rsq )
dxpt = 'LMCI'
fl=adnimerge$VISCODE == 'm24' & !is.na( adnimerge$Entorhinal ) &
  !is.na( adnimerge$Entorhinal.bl ) & ( adnimerge$DX.bl %in% c("CN",dxpt)  )
mysubjects = unique( adnimerge$PTID[fl] )
bl=adnimerge$VISCODE == 'bl' & (adnimerge$PTID %in% mysubjects )
studyDF = data.frame(
  age = adnimerge$AGE[bl],
  gen = adnimerge$PTGENDER[bl],
  apoe = factor( adnimerge$APOE4[bl] ),
  icv = adnimerge$ICV.bl[bl],
  dx  = factor(as.numeric( adnimerge$DX.bl[bl] == dxpt )),
  erb  = adnimerge$Entorhinal.bl[bl],
  er  = adnimerge$Entorhinal[fl] )
studyDF$erd = studyDF$er - studyDF$erb
bmdl = glm( dx ~ age + gen + apoe + icv + erb, data = studyDF, family='binomial' )
fmdl = glm( dx ~ age + gen + apoe + icv + erb + erd, data = studyDF, family='binomial' )
print( table( studyDF$dx, predict(fmdl, type='response')>=0.5) )
pander::pander( summary( fmdl ) )
with<-rsq(fmdl)
without<-rsq(bmdl)
powerSummary = ss.power.reg.coef(
        Rho2.Y_X = with,
        Rho2.Y_X.without = without,
        alpha.level=0.05,
        desired.power=0.8, p=5 )
necessarysubjects<-powerSummary$Necessary.Sample.Size
print( necessarysubjects )

```
~~~~

or maybe

~~~~
```{r basicADNIpower2}
library( ADNIMERGE )
library( MBESS )
library( rsq )
~~~~

```
fl=adnimerge$VISCODE == 'm24' & !is.na( adnimerge$Entorhinal ) &
  !is.na( adnimerge$ADAS13) &
  !is.na( adnimerge$ADAS13.bl) &
  !is.na( adnimerge$Entorhinal.bl ) & ( adnimerge$DX.bl %in% c("EMCI","LMCI")  )
mysubjects = unique( adnimerge$PTID[fl] )
bl=adnimerge$VISCODE == 'bl' & (adnimerge$PTID %in% mysubjects )
studyDF = data.frame(
  age = adnimerge$AGE[bl],
  gen = adnimerge$PTGENDER[bl],
  apoe = factor( adnimerge$APOE4[bl] ),
  icv = adnimerge$ICV.bl[bl],
  dx  = factor(as.numeric( adnimerge$DX.bl[bl] == dxpt )),
  erb  = adnimerge$Entorhinal.bl[bl],
  er  = adnimerge$Entorhinal[fl],
  adab  = adnimerge$ADAS13.bl[bl],
  ada  = adnimerge$ADAS13[fl] )
studyDF$erd = studyDF$er - studyDF$erb
studyDF$adad = studyDF$ada - studyDF$adab
bmdl = glm( adad ~ age + gen + apoe + icv + erb, data = studyDF, family='gaussian' )
fmdl = glm( adad ~ age + gen + apoe + icv + erb + erd, data = studyDF, family='gaussian' )
print( cor.test( studyDF$adad, predict(fmdl, type='response') ) )
plot( studyDF$adad, predict(fmdl, type='response') )
pander::pander( summary( fmdl ) )
with<-rsq(fmdl)
without<-rsq(bmdl)
powerSummary = ss.power.reg.coef(
      Rho2.Y_X = with,
      Rho2.Y_X.without = without,
      alpha.level=0.05,
      desired.power=0.8, p=5 )
necessarysubjects<-powerSummary$Necessary.Sample.Size
print( necessarysubjects )

```

as well as maybe an annualized change estimate (can be derived from above)

side note: i've found this code helpful in improving longitudinal detection power:

```
#  apply smoothing matrix based on each subject's temporal visits
for ( s in unique( subdemog$PTID ) ) {
  rowsel = subdemog$PTID == s
if ( sum( rowsel ) > 1 ) {
 subjectMatrix = submat[ rowsel, ]
 subjectDays = matrix(subdemog$Days[ rowsel ],ncol=1)
 daydist = as.matrix(
  dist( subjectDays, method = "euclidean", diag = TRUE, upper = TRUE, p = 2) )
 gaussdist = as.matrix( exp( -1.0 * daydist / gaussSig ) )
 gaussdist = gaussdist / Matrix::rowSums( gaussdist )
 smoothed = gaussdist %*% subjectMatrix
 submat[ rowsel, ] = smoothed
 }
}
```