Reviewers' comments

Handling Editor

The revised submission seems to have generated as many new issues as it resolved. Some significant concerns about the evaluation metrics, the evaluation of longitudinal accuracy, and several other issues require a substantial revision and response. Without detailed, transparent and convincing response, the paper is below acceptance levels

> *We appreciate the previous efforts of the Editors and Reviewers. We believe that the first two rounds of reviews were extremely constructive resulting in the much improved manuscript in its current form and we are grateful to the Reviewers for their significant contribution. Many of the current issues, however, seem to stem from an uncharitable reading of the current manuscript and of our previous responses (e.g., accusations of possible "data massaging") as well as conflicting directives. Although we question the current impartiality of one or more of the Reviewers and are, therefore, somewhat dubious concerning the possibility of a placating rejoinder on our part, the following general response and subsequent point-by-point responses to each of the Reviewers is honestly a good faith attempt.*
>
> *To address some broader issues, we would first like to summarize the overarching narrative of the manuscript:*
>
> 1) *A new framework is presented for longitudinal cortical thickness processing. This previously unpublished (in journal form) framework is developed and made available within the well-vetted and popular open-source ANTs toolkit.*
> 2) *In contrast to many publications which propose validation/evaluation through the use of one or more domain-specific experiments, we use a generic criterion, well-established within the statistical literature, to provide a general comparative performance assessment with other popular alternatives (i.e., FreeSurfer).*
> 3) *Further evidence for the utility of our work is supported by a relatively straightforward experiment involving specific aspects of the selected ADNI data set meant to reflect a typical use case scenario. This experimental evaluation is also an attempt to incorporate a response to previous criticisms (Reviewer 2: "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.") with the same evaluation criteria as before (How well do cortical thickness measurements between the various pipelines differentiate clinically derived diagnoses?--since both Reviewers 2 and 3 are concerned about group distinctions specific to the ADNI data).*

> *Both 1) and 2) have remained more-or-less consistent through the different revisions whereas 3) has changed for each version of the manuscript. We are satisfied with our current choice for 3) given that it most accurately reflects our intent of showcasing our measurements in a more typical, real-world study scenario while simultaneously minimizing deviations from the specified directives of Reviewers 2 and 3. Previous choices did not reflect an ideal use of the longitudinal data, as mentioned in the review. The current version directly assesses change in thickness as it relates to diagnosis, as requested in prior rounds.*

Reviewer 1

The authors still didn't fully address the comment. As suggested, for the convenience of inspection of thickness values and changes ratio, it is important to show their values projected onto the corresponding brain regions, as in Fig 8. It is very hard to check these values when using tables.

> *As we also wrote in our previous response, we have made these values and corresponding plots available as supplemental material in the specified GitHub repository and added text to manuscript directing the reader to this material, if interested.*

Reviewer 2

In this second revision, the authors have removed what I considered was a dubious experiment with a hard-to-believe outcome (namely, that the age of old people can be predicted from temporal non-linearities in their brain atrophy progression), and, most importantly, clarified that the reduced ability to detect atrophy rate differences with the proposed method, which was reported in the previous revision, was the result of an accidental mix-up of column labels in the provided tables. I have several doubts about this revision:

* Instead of simply correcting the mislabeling of the columns, the authors have now inexplicably also changed the entire experiment using a much more complicated model, and we never get to know what the results were with the proper labeling of the original, easy-to-interpret experiment. None of the reviewers asked for this, and therefore I think it is crucial that the corrected version of the original experiment is disclosed so that the possibility of "data massaging" on the part of the authors can be excluded.

As far as we are aware, it is perfectly legitimate for the authors to improve the manuscript during the course of a review without the permission and/or suggestion of the reviewers.

*As we pointed out above, this modification was, in large part, a direct response to the criticism of Reviewer 2 in the previous round---"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." In addition to trying to accommodate the Reviewer, we wanted to maintain the same evaluation criteria in an ADNI context (How well do cortical thickness measurements between the various pipelines differentiate clinically derived diagnoses?) while incorporating the disease and population relevant covariates for a more sound experimental design.*

With respect to the accusations of possible "data massaging," we would like to assume that we deserve the benefit of the doubt. We are more than aware of such issues having written on related topics and also having a long history of supporting open-science with our open-source software efforts. More importantly, though, we have been completely transparent during the entire review process having made our data and analysis scripts available through our GitHub repository from the beginning.

* Assuming there is good explanation for the change in experimental set-up (in which case the authors should clearly justify it), the execution raises many questions:

- The additional inclusion of ICV muddles the experiment, as different tools will measure this quantity differently and with varying accuracy. This in turn risks conflating the reported differences in the ability to detect atrophy rates with issues that are orthogonal to the experiment being performed.

Nowhere did we mention ICV being derived from different tools in the previous version of the manuscript. We have since added the following:

> Pipeline-specific LME models were constructed for each DKT region relating the change in cortical thickness to diagnosis and other pertinent covariates taken directly from the ADNIMERGE package.

This should satisfy any concerns of any confounding effects (i.e., "muddling") of the use of different ICV measuring tools. The Reviewer can verify this in our analysis starting from the following line:

> https://github.com/ntustison/CrossLong/blob/master/Scripts/Analysis/diagnosticSeparation.R#L83

- Equation (3) uses non-standard notation in which regression coefficients are omitted and a variety of symbols are used without precisely explaining their meaning. Even after digging up the 1973 paper (reference [82]), I still don't understand what "VISIT:DIAGNOSIS" means. More importantly: what is "(1|ID)"? It sounds like a one-hot basis function is used (zero for all subjects except for the one in which atrophy rate is being predicted), but since each subject has only one atrophy rate measurement this would be nonsensical (?)

We have rewritten Equation (3). The section in question now reads

> Pipeline-specific LME models were constructed for each DKT region relating the change in cortical thickness to diagnosis. These regional LME models are defined as:
>
> $$\begin{aligned}\Delta Y_{ij}^{k} = \beta_0 + Y_{i,bl}^{k}\beta_1 + AGE_{i,bl}\beta_2 + ICV_{i,bl}\beta_3 + APOE_i\beta_4 + GENDER_i\beta_5 \\ + DIAGNOSIS_i\beta_6 + VISIT_{ij}\beta_7 + VISIT_{ij} \times DIAGNOSIS_i\beta_8 \\ + \alpha_i^k + \gamma_s^k + \epsilon_{ij}^k\,. \end{aligned} \tag{3}$$
>
> Here, ΔY_{ij}^{k} is the change in thickness of the k^{th} DKT region from baseline (bl) thickness measurement $Y^{k}_{i,bl}$ for the i^{th} subject at the j^{th} time point. The subject-specific covariates (common to many ADNI-based studies) *AGE*, *APOE* status, *GENDER*, *DIAGNOSIS*, and *VISIT* were taken directly from the ADNIMERGE package. α_{ik}, γ_{sk}, and ε^{k}_{ij} are independent, mean zero random variables representing individual-specific random intercepts, site-specific (indexed by *s*) random intercepts, and residual errors, respectively.

- Despite the experiment predicting age from atrophy rate differences being removed from the current revision, the model in equation (3) still tries to predict atrophy rate from age. Is this really meaningful, and if so: why?

*The model includes **baseline** age as a covariate to control for age related effects independent of disease. Given the fairly broad age distribution in Alzheimer's disease and related disorders, we believe this is a sensible covariate to include.*

- Tables 2-3 report "difference[s] in slope values", but with the more complicated model of equation (3) it is not entirely clear what is being shown. Presumably differences in the (omitted) coefficients of the DIAGNOSIS basis functions?

This was an oversight on our part. The caption for both Tables now reads "95% confidence intervals for the diagnostic contrasts (LMCI–CN, AD–LMCI, AD–CN) of the ADNI- 1 data set for each DKT region…"

- Not sure showing log-scaled p-values (Figure 9) is really something to be encouraged. Why not directly show average/median slope differences?

> *As discussed in Boos and Stefanski, P-Value Precision and Reproducibility, Am Stat. 2011; 65(4): 213–221.,*
>
> > The analysis of *p*-value and log10(*p*-value) standard errors for the Miller data illustrates the generally observed phenomenon that the large variability of the *p*-values implies that only the *magnitude* of the *p*-value is accurate enough to be reliably reported. This observation runs counter to the current emphasis on reporting exact *p*-values, but it roughly coincides with use of *, **, and *** to denote levels of statistical significance often found in subject-matter journals.
>
> *Given the large variability associated with estimated p-values, comparisons are facilitated via the log transformation i.e.,*
>
> > However, because the distribution of the *p*-value is highly skewed, it is preferable to work on the log scale.
>
> *The reason why we do not show absolute values of average/median slope differences is because such absolute measures, in and of themselves, do not provide a relative performance assessment. However, as we have directed the interested reader, such values are easily obtainable from the material included in the GitHub repository.*

* Since my initial review of the first submission, I've been trying to point out that the proposed ratio between between-subject variability and residual variability is an awkward metric for longitudinal processing methods, since it will show excellent values for methods that over-regularize across time (a number one concern for longitudinal methods, since they intentionally introduce such regularization). Even though at least one other reviewer also pointed out that more straightforward validation metrics exist, and I have additionally pointed out that the proposed measure dubiously lumps together the different atrophy rates of various disease groups, the authors insist in putting it forward as the main validation outcome, which I think is unfortunate when such clear and easy-to-interpret alternatives exist.

> *The current version of the manuscript tries to put all evaluation criteria on the same footing. Some additional time is spent discussing variance ratio because its application in neuroimaging is relatively new (though it is well established in the statistical literature) and this warrants some additional explanation in this context. We have made attempts to explain the validity of this measurement in response to the Reviewer's repeated assertions concerning the use of the combined between-subject and residual variabilities. However, the Reviewer seems to gloss over our responses (including discussion of precedent in the statistical literature) despite our repeated attempts to*

address the Reviewer's concerns. Perhaps a line-by-line examination would be more effective:

> Since my initial review of the first submission, I've been trying to point out that the proposed ratio between between-subject variability and residual variability is an awkward metric [*Despite assertions to the contrary, the variance ratio is a completely valid metric to use in the given context. It draws support from the citations and corresponding discussion of the statistical literature which the Reviewer continues to gloss over or ignore.*] for longitudinal processing methods, since it will show excellent values for methods that over-regularize across time [*The Reviewer continues to assert this without evidence or citation despite repeated explanations of why this is not the case. If we limited the variance ratio to the denominator i.e., the residual variability, then the Reviewer would have a point. We could simply duplicate the baseline values and the residual variability would be zero resulting in infinite values for our proposed metric. Obviously, this is not the case as we are also including the between-subject variability. It is also clear that this hypothetical degenerate scenario is not applicable as we are actually demonstrating feasible results on the follow-up experiment dealing with diagnostic differentiation using the cortical thickness measures.*] since it will show excellent values for methods that over-regularize across time (a number one concern for longitudinal methods, since they intentionally introduce such regularization) [*Longitudinal methods do not necessarily "intentionally introduce such regularization." Neither the longitudinal FreeSurfer stream nor any of the ANTs longitudinal pipelines introduce any type of regularization on the cortical thickness measurements. We explicitly touch on this issue of over-regularization when we write in the Introduction (which has been there for at least a couple rounds):*
>
> > The ANTs framework also permits rotation of the individual time point image data to the SST, similar to FreeSurfer, for reducing variability, minimizing or eliminating possible orientation bias, and permitting a 4-D segmentation given that the Atropos segmentation implementation is dimensionality-agnostic [39]. Regarding the 4-D brain segmentation, any possible benefit is potentially outweighed by the occurrence of "over-regularization" [12] whereby smoothing across time reduces detection ability of large time-point changes. Additionally, it is less than straightforward to accommodate irregular temporal sampling such as the acquisition schedule of the ADNI-1 protocol.
>
>]. Even though at least one other reviewer [*a small interjection to point out that while Reviewer 3 also has serious issues with our use of the variance ratio, Reviewer 3 has actually engaged directly with what we have written and does not seem to gloss over our previous responses.*] also pointed out that more straightforward validation metrics exist, and I have additionally pointed out that the proposed measure dubiously lumps together the different atrophy rates of various disease groups, the authors insist in putting it forward as the main validation outcome, which I think is unfortunate when such clear and easy-to-interpret alternatives exist. [*We readily acknowledge the existence of alternative approaches for validation/evaluation. Although potentially perceived as more "straightforward" and "clear" and "easy-to-interpret", as we have pointed out, these performance measures are limited in that they are experiment-specific which calls into question their carry-over utility to other studies hence the motivation for the variance ratio--an approach which is well-established in the statistical literature.*]

Reviewer 3

Overall evaluation:

The primary changes as compared to the previous version comprise a change of results of the longitudinal analyses (correction of a technical error and apparently also a reanalysis) as well as a shortening of the evaluation section (age and MMSE predictions are dropped). As a result, the evaluation section of the manuscripts now primarily consists of the evaluation of the reliability ratio as well as comparison of significance values obtained from the different variants of the algorithms.

Generally speaking, the addition of a longitudinal analysis stream to the ANTs pipeline is a welcome addition to the field. Further, the authors are to be commended with regard to transparency with regard to data, analysis, and manuscript, primarily by means of the github repository. At the same, I have to say that I still have reservations with regard to the evaluation of the longitudinal analysis pipeline, as detailed below. My overall evaluation of the current revision of the manuscript is therefore that the work has merits with regard to the development of a longitudinal analysis pipeline, but continues to have issues concerning its evaluation.

A. Evaluation of the reliability

A central tenet during the methods section is that various precision-type evaluation criteria do not indicate clinical utility, and that better measures than specific precision indicators are needed to establish the utility of a new measure as a potential biomarker. The authors hence propose to use the variance ratio, defined as the quotient of between-subjects and residual variability. I agree that various measures have their shortcomings, but I disagree that the proposed criterion is a one-cure-for-all-problems solution.

> *We are unaware of any reference to the variance ratio being labeled or implied to be a "one-cure-for-all-problem" solution. What we have continued to assert is that such a measurement has generic utility over many alternative evaluation strategies, most (if not all) of which are domain-specific which may or may not generalize to other experimental designs and/or data and/or applications.*

The use of the variance ratio is motivated by a discussion, by Seber and Lee, of how regression models are affected by departures from underlying statistical assumptions.

Particular attention is given to measurement error in covariates, which can lead to bias and loss of power.

This is useful, since potential shortcomings of an analysis pipeline can statistically be conceptualized as measurement errors. At the same time, the Seber and Lee treatment is not specifically tailored to longitudinal analyses, and it is not obvious how their treatment applies to the analyses presented in the manuscript, where cortical thickness is used as the dependent variable, and not as a predictor/covariate as discussed by Seber and Lee. Conversely, it is unfortunate that the analyses that resemble the scenarios outlined by Seber and Lee (i.e., use of cortical thickness as a predictor) have been removed from the manuscript in its current version.

> *As we mentioned above, we are sincerely trying to address the concerns of the Reviewers. This is obviously difficult when attempting to address perceived conflicting directives from the Reviewers. As we pointed out above, this modification was, in large part, a direct response to the criticism of Reviewer 2 in the previous round---"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." In addition to trying to accommodate Reviewer 2, we wanted to maintain the same evaluation criteria in an ADNI context (How well do cortical thickness measurements between the various pipelines relate to clinically derived diagnoses?---both Reviewer 2 and 3 are concerned over group distinctions) while incorporating the relevant covariates for a more focused and data-appropriate experimental design.*
>
> *As noted by the Reviewer, supporting evidence for establishing precedence in the statistical literature for the variance ratio includes a discussion of the work by Seber and Lee (a general resource for linear regression analysis). However, the Reviewer neglects the two other important sources used in the Discussion which specifically address the topic of measurement error. These are "Measurement Error Models" by Fuller and "Measurement Error in Nonlinear Models: A Modern Perspective." The Reviewer would be served by considering the totality of our discussion and its supporting sources. Again, this manuscript concerns the ANTs longitudinal pipeline---not the validity of the variance ratio which is well-established in the statistical literature. It is certainly not the only approach to assessment of measurement error but it is a sound one.*

As a consequence, I do not see how the variance ratio generalizes to become a 'general desideratum for a variety of possible use cases'. Further, I do not see how the use of this criterion corresponds to the analyses presented in the current form of the manuscript - in other words, the connection between the two parts of the evaluation is not apparent to me.

> *We removed the word 'general'.*

I would suggest to construct an evaluation scenario that is consistent with the theory presented here. This could be an analysis that the slope of the change of cortical thickness, taken as an explanatory variable, distinguishes between groups. An alternative would be to substitute the reliability criterion by a different one. Another alternative would be to demonstrate how the theory relates to the empirical evaluation, i.e. the use of cortical thickness as a dependent variable in a longitudinal context. In part, these are suggestions that I have made during the previous round of revisions.

> *It is difficult to understand what is requested here. The variance ratio is not being evaluated---it is already established in the statistical literature. We maintain its utility based on this literature and find the suggested requests on the part of the Reviewer to seemingly establish its validity to be distracting from the main point of the paper (i.e., the pipeline itself). We count three evaluation criteria for the proposed pipeline. The variance ratio is a measure of data reliability through time, the second experiment addresses Reviewer's 2 concern regarding the use of cortical thickness as an outcome measure and group distinctions mentioned by both Reviewer 2 and 3.*

I will now present the details of my critique according to the development of the methods section:

(2.5)

"For a longitudinal biomarker to be effective at classifying subpopulations, it should have low residual variation and high between-subject variation"

* A high-between subjects variance does not per se allow to distinguish between sub-populations. It may be necessary, but is not sufficient. What is required instead is a high between-group variation.

> *It is true that high between-subjects variation is necessary but not sufficient. Obviously, the absence of differences between groups makes it difficult to detect differences between groups. Nor would we like to do so. We do not claim anything to the contrary.*

"Specifically, we use longitudinal mixed-effects (LME) modeling to quantify pipeline-specific between-subject and residual variabilities where comparative performance is determined by maximizing the ratio between the former and the latter. Such a quantity implies greater within-subject reproducibility while distinguishing between patient sub-populations."

* Using a ratio of variances implies that maximizing the numerator, or minimizing the denominator, or both, will lead to improved performance. Then, it is not logically clear to me how increasing the between-subjects variability would necessarily lead to better performance in terms of group separation.

> *Based on the Reviewer's assumptions, the Reviewer's confusion is warranted---group separation is not necessarily improved by increasing between-subject variability. The variance ratio, as used here, is indeed naive to group differences. This attribute makes it valuable for data inspection in the context of blinded clinical studies. Subject-wise separation with respect to variance ratio is useful to have but, ultimately, must be followed up with unblinded analyses to reveal group differences. For example, we would be surprised if superior cortical thickness measurements (as determined by the variance ratio) provided better separation between those who preferred dark chocolate over milk chocolate. Group considerations are study-specific whereas motivation for the variance ratio is prior to such study-specific considerations. In related terms (and as alluded to by the Reviewer), between-subject distinction is a necessary but not sufficient condition for relevant group separation. The follow-up study wherein the longitudinal regional cortical thickness measurements were used to evaluate clinically determined diagnostic (i.e., grouping) contrasts in the context of ADNI provides relevant supporting experimental evidence of group differentiation with the ANTs pipelines. However, the purpose of the manuscript and the follow-up experiment are not meant to validate the use of the variance ratio which is clearly established in the statistical literature.*

"As such this amounts to higher precision when cortical thickness is used as a predictor variable or model covariate in statistical analyses upstream."

* This is along the lines of Seber and Lee, but it only covers two particular use cases: one where cortical thickness is used to predict a criterion of interest such as clinical group, and one where cortical thickness is used as a mere covariate, i.e. a possible confounder. While the former is an important application scenario, the latter is relatively less important in the current context.

> *Again, the Reviewer seems to have discarded consideration of much of the literature support for our use of the variance ratio which specifically couches our discussion in the context of measurement error considerations. Even if we concede that the "latter" might be "relatively less important in the current context", it is still useful. Covariates can be included in regression models either as predictors of interests, confounders, or precision variables.*

* What I consider an issue here is that one important use case is missing, i.e. the use of cortical thickness as the outcome / criterion, and that b) analyses similar to the first use case have been removed from the manuscript.

> *As we pointed out above, this modification was, in large part, a direct response to the criticism of Reviewer 2 in the previous round---"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." In addition to trying to accommodate the Reviewer, we wanted to maintain the same evaluation criteria in an ADNI context (How well do cortical thickness measurements between the various pipelines differentiate clinically derived diagnoses?) while incorporating the relevant covariates for a more sound experimental design.*

"This ratio is at the heart of classical statistical discrimination methods as it features both in the ANOVA methodology and in Fisher's linear discriminant analysis. These connections are important since the utility of cortical thickness as a biomarker lies in the ability to discriminate between patient sub-populations with respect to clinical outcomes."

* The connection is somewhat far fetched, as different ANOVA models use different variance ratios, and longitudinal models models in particular can be quite different from cross-sectional ones where the above may apply.

> *It is unclear what the reviewer means by "different variance ratios." Statisticians include time and its interactions as covariates in longitudinal LME models. Once the data has been thus de-trended, it is precisely an ANOVA.*

"In particular, [74] (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 residual 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 in statistical analyses."

* Less bias is certainly desirable, but not an indication of reliability. Therefore, it is unclear to me how this would allow for an interpretation of rk in terms of reliability.

> *This is a well accepted interpretation within the measurement error literature. We agree that bias is not good. Also, the relevant discussion of the literature sources seems to have been overlooked by the Reviewer where we write*
>
>> 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 [76], Section 3.3.1). The same authors state that this bias is typical even in generalized linear models (Section 3.6) and use the ratio as a measure of reliability *even in the longitudinal context* (Section 11.9).

"When the explanatory variable is considered random and is measured with error (a common assumption in the measurement error literature [75, 76]), 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."

* Agreed, but this applies only for the case random explanatory variables, whereas in the previous paragraph cortical thickness is considered as a fixed covariate, if I am not mistaken. Trying to make a point for both the fixed and random covariates case is not convincing when only one particular analysis scenario is under consideration.

> *The point is that a larger variance ratio is beneficial regardless of the measurement error paradigm considered. Again, we are not trying to establish the validity of the variance ratio through the follow-up experiment. They are both (one general, one domain-specific) used to explore different performance aspects.*

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

* It is this last half-sentence that is actually the most interesting, i.e. to explain how the proposed criterion can be used as a measure of reliability for longitudinal analyses. This needs to be more precise, and a simple reference is not sufficient.

> *We wrote "and even biases the estimators for other covariates, progression through time included." This is precise. Also, given that we are not trying to establish the validity of the variance ratio but rather point to the literature for the requisite context, it is unclear what exactly is wanted here.*

B. Group discrimination

I have a few relatively minor issues about the analysis regarding the discrimination of diagnostic groups:

* The data that are presented within the updated table appear to result from a reanalysis, not just a simple re-ordering of rows, as I would have presumed based on the previous responses to the reviewers. I could not find any mention of this in the previous response. What has been changed in contrast to the preceding analysis, and why?

Please see the general response above. Our response was based on the previous Reviewers' criticisms and a reassessment of the additional experimental evaluation.

* Aggregating across the brain is convenient to come up with summary measures, but aggregation is less convincing when for a given disease not all regions of the brain are expected to be affected to the same extent. Similarly, whenever single regions are highlighted, such as the lateral occipital gyri, their relevance for the disease under consideration should be made clear.

We disagree that "whenever single regions are highlighted...their relevance for the disease under consideration should be made clear." This is a paper primarily concerning the ANTs pipeline---not AD-specific findings. However, to accommodate the Reviewer, we removed the specific mention of the lateral occipital gyri. The explicit mention of the lateral occipital gyri was simply provided in the context of performance trends. We did, however, make the following AD-specific statement:

> Although ANTs Cross 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 FS Cross and FS Long thus motivating the use of longitudinal considerations for processing of AD data.

* Whenever a novel measure is to be established as a biomarker, higher / lower p-values or wider / narrower confidence intervals alone do not tell the whole story, since statistical significance does not necessarily imply predictive power. If it is the goal of the current work to establish longitudinal cortical thickness changes as determined by the ANTs pipeline as a biomarker, more than (increased) statistical significance needs to be demonstrated, in particular generalization to unseen data, by means of prediction intervals. Otherwise, these necessary steps should at least be discussed as requirements for future work.

We agree with the Reviewer. Establishing biomarkers is a multi-pronged process and will certainly not be restricted to any single manuscript derived from this work (regardless of whether or not it is ultimately published in NeuroImage). As noted by the Reviewer, we did see increased statistical significance in the majority of brain regions in terms of the diagnostic contrasts. In addition, as we have discussed seemingly ad nauseum up to this point, the use of the variance ratio motivated from the statistical literature is simply one component supporting the utility of the proposed ANTs-based longitudinal pipeline. To indicated follow-up work we added the following sentence to the Discussion:

> *Furthermore, future studies, e.g., cross validation and prediction, will provide further understanding of performance characteristics.*

* One final question about consistency: is there any particular reason to employ Bayes methods for the reliability models, and frequentist methods for the evaluation of the group discrimination?

The Bayesian approach described in Section 2.5 facilitates credible intervals for the variance ratio. The frequentist confidence interval requires the delta method which itself requires pen-and-paper calculations we would rather avoid. The more common frequentist approach utilized in Section 2.6 represents typical use-case scenarios. In both cases, the results are valid in terms of methodology.