We appreciate the time spent by the editors and reviewers in assessing our manuscript.

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

Reviewer 1:

I would like to thank the authors for their revisions to the manuscript. I will limit my attention, in this review, to outstanding issues with the content, and with the response to the reviewers:

Despite the reorganization of the manuscript, I find my criticism about the misleading nature of the article still hold. The bulk of the article's technical details concern a validation of ANTSx cortical thickness (indeed, with the inclusion of the longitudinal pipeline this is slightly increased). The details in the methods section describing the training of the component models are not specific enough (barring a few mentions of ANTSx specific function calls) to give the reader a feel for the 'novel aspects of the toolkit' to recommend it above the combination of Tensorflow/Pytorch and ITK/Nibabel (or even ANTS itself) used by hundreds of researchers to submit to MICCAI, MIDL and other venues each year.

It is important to note that we are not claiming "novelty" via training of the component models as we are adapting pre-existing deep learning technology. Nor do we necessarily "recommend it above [other toolkits]"---particularly toolkits focused on the research underlying deep learning developments. What we do claim is that ANTsX (which includes ANTs, ANTsR, ANTsPy, ANTsRNet, and ANTsPyNet) is potentially useful to many researchers---many of whom will never submit to (or even attend) MICCAI, MIDL, etc. but rather are more concerned with using the measurements provided by established processing pipelines which take advantage of deep learning technology.

The authors write that the intended readership is existing users of the ANTS ecosystem, and while I consider that a rather narrow scope for a scientific article I accept that the usership of ANTS is very large and growing. I also agree, wholeheartedly, with the authors, that there are no human-derived ground truths of sufficient quality with which one may assess segmentation performance. However, especially in light of the fact that the authors carefully tuned the segmentation (by altering tissue weights) to produce an output which 'most resembled the traditional Atropos segmentations' it would seem sensible to report statistics, on a large dataset, showing how well that tuning performed: i.e., how close are the segmentations of DeepAtropos to those of Atropos? This is standard practice when building a supervised classifier and presumably part of the ANTSPyNet framework. Indeed, some more information in the paper about how to monitor training, assess performance, etc. in ANTSx specifically would strengthen the assertion that this is a paper about an ecosystem, not an algorithm.

While I also agree that in general, the results of the cortical thickness pipeline may be a truer measure of segmentation quality than a Dice coefficient, surely users of ANTS (and potentially ANTSx) are interested in how well the thicknesses derived from the DeepAtropos segmentation agree with those of Atropos, in both relative and absolute terms? I imagine this would be of more interest to the readers than the feature importance plots in figure 4 (which, unsurprisingly, give different feature importance to different features, since many (most!) of these features will be highly correlated with each other, which can cause instability in variable importance.) Since I was interested, and the authors have made the code available, I ran both the ANTS and ANTSx algorithm over a number of cases from publicly available datasets and found a consistent bias toward smaller thicknesses from the ANTSx pipeline. This is not so surprising, and the consistency of global cortical thickness (as measured by ICC(3,1)) was relatively high (0.89), but it is precisely these analyses I imagine that would be of interest to a user who may think they can simply substitute ANTSx for ANTS in their workflow.

We agree that there are many additional analyses which readers might find interesting, including more in-depth segmentation-specific comparisons. Given the limitations due, at least in part, to submission guidelines, that is one of the many reasons why we make the data publicly available. However, we agree that providing additional insight concerning relative thickness values between the ANTs and ANTsXNet pipelines would indeed be useful. To this end, we replaced the importance plots in Figure 4 with radar plots showing relative thickness values with the following text:

> A comparative illustration of regional thickness measurements between the ANTs and ANTsXNet pipelines is provided in Figure 4 for three different ages spanning the lifespan. Linear models of the form
>
> $$T(DKT_i) \sim GENDER + AGE \quad (2)$$
>
> were created for each of the 62 DKT regions for each pipeline. These models were then used to predict thickness values for each gender at ages of 25 years, 50 years, and 75 years and subsequently plotted relative to the absolute maximum predicted thickness value (ANTs: right entorhinal cortex at 25 years, male). Although there appear to be systematic differences between specific regional predicted thickness values (e.g., $T(ENT)_{ANTs} > T(ENT)_{ANTsXNet}$, $T(pORB)_{ANTs} < T(pORB)_{ANTsXNet}$), a pairwise t-test evidenced no statistically significant difference between the predicted thickness values of the the two pipelines.

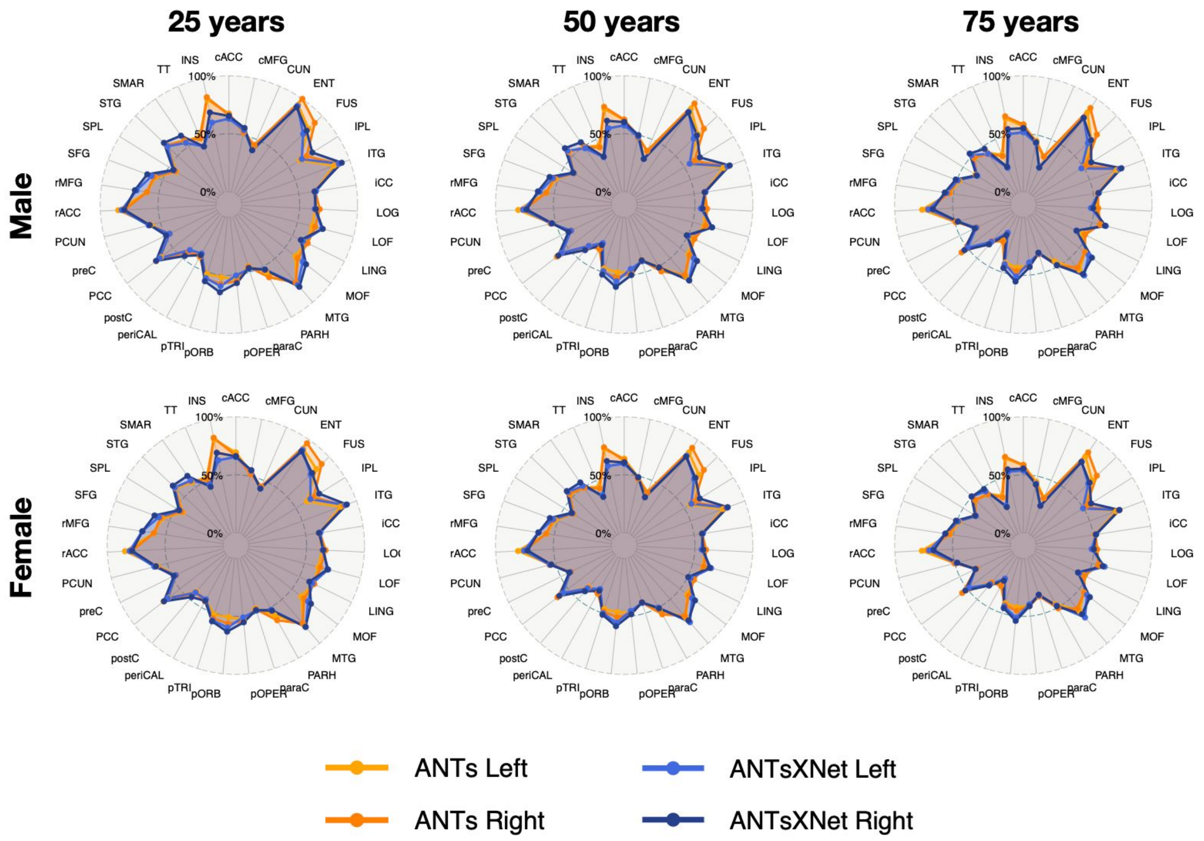

Figure 4: Radar plots enabling comparison of relative values between the ANTs and ANTsXNet cortical thickness pipelines at three different ages sampling the lifespan. See Table 2 for region abbreviations.

My suggestion was to include a comparison to the latest version of FREEsurfer (not FASTsurfer) as the previous publication comparing ANTS to Freesurfer in terms of age regression (used for Cross Sectional performance evaluation) was some time ago.

Yes, this was a mistake on our part due to conflation with the subsequent comment.

Given that usage of FreeSurfer 6 continues to be widespread and that no mention of improved or variation in thickness values is provided in the FreeSurfer release notes for version 7 (https://surfer.nmr.mgh.harvard.edu/fswiki/ReleaseNotes), we find the current comparison to be adequate for this publication. However, we agree with the Reviewer that a comparison with FreeSurfer 7 would be interesting especially given that it now incorporates two fundamental tools used in ANTs-based pipelines: N4 bias correction and adaptive denoising. In fact, one of our colleagues is performing this analysis using a novel performance criterion and plans to submit the work for future publication.

The discussion about 'fine-tuning' I now find difficult to follow: presumably the authors of other deep learning segmentation algorithms invested some effort in making sure their algorithms produced plausible segmentations: I had assumed that some effort was implied in tuning the hyperparameters of DiReCT to the segmentation, but this fine tuning is presumably performed by anyone developing a segmentation algorithm, especially where the results will be used in some downstream analysis.

There was no fine-tuning of the DiReCT algorithm itself. The Reviewer is correct in that the fine-tuning discussed concerns the Deep Atropos segmentations.

Removal of the section regarding the poor performance of the cross-sectional algorithm does not remove the issue: the ANTSx cross sectional algorithm performs poorly (much more poorly than the ANTS algorithm), and this deserves some attention.

*This is not accurate. We used multiple performance criteria for the different cortical thickness pipelines. The ANTsX cross-sectional pipeline performs well in the context of cross-sectional data using both age prediction and in terms of repeatability (i.e., ICC). However, its performance is relatively poor for one of the two criteria *specific to longitudinal data* which is not unexpected as it is not tailored for such data. In fact, its performance, in terms of variance ratio, on longitudinal data is comparable to the other two cross-sectional pipelines, ANTsCross and FSCross, as illustrated in Figure 5 (b) whereas, in terms of diagnostic differentiation, its performance is comparable to the well-performing ANTs longitudinal pipelines (Figure 6).*

In my (admittedly small) set of subjects I found several failures of skull stripping (primarily residual non-brain tissue identified as GM/WM) which are presumably not related to the subject but rather the specific image, and so would be smoothed over by the the longitudinal algorithm, yielding better results.

It is good to read that the Reviewer actually tried the software despite the poor results. While discussion of poor outcome for specific data is handled much better through discussion on the GitHub repository, we point out that poor skull stripping is something that was noticed a while back. In fact, the first author posted this observation as an issue on Jan. 7, 2021 (https://github.com/ANTsX/ANTsPyNet/issues/25) and provided a temporary workaround. The underlying issue is that much public data (including the vast majority of data used for ANTsXNet training) uses various defacing protocols for anonymization purposes. The SRPB1600 data set also uses a defacing protocol so we did not encounter such skull-stripping issues during our evaluation. However, we recently finished processing the OASIS-3 data which uses whole-head images and are currently training with these data to update our brain extraction weights.

Could that be the cause of the poor cross-sectional performance, and what does that say about using the ANTSx algorithm for cross-sectional studies in general?

Again, the cross-sectional ANTsXNet pipeline did not perform poorly for the cross-sectional evaluation. As mentioned above, poor performance of the ANTsXNet cross-sectional pipeline is restricted to one of two performance criteria designed specifically for longitudinal data. Performance of the traditional ANTs cross-sectional and the FreeSurfer cross-sectional pipeline were comparably poor for this performance criterion, which is not unexpected as they are not tailored for longitudinal data. For longitudinal data, we would recommend pipelines tailored for such data.

The authors mention 'repeatability assessment on the MMRR data set yielded ICC values ("average random rater") of 0.99 for both pipelines' Is this ICC for the age prediction? Why not report ICC for the repeated regional thickness measures themselves? If this is an ICC for thickness measures, which ones? It seems implausibly high.

We added the following clarification:

> *Additionally, repeatability assessment on the regional cortical thickness values of the MMRR data set yielded ICC values ("average random rater") of 0.99 for both pipelines.*

We did plot the regional repeated thickness measures in our NeuroImage 2014 paper and reported comparable numbers for both the FreeSurfer and ANTs cortical thickness pipelines. However, we are currently at the limit with respect to display items ("Display items are limited to 8 (figures and/or tables).") and we choose to keep the current set.

We invite the interested reader (including the Reviewer) to run the relevant code available at https://github.com/ntustison/PaperANTsX/blob/master/Data/Scripts/Analysis/repeatabilityKirby.R.

Reviewer 2:

The authors responded satisfactorily to most of my points and restructured the manuscript, which improved its quality. I have the following minor comments that should be addressed before publication, with line numbers provided in square brackets.

[53] Please cite the relevant literature on image registration using deep learning:

- Balakrishnan et al. VoxelMorph: a learning framework for deformable medical image registration. IEEE transactions on medical imaging, 2019;38(8):1788-1800. https://doi.org/10.1109/TMI.2019.2897538

- de Vos et al. A deep learning framework for unsupervised affine and deformable image registration. Medical image analysis, 2019;52: 128-143. https://doi.org/10.1016/j.media.2018.11.010

- Fu et al. DeepReg: a deep learning toolkit for medical image registration. Journal of Open Source Software, 2020;5(55):2705. https://doi.org/10.21105/joss.02705

Done.

[111] "Related work" requires "provides" in the singular form.

Done.

[114] It would seem appropriate to move the description of the original cortical-thickness pipeline from the results section to the introduction or the methods section, since it does not represent a result.

Done.

[145] Similar to the point above, the description of network architectures should be placed in the methods section.

Done.

Do the ANTsXNet networks implement the exact same architectures of references 25 and 42, including the number of filters, layers, etc. or adaptations thereof?

Given the sheer number of hyperparameters and ambiguity in the text of the cited references, our guess is that there exists at least some implementation difference (albeit unintentional) which underscores the importance of including actual code with publications. However, we did include a clarification in **Methods: Training details: T1-weighted brain extraction**

> *The architecture consists of four encoding/decoding layers with eight filters at the base layer which doubled every layer.*

And in **Methods: Training details: Deep Atropos**

> *The architecture consists of four encoding/decoding layers with 16 filters at the base layer which doubled every layer.*

[240] The statement that incompatibility with commercial use is "a common constraint characteristic of other packages" is either too strong or too vague, depending on how it is read. While many packages choose viral copy-left licenses preventing commercial use, others do not, e.g. FreeSurfer, FastSurfer, fMRIPrep and Nobrainer. This should be toned down or made specific.

Sure. The sentence now reads:

> *And given the open-source nature of the ANTsX software, usage is not limited, for example, to non-commercial use—a common constraint characteristic of other packages such as the FMRIB Software Library (https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/Licence).*

[260-63] While the comment about the public availability of DeepSCAN may be justified in itself, it is irrelevant to the authors' argument about the comprehensiveness of ANTsX, which appears to be the main point of the paragraph. I would recommend removing the following wording: "For the reader interested in reproducing the authors' results, they are primarily prevented from doing so due, as far as we can tell, to the lack of the public availability of the DeepSCAN software. However, in addition, [...] also" (leaving the "[...]" intact).

Done.

[275] Please discuss limitations and possible drawbacks of using the ANTsXNet cortical-thickness pipelines.

We added the following to the Results section:

> Although potentially advantageous in terms of such issues as computational efficiency and other performance measures, there are a number of limitations associated with the ANTsXNet pipeline that should be mentioned both to guide potential users and possibly motivate future related research. As is the case with many deep learning models, usage is restricted based on training data. For example, much of the publicly available brain data has been anonymized through various defacing protocols. That is certainly the case with the training data used for the ANTsXNet pipeline which has consequences specific to the brain extraction step which could lead to poor performance. We are currently aware of this issue and have provided a temporary workaround while simultaneously resuming training on whole head data to mitigate this issue. Related, although the ANTsXNet pipeline performs relatively well as assessed across lifespan data, performance might be hampered for specific age ranges (e.g., neonates), whereas the traditional ANTs cortical thickness pipeline is more flexible and might provide better age-targeted performance. This is the subject of ongoing research. Additionally, application of the ANTsXNet pipeline would be limited with high-resolution acquisitions. Due to the heavy memory requirements associated with deep learning training, the utility of any resolution greater than ~1 mm isotropic would not be leveraged by the existing pipeline. However, there is a potential pipeline variation (akin to the longitudinal variant) that would be worth exploring where Deep Atropos is used only to provide the priors for a subsequent traditional Atropos segmentation on high-resolution data.

[410] The word "Weights" requires "were", in the plural form.

Done.