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.

Editorial Board Member:

Please consider all the reviewers' comments in revising and restructuring your manuscript. May I also suggest to keep in the text the parts of the criticism of the existing work that are directly related to the proposed method, and to move the rest to a footnote.

In our original write-up we placed the criticism of existing work in a footnote as we think that, while it is important, it is not the main focus of the manuscript. However, this footnote was moved to the main text upon reading the submission guidelines for Nature: Scientific Reports[1]*, specifically where it reads: "Please note, footnotes should not be used." Please advise as we are more than willing to accommodate.*

Reviewer 1:

The authors are the principal developers of an ecosystem (ANTs) for the processing of structural brain images. Recently they have provided packages for accessing ANTs methods inside R (ANTsR) and Python (ANTsPy), with corresponding packages to allow TensorFlow/Keras models to be trained on ANTs images.

The current paper describes a recent adaptation of an existing ANTs method (Cortical Thickness) to use deep-learning-based segmentation rather than the author's own intensity-plus-prior method (Atropos).

Summary:

I recommend a thorough rewrite of this paper, focusing on the scientific content, before publication can be considered. Excessive material on unrelated ANTs modules must be removed. Unfounded 'editorializing' on how other researchers should perform their work should be removed. Experiments must be updated to directly compare to recent versions of existing tools. Results of proposed segmentation algorithms (the contribution of the paper) must be presented.

We appreciate the comments of the Reviewer and hope that our self-initiated changes (e.g., the introduction of an ANTsXNet longitudinal variant) and the edits based on specific

[1] *https://www.nature.com/srep/author-instructions/submission-guidelines*

recommendations from the Reviewers have improved the manuscript. The manuscript has been reorganized as follows:

- The ANTsX ecosystem: A brief overview
  - Image registration origin
  - Current developments
  - The ANTsXNet cortical thickness pipeline
- Results
  - The original ANTs cortical thickness pipeline
  - Overview of cortical thickness via ANTsXNet
  - Cross-sectional performance evaluation
  - Longitudinal performance evaluation
- Discussion
- Methods
  - Implementation
  - Training details
    - T1-weighted brain extraction
    - Deep Atropos
    - Desikan-Killiany-Tourville parcellation

Issues with scientific content:

The scientific content of the paper consists of the description of the deep-learning modules on which the new method is based, description of their training, and an assessment of the performance of cortical thickness based on these new segmentation methods. For this assessment the authors reproduce two experiments from previous work: [20] provides an experiment purporting to gauge the performance of cross-sectional performance, by predicting age and gender from regional cortical thickness measures, while [33] provides an assessment of longitudinal performance based on the ability to distinguish degrees of cognitive impairment, as well as the ratio between residual and between-subject variability. The new method performs comparably-to-slightly-better when assessed cross-sectionally and in distinguishing cognitive impairment, but dramatically worse in the ratio between residual and between-subject variability.

We appreciate the Reviewer’s synopsis of the submitted manuscript.

It is unclear if the authors intend this to be a paper about segmentation algorithms or a cortical thickness algorithm. Despite the fact that the only original contribution of the paper is a collection of segmentation algorithms, no attempt is made to present or assess the performance of these algorithms themselves, except via the cortical thickness algorithm. This is in particular problematic regarding the 'deep Atropos' algorithm, since the authors themselves have claimed in the past that the quality of their DiReCT cortical thickness method depends heavily on segmentation quality. I would therefore have appreciated at a minimum a statement of Dice coefficients per region, and preferably also some images showing representative performance from among the datasets used for testing, comparing performance of the deep and non-deep Atropos methods. An alternative approach would be to focus on the performance of the DiReCT algorithm using different segmentation algorithms as input, but this would warrant the inclusion of segmentations from algorithms outside the ANTs framework, and would also entail a comparison of the segmentation results themselves. Similarly (since authors who use the existing ANTs framework may want to know if they need to re-run their analysis) a comparison of the cortical thickness measures between the old and new cortical thickness pipelines would be sensible.

The Reviewer is correct in that we use cortical thickness values as the primary assessment measurement for determining comparative performance. The Reviewer is also correct in that the principal changes to the deep learning-based cortical thickness pipeline involve segmentation (i.e., brain extraction, deep Atropos, and a DKT labeling protocol). We find that the evaluation strategy employed in the current manuscript yields advantages over traditional segmentation algorithmic assessment strategies (e.g., regional Dice coefficients). First, as this is a manuscript describing the ANTsX ecosystem, our intended audience are current and potential users of ANTsX. As such, the primary use case for our segmentation algorithms are through applications such as the cortical thickness pipeline and not with the segmentation algorithms directly. In other words, such readers overwhelmingly interface with the cortical thickness measurements as opposed to the regional volumes or surface areas produced by the segmentations. Second, as we have pointed out in previous publications [e.g. [2]], we find biomarker-based evaluations more convincing as to algorithmic relevance/utility over other measurements such as the Dice coefficient, especially given biomarker measurement independence[3] and the lack of availability of relevant ground truth segmentations of sufficient quality[4]. Finally, the Reviewer should be aware of the hard limits imposed by the submission guidelines of Scientific Reports[5] *and note that such additional experimental investigations would put us over those limits.*

It would also be sensible to compare CT measures yielded by the proposed method to those derived from the *latest* version of Freesurfer.

[2] *https://pubmed.ncbi.nlm.nih.gov/24879923/*

[3] *By "independence", we are contrasting with the practice of comparing proposed measurements with established measurements (of limited and unknown accuracy) used to gauge performance. For example, in the FastSurfer paper, Section 3.1 (titled "Accuracy"), the authors use FreeSurfer segmentations as the reference segmentation upon which are based the Dice coefficients for certain regions. While we certainly do not find anything inherently flawed in this approach, we find such an evaluation informatively limited given the relatively small measurements involved (i.e., 2-4 mm for cortical thickness over the typical image-based sampling of ~1 mm^3).*

[4] *Given the relatively small measurements involved and the variability associated with manual segmentations, labeling existing data collections as sufficient for precise measurements is not without controversy.*

[5] *https://www.nature.com/srep/author-instructions/submission-guidelines*

As FastSurfer experiences increased uptake within the research community and becomes an established tool for cortical thickness measurement, we will most likely perform such a comparison. However, given its lack of compatibility with FreeSurfer versions > 6.0[6], the fact that current FastSurfer performance with the established FreeSurfer pipeline is quite similar (in terms of accuracy), and that we are not making any computational efficiency claims relative to FastSurfer, we choose not to include it in this work.

The authors claim 'significantly improve[d] computational efficiency', but do not quantify this improvement in terms of time to yield cortical thickness measurements. This is in contrast to, for example, Fastsurfer [35], where real, substantial improvements in speed were yielded, not only because the segmentation algorithm was faster, but because additional steps in the algorithm were rendered unnecessary by improvements in the segmentation quality.

Although we previously provided an estimate of computational time in the Methods section ("Computation time on a CPU-only platform is ~1 hour primarily due to the ants.kelly_kapowski function."), we expand our discussion in the revised manuscript where we explicitly outline the steps which are rendered obsolete within the deep learning workflow while providing additional timing information. In the introductory section, we include the following figure:

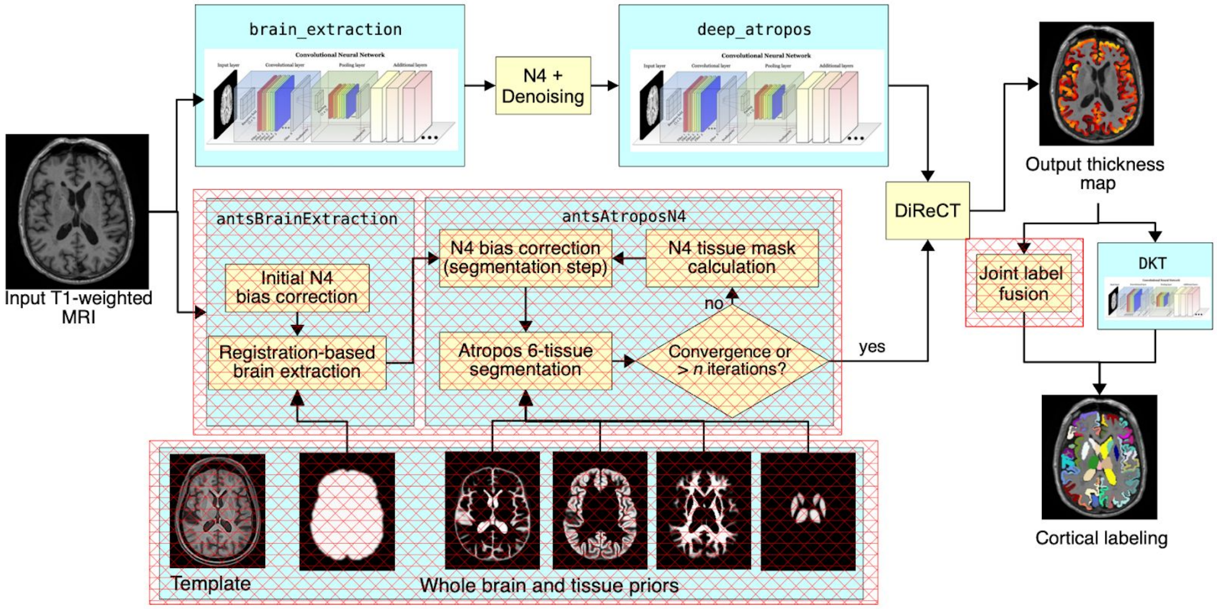

brain_extraction
N4 +
Denoising
deep_atropos
Output thickness map
Input T1-weighted MRI
antsBrainExtraction
antsAtroposN4
Initial N4 bias correction
Registration-based brain extraction
N4 bias correction (segmentation step)
N4 tissue mask calculation
no
Atropos 6-tissue segmentation
Convergence or > n iterations?
yes
DiReCT
Joint label fusion
DKT
Cortical labeling
Template
Whole brain and tissue priors

Figure 2: Illustration of the ANTsXNet cortical thickness pipeline and the relationship to its traditional ANTs analog. The hash-designated sections denote pipeline steps which have been obviated by the deep learning approach. These include template-based brain extraction, template-based *n*-tissue

[6] *https://github.com/Deep-MI/FastSurfer (retrieved January 2, 2021).*

> segmentation, and joint label fusion for cortical labeling.

with the associated text:

> The most recent ANTsX innovation involves the development of a deep learning analog of our popular ANTs cortical thickness cross-sectional[16] and longitudinal[29] pipelines within the ANTsXNet framework for, amongst other potential benefits, increased computational efficiency. Figure 2, adapted from our previous work16 illustrates some of the major changes associated with the single-subject pipeline. It should be noted that this improvement in efficiency is principally a result of eliminating deformable image registration from the pipeline—a step which has historically been used to propagate prior, population-based information (e.g., tissue maps) to individual subjects for such tasks as brain extraction[30] and tissue segmentation[10] which is now configured within the neural networks.

In addition, we have augmented the previous timing information in the ***Methods*** *section:*

> Computational time on a CPU-only platform is approximately 1 hour primarily due to ants.kelly_kapowski processing. Other preprocessing steps, i.e., bias correction and denoising, are on the order of a couple minutes. This total time should be compared with 4 − 5 hours using the traditional pipeline employing the quick registration option or 10 − 15 hours with the more comprehensive registration parameters employed). As mentioned previously, elimination of the registration-based propagation of prior probability images to individual subjects is the principal source of reduced computational time. For ROI-based analyses, this is in addition to the elimination of the optional generation of a population-specific template. Additionally, the use of antspynet.desikan_killiany_tourville_labeling for cortical labeling (which completes in less than five minutes) eliminates the need for joint label fusion which requires multiple pairwise registrations for each subject in addition to the fusion algorithm itself.

One could imagine that bias field correction and even brain extraction are no longer necessary, given a high performing deep neural network trained on 1200 cases, but other than segmentation the authors leave the rest of their algorithm intact.

In our initial foray to deep learning applications, we believed similarly that many, if not all preprocessing steps, would be rendered obsolete. However, this has not been our experience, generally. Indeed, it seems to not be the experience of many groups which continue to perform various preprocessing steps as part of their deep learning workflows, including bias correction and brain extraction.

This is surprising, in light of the author's criticism that 'fine-tuning is often omitted or not considered when other groups use components of our cortical thickness pipeline'. Apparently no fine tuning was necessary here, but this is not mentioned.

Fine-tuning was necessary in our experience. Indeed we direct the Reviewer's attention to the sentence immediately following the quoted sentence:

> This fine-tuning is often omitted or not considered when other groups use components of our cortical thickness pipeline which can be potentially problematic. *Fine-tuning for this particular workflow was also performed between the first and last authors using manual variation of the weights in the weighted categorical cross entropy.* (emphasis added)

We further emphasized this point in the revised manuscript by adding:

> Specifically, the weights of each tissue type was altered in order to produce segmentations which most resembled the traditional Atropos segmentations.

This was similar to the development of the original cortical thickness pipeline where hyperparameters were varied to produce quality segmentations, according to the first and last authors who tend to work closely together on the development of these ANTs pipelines.

Assessment of the performance of the new cortical thickness their method by using repeating previously published experiments would be an excellent decision if the new pipeline were completely independent of the old, but this is not the case: the algorithm for calculating thickness is identical, and the new segmentation algorithm is trained to reproduce the output of the old. This may entail a form of circularity: by training a method on their existing Atropos method, and repeating experiments on which their Atropos+DiReCT method has previously outperformed other algorithms, it is perhaps not surprising that the new method also performs well.

*In mentioning "circularity", we assume the reviewer is referring to forms of selection bias described by Kriegeskorte[7]. It should be noted that valid criticisms of flawed methodology based on such circularity also entail consideration of the assumptions made for hypothesis testing (formal or informal). *If* we had proposed a generalized deep learning framework whereby we claim to being capable of producing superior results from training data derived from segmentations from any number of established pipelines, *and if* we only used Atropos results to train our model, then, yes, the Reviewer would have a point that such an analysis (with its underlying hypothesis-based assumptions) would constitute a form of circularity. However, that is not what is claimed. It could very well be that if one were to use FreeSurfer, FSL, or other pipelines (or some combination thereof) to produce training data, it is certainly possible that the resulting model could outperform the proposed model based on neutral, biomarker-based criteria. Instead, we used Atropos to produce training data for our segmentation model due to the simple fact that it is an ANTs-based tool (for an ANTsX-based manuscript), that it has an established track record of performance, and that it is known to work well with the DiReCT algorithm.*

The authors attribute the poor performance of the ratio between residual and between-subject variability as a property of deep networks in general. This is a big stretch: to justify this claim

[7] *https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2841687/*

would require demonstrating that this effect persists over different models (also, ideally, trained over different segmentations than just Atropos). Also I do not understand the claim that this effect has something to do with the bias-variance tradeoff, which refers to variance over the choice of training set and has nothing per-se to do with variance between or within observations of individuals.

With the inclusion of the longitudinal variant of the ANTsXNet pipeline, we removed the section in question.

Issues with the Introduction and Discussion:

In general the majority of the introduction and discussion, as well as the title and abstract, have little to do with the scientific content of the paper. A bit of context is fine, of course, and the authors should be commended for building a heavily used suite of tools, but the number of citations to previous modules (Table 1) are not relevant to the performance of the cortical thickness module. In addition, the repeated emphasis on the high performance of registration algorithms such as SyN included in ANTs and the emphasis on tight integration in ANTs may give readers the false impression that the DiReCT CT algorithm makes use of these methods, while in fact the diffeomorphic registration algorithm used by DiReCT is completely separate from those methods. The title and abstract must reflect the scientific content of the article.

We suggest that the reorganization of the current manuscript from the previous version should mitigate confusion concerning manuscript scope. The focus of the paper is not the cortical thickness module, per se, but rather it is about the ANTsX software ecosystem of which the newly developed deep learning cortical thickness module is only a recent addition. And it is this addition which is used to showcase certain novel aspects of the toolkit---hence, the title referencing the software ecosystem and the relevant portion of the abstract which reads:

> The Advanced Normalizations Tools ecosystem, known as ANTsX, consists of multiple open-source software libraries which house top-performing algorithms used worldwide by scientific and research communities for processing and analyzing biological and medical imaging data. The base software library, ANTs, is built upon, and contributes to, the NIH-sponsored Insight Toolkit. Founded in 2008 with the highly regarded Symmetric Normalization image registration framework, the ANTs library has since grown to include additional functionality. Recent enhancements include statistical, visualization, and deep learning capabilities through interfacing with both the R statistical project (ANTsR) and Python (ANTsPy). Additionally, the corresponding deep learning extensions ANTsRNet and ANTsPyNet (built on the popular TensorFlow/Keras libraries) contain several popular network architectures and trained models for specific applications. *One such comprehensive application is a deep learning analog for generating cortical thickness data from structural T1-weighted brain MRI.* Not only does this significantly improve computational efficiency and provide comparable-to-superior accuracy over multiple criteria relative to the existing ANTs pipelines but it also illustrates the

> importance of the comprehensive ANTsX approach as a framework for medical image analysis. (emphasis added)

The authors make much in the discussion of the benefits of 'comprehensiveness' and 'interoperability' of ANTs. I can well believe that this makes it convenient for some users, but the authors do not make a substantial case that this is the best or only way to build neuroimaging pipelines. Indeed, the popular nipype python package exists precisely because many researchers want to "mix and match" between, for example, SPM, FSL, Freesurfer and ANTs. Similarly, while tight integration of neuroimaging and deep learning may be convenient for some, it is by no means superior to working directly, as most researchers do with Tensorflow or Pytorch.

It is a bit difficult to follow the Reviewer's criticisms here in the sense that they would require a response. The Reviewer accepts that some users find the comprehensiveness and interoperability of ANTs "convenient"---with which we do not disagree. For example, in a recent ANTs-based NIH grant submission, we solicited letters from the medical research community and within two weeks we had well over 100 letters of support attesting to, in part, the convenience of ANTs. The Reviewer's follow-up criticisms of "the authors do not make a substantial case that this is the best or only way to build neuroimaging pipelines" and "[ANTsX] is by no means superior to working directly, as most researchers do with Tensorflow or Pytorch" is absolutely correct. We did not (nor did we intend to) say anything to the contrary. As important as we think ANTsX is to the research community (including our own research), we fully recognize that it is a collection of processing and measurement tools which complements other pipelines/approaches, all of which have their own merits and weaknesses.

Reviewer 2:

The manuscript reviews the ANTsX ecosystem for registration, segmentation, template building, label fusion and more, with a main focus on the reimplementation of the ANTs cortical thickness pipeline within the ANTsXNet deep-learning (DL) framework. Although the text is very well and carefully written at the micro level, its overall structure is confusing.

About one third of the introduction is about the (impressive) achievements of the SyN registration algorithm, but the remainder of the paper is not about registration. Half of the results section evaluates the pipeline in longitudinal data, although ANTsXNet includes no longitudinal considerations (and the results are mixed).

The section headings are misleading, beginning with results, then discussion, then methods. Although there may sometimes be good reason to do this, here, the "results" section presents a mixture of methods and results, and the "methods" section includes a code snippet with an instructive explanation. The traditional methods-results-discussion sections may not lend itself to a toolbox review paper. In addition, including these sections here gives the impression of a standard methodological contribution, which adds to the confusion.

For a toolbox paper, the authors are advised to deviate completely from the methods-results-discussion sections and use more specific headings reflecting their actual content.

We agree. This manuscript was originally prepared for the new journal Nature: Computational Science *which has as one of its stated Scope and Aims "Methods, Tools and Platforms for Computational Science*[8]*." We specifically targeted the manuscript as a* Resource *content type ("A Resource is typically used to report on and present a large dataset, a new tool, or a new framework of broad interest for the computational science community")*[9] *which has specific formatting instructions:*

> An introduction (without heading) is followed by sections headed Results, Discussion and Methods. The Results and Methods should be divided by topical subheadings; the Discussion does not contain subheadings. As a guideline, Resources allow up to 50 references.

The pre-screening of the manuscript referred us to Nature: Scientific Reports. *In our follow-up preparation of the submitted manuscript under review, we followed the suggested organization*[10]*:*

> For the main body of the text, there are no specific requirements. You can organise it in a way that best suits your research. However, the following structure will be suitable in many cases:
>
> Introduction
> Results (with subheadings)
> Discussion (without subheadings)
> Methods

as it conformed with our previous submission and deferred any potential formatting changes to the requests of the Editors/Reviewers.

However, given the flexibility afforded by Scientific Reports, the Reviewers' recommendations, and our own preferences, we have significantly altered the organization as follows:

> The ANTsX ecosystem: A brief overview
> > Image registration origin
> > Current developments
> > The ANTsXNet cortical thickness pipeline
>
> Results
> > The original ANTs cortical thickness pipeline
> > Overview of cortical thickness via ANTsXNet

[8] *https://www.nature.com/natcomputsci/about/aims*

[9] *https://www.nature.com/natcomputsci/about/content*

[10] *https://www.nature.com/srep/author-instructions/submission-guidelines*

Cross-sectional performance evaluation
Longitudinal performance evaluation
Discussion
Methods
Implementation
Training details
T1-weighted brain extraction
Deep Atropos
Desikan-Killiany-Tourville parcellation

Further specific points:

1) Line/page numbers would have been appreciated.

Done.

2) There is no introduction heading.

Fixed. See above for explanation.

3) The presented achievements of SyN span the years 2010-2014, but later years are not mentioned. In addition, DL developments are discussed in various sections but not for registration applications. This discussion is somewhat long for a paper that is not about registration. However, for a fairness, outperformance of SyN by competing DL methods should be acknowledged, e.g. deepregnet/UCL, de Vos/Utrecht, voxelmorph/MIT.

Reviewing the original manuscript, we find that too much space was spent on ANTs image registration thus agreeing with the reviewer that "This discussion is somewhat long for a paper that is not about a registration." As a result, we removed the majority of the registration discussion. The following constitutes the entirety of the image registration historical introductory section:

> ### Image registration origins
>
> The Advanced Normalization Tools (ANTs) is a state-of-the-art, open-source software toolkit for image registration, segmentation, and other functionality for comprehensive biological and medical image analysis. Historically, ANTs is rooted in advanced image registration techniques which have been at the forefront of the field due to seminal contributions that date back to the original elastic matching method of Bajcsy and co-investigators[1–3]. Various independent platforms have been used to evaluate ANTs tools since their early development. In a landmark paper[4], the authors reported an extensive evaluation using multiple neuroimaging datasets analyzed by fourteen different registration tools, including the Symmetric Normalization (SyN) algorithm[5], and found that "ART, SyN, IRTK, and SPM's DARTEL Toolbox gave the best results according to overlap and distance measures, with ART and SyN delivering the most consistently high accuracy across subjects and label sets." Participation

> in other independent competitions[7,8] provided additional evidence of the utility of ANTs registration and other tools. Despite the extremely significant potential of deep learning for image registration algorithmic development[9], ANTs registration tools continue to find application in the various biomedical imaging research communities.

4) ANTs is vetted by users and developers from all over the world: some numbers showing the downloads/users/contributions and their evolution over time would both be interesting and back this claim (other than publications).

Sure. We added the following text in the introductory section:

> According to GitHub, recent unique "clones" have averaged 34 per day with the total number of clones being approximately twice that many. 50 unique contributors to the ANTs library have made a total of over 4500 commits. Additional insights into usage can be viewed at the ANTs GitHub website.

5) ANTs facilitated workflows which were not previously possible: please explain how.

We replaced "which were not previously possible" with "which leverage the advantages of these high level programming languages."

6) The ANTs cortical thickness pipeline is CURRENTLY available within the ANTsXNet libraries: this suggests is it going to be removed?

We recognize that this sentence could be read as implying removal in the future. However, that is not how we intended it to be read. As opposed to many methodology-based articles which simply describe their contributions (with varying degrees of detail) and only occasionally suggest public availability (oftentimes only once the manuscript is published), we make our tools publicly available even prior to publication, as is the case here. That is what is meant by the use of "currently."

7) DTK is conducive to voxel-based analyses: please briefly explain the DTK protocol.

The Desikan-Killiany-Tourville (DKT) labeling protocol is a cortical parcellation scheme. We briefly explain this in the Results section:

> Although the resulting thickness maps are conducive to voxel-based[38] and related analyses[39], where we employ the well-known Desikan-Killiany-Tourville (DKT)[40] labeling protocol (31 labels per hemisphere) to parcellate the cortex for averaging thickness values regionally. This allows us to 1) be consistent in our evaluation strategy for comparison with our previous work[20,33] and 2) leverage an additional deep learning-based substitution within the proposed pipeline.

As the DKT protocol provides an anatomically-based regional parcellation of the cortex where each region comprises multiple voxels, this would exclude the type of voxel-based analysis to which we refer in the above paragraph.

8) A SINGLE model was created for each of these steps: this is unclear. Is it a single model for all of these steps or separate models for each of these steps?

Clarification is provided in the previous sentence:

> *Brain extraction employs a traditional 3-D U-net network[28] with whole brain, template-based data augmentation[45] whereas brain segmentation and DKT parcellation are processed via 3-D U-net networks with attention gating[46] on image octant-based batches.* We emphasize that a single model was created for each of these steps and was used for all the experiments described below. (emphasis added)

We were making the distinction with ensemble approaches where multiple models are used to derive a single solution. For further clarification, we edited the above paragraph to include this distinction:

> Brain extraction employs a traditional 3-D U-net network[28] with whole brain, template-based data augmentation[45] whereas brain segmentation and DKT parcellation are processed via 3-D U-net networks with attention gating[46] on image octant-based batches. We emphasize that a single model (as opposed to ensemble approaches where multiple models are employed, e.g.,[31]) was created for each of these steps and was used for all the experiments described below.

9) Processing data through components of the pipeline that were used to train the same pipeline: how is this the case given the presented cross-validation strategy, meaning training on part of the data and evaluating on another? Are the authors referring to hyperparameter choices?

Perhaps further details would clarify what was meant. The entire ~1200-subject training cohort (IXI, MMRR, NKI, and Oasis) was used in a typical model fitting scenario explicitly supported by the Keras framework. We used a conventional 80/20 split where an initial random selection of 80% of the ~1200 subjects were used for model fitting and 20% were used as "validation_data" in the context of the keras.fit_generator function. All ~1200 subjects were processed through the resulting ANTsXNet pipeline for comparison with the traditional ANTs-based pipeline. Given the well-known loose analogy of the encoder-decoder structure with certain dimensionality reduction strategies (e.g., PCA), this can be seen as an additional step in a combined processing pipeline. We thought it might be of interest to the reader to report these results, in addition to reporting results on a completely separate data set (SRPB)---the latter providing evidence of performance generalizability.

10) The heading "longitudinal cortical thickness" is slightly misleading as the ANTsXNet pipeline is not specifically tailored to longitudinal analysis.

Note that we added a longitudinal variant of the ANTsXNet pipeline but have also completely reorganized the current manuscript as mentioned previously to alleviate confusion.

11) Superior computational efficiency of the ANTsXNet pipeline: please compare typical runtimes to the classical pipeline.

Although we previously provided an estimate of computational time in the Methods section (“Computation time on a CPU-only platform is ~1 hour primarily due to the ants.kelly_kapowski function.”), we expand our discussion in the revised manuscript where we explicitly outline the steps which are rendered obsolete within the deep learning workflow while providing additional timing information. In the introductory section, we include the following figure:

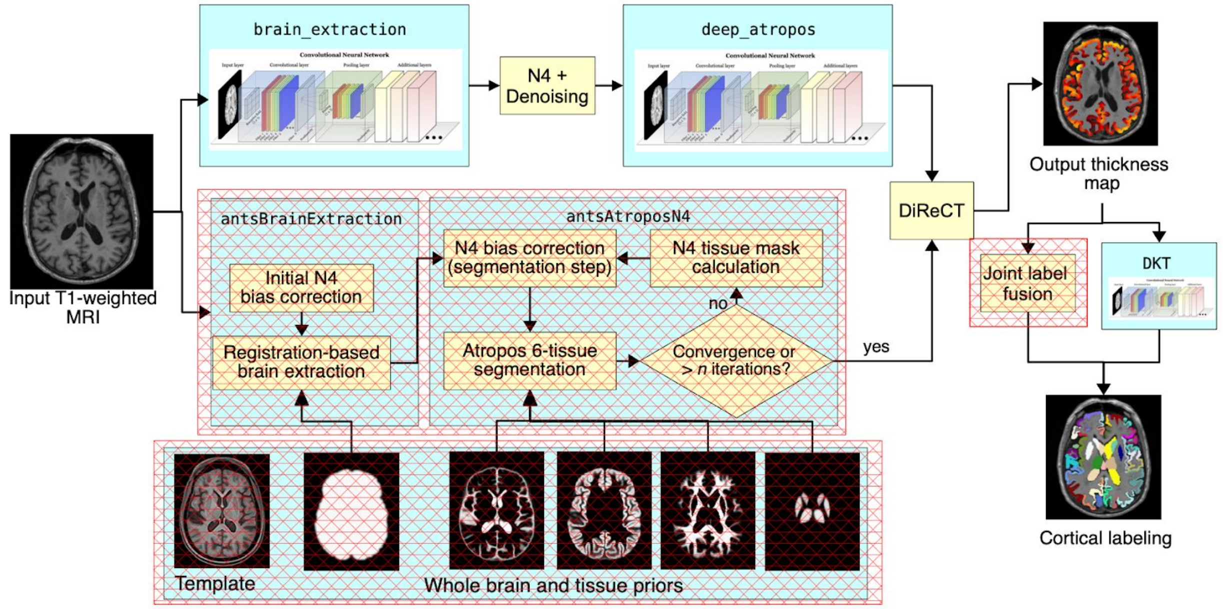


Figure 2: Illustration of the ANTsXNet cortical thickness pipeline and the relationship to its traditional ANTs analog. The hash-designated sections denote pipeline steps which have been obviated by the deep learning approach. These include template-based brain extraction, template-based n-tissue segmentation, and joint label fusion for cortical labeling.

with the associated text:

> The most recent ANTsX innovation involves the development of a deep learning analog of our popular ANTs cortical thickness cross-sectional[16] and longitudinal[29] pipelines within the ANTsXNet framework for, amongst other potential benefits, increased computational efficiency. Figure 2, adapted from our previous work16 illustrates some of the major changes associated with the single-subject pipeline. It should be noted that this improvement in efficiency is principally a result of eliminating deformable image registration from the pipeline— a step which has historically been used to propagate prior, population-based information (e.g., tissue maps) to individual subjects for such tasks as brain extraction[30] and tissue segmentation[10] which is now configured within the neural networks.

In addition, we have augmented the previous timing information in the **Methods** *section:*

> Computational time on a CPU-only platform is approximately 1 hour primarily due to the ants.kelly_kapowski function. Other preprocessing steps, bias correction and denoising, are on the order of a couple minutes. This total time should be compared with 4 – 5 hours using the traditional pipeline employing the quick registration option or 10 – 15 hours with the more comprehensive registration parameters employed). As mentioned previously, elimination of the registration-based propagation of prior probability images to individual subjects is the principal source of reduced computational time. For ROI-based analyses, this is in addition to the elimination of the optional generation of a population-specific template. Additionally, the use of antspynet.desikan_killiany_tourville_labeling, for cortical labeling (which completes in less than five minutes) eliminates the need for joint label fusion which requires multiple pairwise registrations for each subject in addition to the fusion algorithm itself.

12) This is provided by the following LME model: please define the abbreviation before using it (the definition is on the next page).

Done.

13) We prefer to see lower residual variability and higher between-subject variability, leading to a LARGER variance ratio: equation 4 defines r as within-subject residuals / between-subject variability. I would suggest that lower residuals with higher between-subject variability decrease r?

We mistakenly swapped the symbols for residual- and between-subject variabilities. This has been fixed and the plots should make sense now.

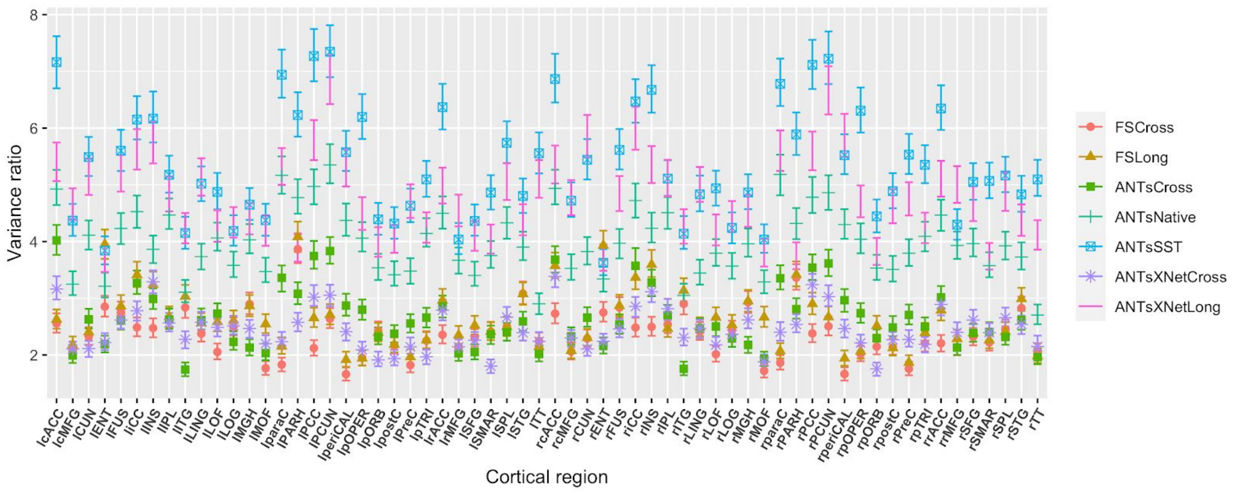
14) Figure 4(b): the right-most panel appears to show 1/r instead of the variance ratio r. To avoid confusion, it would be helpful to plot r and specifically say "r" in the subfigure title or caption.

See previous item.

15) Would it be possible to provide longitudinal plots for specific brain regions? It would be interesting to see if the differences are global or spatially localized.

We agree that such information is interesting. In fact, there are at least two separate manuscripts in preparation examining in greater detail the longitudinal-based performance of these pipelines. Instead of longitudinal plots (which can be easily reconstructed using data and a spaghetti plot script available in one of the specified GitHub repositories[11]), we summarize the regionally-specific performance information using the following figure which has been added to the revised manuscript:

[11] *https://github.com/ntustison/CrossLong/blob/master/Scripts/Analysis/plotSpaghetti.R*



16) ANTs at the forefront of image registration: ANTs is great but again, what about competing DL methods?

Please see our response above to item 3.

17) Comprehensiveness of ANTsX vs DeepSCAN: it is unclear why the authors compare the comprehensiveness of ANTsX to DeepSCAN, as they claim the key point preventing repeatability is the public unavailability of this algorithm.

The "comprehensiveness of ANTsX" is in no way meant to be a direct contrast with DeepSCAN. We rewrote this section as follows:

> It is the comprehensiveness of ANTsX that provides significant advantages over much of the deep learning work that is currently taking place in medical imaging. In other words, various steps in the deep learning training processing (e.g., data augmentation, preprocessing) can all be performed within the same framework where such important details as header information for image geometry are treated the same. In contrast, related work[32] described and evaluated a similar thickness measurement pipeline. However, due to the lack of a complete processing and analysis framework, training data was generated using the FreeSurfer stream, deep learning-based brain segmentation employed DeepSCAN[51] (in-house software), and cortical thickness estimation[16] was generated using the ANTs toolkit. 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, the interested reader must also ensure the consistency of the input/output interface between packages (a task for which the Nipype development team is quite familiar.)

18) While it may be cumbersome to install or compile various packages, please explain what is meant by interoperability issues and how this may inhibit usage.

This was in reference to secondary concerns associated with previous work. Due to its ambiguity, we chose to remove this sentence. However, interoperability, in general, is a concern across software packages. For example, the way different packages utilize the Nifti header (e.g., ITK vs. SPM) has caused interoperability issues for ANTs (and SPM) users. Keeping everything within a single framework tends to mitigate such problems.

19) Comparable-to-superior longitudinal performance depending on evaluation metric: actually, performance is also comparable-to-worse, depending on the evaluation metric. Please be specific.

We added additional details at the end of the **Discussion** *section:*

> Performance is mostly comparable-to-superior relative to existing pipelines depending on the evaluation metric. Specifically, the ANTsXNet cross-sectional pipeline does well for the age prediction performance framework and in terms of the ICC. Additionally, this pipeline performs relatively well for longitudinal ADNI data for disease differentiation but not so much in terms of the generic variance ratio criterion. However, for such longitudinal-specific studies, the ANTsXNet longitudinal variant performs well for both performance measures. We see possible additional longitudinal extensions incorporating subject ID and months as additional network inputs.

20) T1-weighted brain extraction: it is unclear what the outputs of the network are and what is being compared in the loss function.

We clarified by adding to the subsection on the T1-weighted brain extraction process:

> **T1-weighted brain extraction.** A whole-image 3-D U-net model[22] was used in conjunction with multiple training sessions employing a Dice loss function followed by categorical cross entropy. *Training data was derived from the same multi-site data described previously processed through our registration-based approach[31]*. A center-of-mass-based transformation to a standard template was used to standardize such parameters as orientation and voxel size. However, to account for possible different header orientations of input data, a template-based data augmentation scheme was used[39] whereby forward and inverse transforms are used to randomly warp batch images between members of the training population (followed by reorientation to the standard template). A digital random coin flipping for possible histogram matching[50] between source and target images further increased data augmentation. *The output of the network is a probabilistic mask of the brain*. Although not detailed here, training for brain extraction in other modalities was performed similarly. (emphasis added to denote revisions).

21) Further increased possible data augmentation: why possible augmentation? Was randomized histogram matching used for training or not?

We removed the word "possible" to remove the ambiguity that random histogram matching was used for data augmentation purposes.

22) Other hyperparameters: what other hyperparameters are there and why are they not reported?

We listed the major hyperparameters in the paper consistent with the related literature. We understand that, for purposes of reproducibility, having access to the totality of information is ideal (and necessary). However, instead of trying to exhaustively list every hyperparameter used in the different architectures for the proposed pipeline, we agree with Linus Torvalds' dictum of "show me the code" and actually point the reader to where they can find the set of hyperparameters in a context that is most informative---in the actual code. If the Reviewer is looking for a set of specific hyperparameters, however, that they believe would be useful to the general reader, we would be happy to also include those within the text.