

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 see the Reviewer's comments.

We read the Reviewer's comments. While we can respect Reviewer 1's preferences, it seems as though we are at the point of "agreeing to disagree." We have sincerely tried to accommodate the recommendations of Reviewer 1 but we are unable to transform the paper into something that it is not intended to be. However, we did respond to the specific suggestions pointed to by the editorial board member.

Specifically, please tone down the phrase about the "advantages over ...",

*The phrase "advantages over" appears in the **Discussion** as follows:*

It is the comprehensiveness of ANTsX that provides several 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 ecosystem where such important details as header information for image geometry are treated the same.

We have modified the text to remove the phrase:

Using ANTsX, various steps in the deep learning training processing (e.g., data augmentation, preprocessing) can all be performed within the same ecosystem where such important details as header information for image geometry are treated the same.

and also provide a statement about the comparability of segmentations between the new CNN and the old Atropos method.

*We added the following to the **Discussion** section:*

Although direct evaluation by the principal co-authors of the ANTs toolkit, the similarity in resulting cortical thickness values, as indicated by Figure 4, and considerations of the training data origins all strongly suggest similarity between Atropos and Deep Atropos output, further evaluation is certainly warranted and would benefit other potential applications.

Reviewer 1

I am afraid that I cannot accept the manuscript. The paper remains an uncomfortable mix between a framework paper and a paper about a specific application of that framework. It unfortunately performs the job of neither: the framework part does not adequately describe

much of the functionality and the "advantages over much of the deep learning work that is currently taking place in medical imaging" are not convincingly explained: indeed, in the comments to reviewers the authors admit that they do not target deep learning practitioners with their framework, but rather "researchers (who) are more concerned with using the measurements provided by established processing pipelines which take advantage of deep learning technology".

As we indicated to the editorial board member, while we can respect the Reviewer's preferences, it seems as though we are at the point of "agreeing to disagree."

I appreciate the gestures made by the authors toward fulfilling my second primary objection (the insufficient validation of the new segmentation method which is the only contribution of the paper), but the authors continue to fail to provide even a cursory statement about the comparability of segmentations between the new CNN and the old Atropos method (which could be addressed in one or two lines of text if necessary).

*Despite revisiting the previous rounds of reviews, we do not recall ever being asked for a "cursory statement about the comparability of segmentations between the new CNN and the old Atropos method (which could be addressed in one or two lines of text if necessary)." However, we are more than happy to comply with this request. We added the following to the **Discussion** section:*

Although direct evaluation by the principal co-authors of the ANTs toolkit, the similarity in resulting cortical thickness values, as indicated by Figure 4, and considerations of the training data origins all strongly suggest similarity between Atropos and Deep Atropos output, further evaluation is certainly warranted and would benefit other potential applications.

Instead the authors fixate on their own cortical thickness method as the only discriminator of segmentation quality, despite the fact that users of their method may apply it to uses beyond those they have imagined.

The suggestion of "our own cortical thickness method as the only discriminator of segmentation quality" is pretty strong and one which we do not recognize as originating from anything we wrote. We did provide reasoning as to why, in the absence of ground-truth segmentations, we prefer our approach but that is significantly different from being "the only discriminator."

I suggest that the authors separate the content of the paper into a pure framework paper (which should describe in detail the framework and its supposed advantages, with multiple use cases (ideally also from outside of the ANTSX development team) and a paper which evaluates their segmentation method in sufficient detail.

While we appreciate the suggestion, as mentioned above, we will have to agree to disagree as to the preferred method of presenting the work contained in the current manuscript.