David Kennedy, Ph.D.,
Co-Editor-in-Chief, Neuroinformatics

Dear Dr. Kennedy and Reviewers,

My coauthors and I would like to thank the editors and reviewers for handling our manuscript entitled "An Open Source Multivariate Framework for n-Tissue Segmentation with Evaluation on Public Data." The reviews were extremely helpful in highlighting both the strengths and weaknesses of the previous version of the paper. In summary, the review found the work to be worthy of publication, with revisions that focused on improving clarity of some key points. We make every effort, in this response letter and the associated revision, to clarify our work and improve the paper according to the reviewers' suggestions. We believe that we carefully revised the manuscript to address all concerns and provide responses to the reviewers' comments, which you will find below. We also elected to add detail regarding the ICM strategy available in Atropos. Thus, we would like to resubmit the revised manuscript which we hope will be suitable for publication.

The revision of the paper includes:

1. A new figure revision highlighting ICM and providing a simple starter example for Atropos usage.
2. Typo fixes in the equations.
3. Many minor changes in the body of the text.
4. The references were altered according to reviewer suggestions.
5. An information sharing statement.

We express our sincere gratitude to the editorial staff and to the reviewers for their efforts in evaluating this paper.

Sincerely,

Brian Avants, Ph.D.
Penn Image Computing and Science Laboratory
University of Pennsylvania
3600 Market Street, Ste 370
Philadelphia, PA 19104

Response To Reviewers

Reviewer 1

This paper summarised a semi-automatic segmentation tool called Atropos primarily targeting the brain. It uses EM based methods and can deal with a high number of labels. Overall it is well written, and the availability of the source code is great. Although the techniques used are not novel, the paper is primarily a description of the Atropos tool and does this very well. It gives good background information as well as explains the theoretical aspects of the work clearly.

Furthermore, I obtained the source code and compiled it successfully. Unfortunately, I could not get it to work on our in-house data before submitting this review.

Please let us know through Sourceforge (http://sourceforge.net/projects/advants/) if you run into any compilation troubles and/or troubles running the code. In addition to Atropos adhering to strict ITK coding practices and standards, we recently turned ensured that all compilation warnings using gcc 4.2.1 on Mac OS X 10.6 were completely resolved. One of our motivations for making Atropos open source is the same for most of software—interaction with and feedback from users can only make our code better and more useful to us.

Below are my comments split into two sections. The first are primarily technical whereas the second section contains grammatical corrections some of which may be ignored at the authors discretion.

We thank reviewer 1 for his/her careful reading for technical content and for helping us to improve our work.

1) Page 5, 2nd line from bottom. What are silver standard labellings?

We realize that the use of this term in distinguishing levels of confidence in a given data set might be a source of unnecessary confusion so we have eliminated its use. The relevant passage currently reads (first paragraph, page 6):

> Although Atropos may be applied to multivariate data from arbitrary modalities, we limit our evaluation to tissue classification in T1 neuroimaging in part due to the abundance of gold-standard data for this modality.

2) Page 10, Equations 8–11. These follow standard approaches. However, Equation 11 does not seem consistent with [8] and [10]. If delta(x_i,x_j)=1 for x_i=x_j, then U(x) as defined in [10] increases, and p(x) as defined in [8] decreases - hence labelling similar regions seems to be penalised. However, things seem consistent if equation 11 is changed as in the standard graph cuts approaches - e.g. see equation 3 Boykov & Jolly ICCV 2001.

We appreciate the reviewer catching this error in the equation. If $x_i = x_j$ then $V = 0$ and is 1 otherwise, as is standard in Besag's approach.

3) Page 10, Equation 11. Computing the value of U(x) only considers if neighbours are the same or not. However, as this paper deals with segmenting a large number of labels why was an approach such as that taken in Noe & Gee IPMI 2001 not considered - i.e. learning the probability of labels being neighbours and using these?

This is an interesting point and we intend to investigate this direction of research in future work. In fact, some basic functionality along these lines is available within the guts of the code but additional comparative research into the myriad partial volume estimation models that have been proposed is required to determine the best approach. Thus, we added comments in the discussion to indicate our interest in pursuing partial volume models but elect not to pursue this task for this resubmission.

4) Page 12, lines 10-14. Authors comment that applying graph cuts to more than one label is not globally optimal and that EM is a reasonable and efficient alternative. Although this is true,another property of graph cuts is that it is fast. How does the EM approach compare in terms of speed?

Fast is a relative term and we cannot comment on the speed of a method for which we do not have an optimal implementation. The main computational challenge with graph cuts is memory. Because one needs to solve multiple shortest path problems—and store the solution of each—the potential for memory bloat is high. Even in EM, memory concerns emerge when enough labels are used (e.g. in our examples). In theory, though, EM computations are embarrassingly parallel which is not a property shared by graph-based methods. As such, one would expect even an optimally implemented multi-label graph cut method to lag in speed behind an optimally implemented EM segmentation.

We would also like to point out that we have had prior experience with graph cuts (see, for example, our implementation available at http://hdl.handle.net/1926/1503 and related publications: Zhuang Song, Nicholas J. Tustison, Brian B. Avants, James C. Gee: Integrated Graph Cuts for Brain MRI Segmentation. MICCAI (2) 2006: 831-838 and Zhuang Song, Nicholas J. Tustison, Brian B. Avants, James C. Gee: Integrated Graph Cuts for Brain MRI Segmentation. MICCAI (2) 2006: 831-838) and our preference for our current EM-based approach is certainly not restricted to speed considerations. As enthusiastic

supporters of open source software, patent issues constrain the use of graph cuts where no such limitations exist for EM-based approaches.

5) Page 13, Equation 17. This looks like a multivariate equation to me. I'm also not familiar with the form of Equation 19 - why is it different from 17?'

The difference between equations 17 and 19 is that the input to the data term is a vector (rather than scalar) and the scalar variance parameter is replaced by the covariance matrix. This is a standard extension from univariate to multivariate probability.

6) Page 17. Penultimate line. "Three likelihood classes have been developed, one parametric and one ...". Should this be "Three likelihood classes have been developed, one parametric and [two] ..."?'

Yes, thank you. The fix is made.

7) Page 26 Figure 5. In the sign column the signs for the Left and Right Caudate Nucleus and Thalamus should be -ve as according to the data in the table, Atropos performed significantly worse than the majority vote.

Fixed.

8) I could not verify the validity of all the references as they were not numbered neither were they in alphabetical order.

We apologize. In preparing the original manuscript, we were unable to find a .bst for the Neuroinformatics journal so we created our own using natbib and did not use all the options to facilitate the reviewing process. The references are now numbered in the manuscript and in the bibliography section and are ordered as they appear in the manuscript.

The reviewer also mentioned not being able to locate the 19 Hammers datasets. It is unfortunate that the brain-development website is not well annotated. We suspect the reviewer looked at the IXI data which is not the Hammers data. We suggest the reviewer look at the adult atlas section of the website which, currently, contains the data we used. This is noted in the revised text.

We also followed each of the reviewer's suggestions on the structure and grammar of the text.

We appreciate reviewer 1's time and effort.

Reviewer 2

We thank reviewer 2 for his/her attention to the manuscript and overall positive comments. We made the reviewer's suggested minor changes in this revision. The following comments were addressed:

I was very pleased with the clarity of most of the article, particularly with the careful introductory explanations, demonstrations of the command-line arguments for Atropos, and the excellent review of the literature. I was also happy to see the multivariate results in Figure 2, and would have liked to have seen accompanying quantitative results. If such results were to assert the significant improvement possible when including a complementary modality (i.e., without bias), that would be a very nice contribution in itself.

We agree with the reviewer that establishing the benefit of multivariate segmentation in the presence of bias would be significant. Such exploration would probably lengthen the current manuscript beyond its intended scope. However, we did want to provide suggestive results with the single data set that we currently have which will obviously segue into a possible future publication focusing on multivariate segmentation with Atropos on specific data sets for which multi-modal data sets are essential.

I have only two concerns with the article. First, there is almost no explanation of any of the figures in the body of the text, which raises questions about the significance of the findings, such as those of Figure 4 ("The results show that the PriorProbabilityMaps with $w = 0.5$ (far right) gives the best performance for all tissues.").

The revision adds explanatory material in the body of the text associated with each figure. These revisions are printed in blue font and are sprinkled throughout the manuscript.

Second, why is it that "In this study, we labeled the brain web (sic) template very sparsely, quickly and crudely and did not expect highly accurate results." The results depending on these manual prior labels aren't informative – they don't give good results, and don't tell us whether the results would be better if better manual labels were used.

We agree with the reviewer and removed this comment and the evaluation results. We added (bottom of page 12):

> We provide this capability to allow the user to implement an interactive editing and segmentation loop. The user may run Atropos with sparse manual label guidance, evaluate the results, update the manual labels and repeat until

achieving the desired outcome. This processing loop may be performed easily with, e.g., ITK-SNAP.

On page 22: "The advancements introduced with N4 permit such an adaptive integration with Atropos." What relevant advances are introduced with respect to N3?

We added the following elaboration on the benefits of N4 in section 3.4.3 (page 21):

> N4 extends the popular nonparametric nonuniform intensity normalization (N3) algorithm [63] in two principal ways:
>
> - We replace the least squares B-spline fitting with a parallelizable alternative (which we also made publicly available in the Insight Toolkit)— the advantages being that 1) computation is much faster and 2) smoothing is not susceptible to outliers as is characteristic with standard least squares fitting algorithms.
> - The fitting algorithm permits a multi-resolution approach so whereas standard N3 practice is to select a single resolution at which bias correction occurs, the N4 framework permits a multi-resolution correction where a base resolution is chosen and correction can then occur at multiple resolution levels each resolution being twice the resolution of the previous level.
>
> Specifically, with respect to segmentation, there exists a third advantage with N4 over N3 in that the former permits the specification of a probabilistic mask as opposed to a binary mask. Recent demonstrations suggest improved white matter segmentation produces better gain field estimates using N3 [64]. Thus, when performing 3-tissue segmentation, we may opt to use, for instance, the posterior probability map of white matter at the current iteration as a weighted mask for input to N4.

Equations: Please remove punctuation following equations, because it came sometimes be distracting. For example, on pg.9, eq.1's comma might also appear to be a prime/apostrophe for the variable y.

Done except in cases where the equation ends a sentence.

Again, we thank each of the reviewers and the editor for their efforts. We truly feel that their reading of the paper was excellent and greatly improved the work.