DISCLAIMER: These comments are provided in the spirit of open science and open peer review. They are not meant to comprise any form of "traditional" peer review, wherein a judgement (neither personal nor professional) on technical quality, scientific impact and community interest is delivered. I do not intend to imply any of that. I would like these to be seen as merely open discussion and, when appropriate, suggestions for improvement if possible. Because I find this work interesting, I am offering my comments, and am excited to engage with and learn from the authors in various ways, including the interesting technical and scientific developments presented here.

Preprint Title: The ANTs Longitudinal Cortical Thickness Pipeline
From http://www.biorxiv.org/content/early/2017/07/30/170209 on Aug 10, 2017.

Background
The need for accurate estimation of cortical thickness had always been strong, owing to its sensitivity to detect disease related change and as well challenges associated with it. Automated tools like Freesurfer, CIVET and ANTs, which provide cortical thickness, requiring minimal manual QC, do enable other researchers to focus on other challenging parts in the pipeline (e.g. development of imaging biomarkers). Key parts of that workflow include but not limited to better feature extraction methods (structural covariance) or classification algorithms for early detection and/or engage in generating insights into how they change or correlate with other variables of interest, including neuropsychological assessments. Much of the sensitivity of these subsequent parts of the workflow depend on the quality and accuracy of the underlying features they are based upon. This is especially true for cortical thickness, where there is no ground truth (other than physically measuring the thickness of the cortex with a tape post-mortem) to validate the automated estimates. In that context, longitudinal methods - being intrinsic to a given subject - offer a significant advantage in reducing the estimation error as well as improving the sensitivity to detect the disease-related changes over time. Hence, this particular pipeline is of considerable interest to the community. This study presents methodical approach to the evaluation of the presented pipeline and a nice comparison to the popular Freesurfer package.

TL;DR: Major points:

1. In the absence of physical ground truth for cortical thickness for validation, should we really be maximizing inter-subject variability?
2. How effective is this longitudinal pipeline to improve the estimates of (or reduce the error) in cortical thinning? how does it compare with FS in this measure?
3. What's *truly longitudinal* about this pipeline? Generation of subject-specific template (SST)?
4. Do the results replicate on another dataset? on another disease? Esp. in a developmental dataset, where the need for longitudinal processing is even greater.

Comments

Overall, the paper is very well-written, covers the previous literature very well, analyzes a large publicly available dataset and produces and shares useful set of results to demonstrate the utility of the proposed pipeline.

I especially commend the authors on the public sharing of the software as well as useful features for many datasets (including from earlier publications).

- The introduction is great in identifying the problems and motivating the current study, although I wish they delve at least little into the challenges in the baseline thickness estimation itself, to educate a novice reader. Moreover, the second para in Page 3 could be broken into two, and expanding the first half further to describe the "relevant statistical issues" in greater detail (what they are, what did ADNI do to mitigate them), for the benefit of the reader.
- It isn't very clear to me yet what's so longitudinal about this longitudinal pipeline if the core thickness estimation still happens at a single time-point? Is the SST creation what differentiates the cross-sectional and longitudinal pipelines? What else could be leveraged in this process to improve the thickness estimation? Fitting smooth splines over time, once vertex- or ROI-wise correspondence is achieved? Would that result in more sensitive thinning measures? I do appreciate the challenges involved in the 4D-segmentation and challenges due to the irregular temporal sampling in ADNI. However, within a smaller sample with consistent temporal sampling, would ANTs be able to leverage it?
- I agree with the authors that "the utility of cortical thickness as a biomarker lies in the ability to discriminate between patient sub-populations with respect to clinical outcomes". However, that [given it is a group-level summary of differences] is not nearly sufficient justification for single-subject accurate estimation of thickness values, neither at baseline nor longitudinally. Hence, I believe more explanation and justification is needed to make the ratio of between-subject to intra-subject variability as the metric to maximize. Resources permitting, other relevant metrics [manual validation on a small subset?] need to be given due consideration also.
- The above point is further reinforced when the authors note that the longitudinal pipeline is completely agnostic to the ordering of time-points. The temporal order could be potentially important prior info that is left un-leveraged.
- Regarding the 52 cognitively normal ADNI-1 subjects chosen for the group template, how were they selected? what are their demographics? Apoe status? Are the results expected to change with a different choice of control subjects? Why 52?
- How does the above 52 atlases differ from the 20 OASIS atlases? Do their demographics different significantly from the 52 from ADNI-1?

- First para in Section 2.3.1 needs to be developed and backed up further, with citations or data. It's not immediately clear to me why should we be maximizing the proposed ratio? Given the absence of the clear characterization of population variability in thickness (except that AD patients would have thinner cortices than health controls), I am concerned such a comparison may not be clinically or algorithmically meaningful.
- Figure 5: it would be great if you share the full data comprising this figure (or point to where this is in the repository). I personally would like to see the full distribution (in terms of violin plots).
- Fig. 6: I'd suggest replacing box plots with violin plots (or something similar). The caption should be expanded to describe exactly what is being plotted here.
- I would add few more plots, to show the data in Figures 6 and 7, broken down by clinical diagnosis, which helps us to see if patterns are specific to health or disease.
- The need for Equation (3) isn't clear to me. Perhaps more data and explanation will be helpful.
- Caption for Table 2 needs to be significantly expanded to walk the reader through the table. Note what each parameter stands for. What does normalized refer to here? How is it done? I'd consider using lighter colours.
- No explanation is given for why "all three processing methods achieve roughly the same amount of total variability,"
- I am not sure if "tighter confidence intervals in calculated mean trends" translate to "greater interpretability". This point needs to be better developed or illustrated. Also, it would be helpful to unambiguously define what you mean by "interpretability" in this context.

Minor comments:

- Many sentences are too long! The abstract is full of them. I suggest breaking them down to multiple sentences, each no longer than 25 words.
- Figure 1:I don't see the need for bar plots. Tables are more helpful, so you can remember sample sizes better. You could present percentages for gender- and class-imbalances, to help the reader to get better insight into what data they are interpreting.
- Figure 2 can have all the 4 panels on the same row, which makes the trends in MMSE even more clear. Colorbar could use more colours to highlight frequent bins.
- Instead of (cf Figure 1 of [29]), you could reproduce it, or sufficiently different schematic of it.
- Figure 4: what do different colours stand for? I can guess, but they need to be labelled.
- Typo: agnostic to the; not concerning in "longitudinal pipeline is completely agnostic concerning ordering of the input time-point images"
- Some citations are incomplete: 64,

- Obtain DOIs (via Zenodo or figshare) for the data/code shared via GitHub repos, so future users can uniquely cite it. Use the above to cite them better, instead of merely providing a link.

Desirable future work:

- Do the results replicate on another similar dataset, say AIBL or PPMI etc? Or the others used in the original paper: OASIS, IXI, MMRR, NKI?
- As the authors already note, "interpolation potentially has a systematic but regionally varying effect". I would love to see the "interpolation errors" quantified under different processing streams (in mm), and compared with the actual value of thickness (in mm) in that particular area. A 3d map of vertex-wise/ROI-wise median interpolation error (grouped by clinical diagnosis) would be very interesting to see.
- As the authors note in the abstract, some of the significant benefits of the longitudinal estimation of thickness include "more consistent estimates of intra-subject measurements while retaining predictive power", and hence showing the latter in a quantitative sense (do the long. estimates result in better cross-validated accuracies in CN vs MCI or CN vs. AD classifications?) would be interesting.
- If one were to pick a different brain disorder (say MS, FTD or Aphasia), how do the sensitivity and results change?
- Given the ROIs are not small and the possibility of averaging out potentially important signal, it may be interesting to analyze how do the results change with a larger number of ROIs? Or when you subdivide these ROIs further into smaller patches? We notice an increased sensitivity in detecting MCIc and AD with that approach, either using patch-wise median thicknesses or patch-wise covariances (Raamana, 2014, Neuroimage Clinical).

It was a pleasure to read this paper and am looking forward to use the pipeline and the shared data when possible in my research.

Thanks,
Pradeep Reddy Raamana
Postdoctoral fellow,
Rotman Research Institute,
Baycrest Health Sciences,
University of Toronto,
Toronto, Canada.
crossinvalidation.com