# Tilting at Algorithmic Windmills

Nicholas J. Tustison[a,1], Brian B. Avants (combined first author, second author or senior author?)[b]

[a]*Department of Radiology and Medical Imaging, University of Virginia, Charlottesville, VA*
[b]*Penn Image Computing and Science Laboratory, University of Pennsylvania, Philadelphia, PA*

**Abstract**

Exploration of neuroscience hypotheses have been enhanced by the increased availability of high-performance computational resources, large-scale communal efforts such as the Insight Toolkit and the R statistical project and algorithmic advancements for data transformations and processing. An integral component of the vetting process for novel algorithmic techniques includes comparison with other methods previously established within the community. Public availability, including open source distribution, of established packages has significantly facilitated these comparative evaluations. However, complementing the recent set of papers pointing to serious methodological and statistical bias considerations in neuroimaging research, we point out potential issues associated with a type of measurement bias in comparative algorithmic evaluations and propose a set of guidelines for authors and reviewers to minimize this confound.

*Keywords:* open science, reproducibility

## 1. Introduction

The neuroimaging community has benefited from the proliferation of imaging software. Established packages such as SPM from the Wellcome Trust Centre for Neuroimaging [1], the FMRIB Software Library (FSL) [3], and the AFNI toolkit [2] have aided neuroimaging researchers around the world in performing complex analyses as part of ongoing neuroscience research. Parallel to this line of inquiry is continued algorithmic innovation for improvement in analysis tools.

As fellow scientists who actively participate in this type of research, we have noticed several publications in which a principle component involves the comparison of algorithms. One of our concerns is the lack of detail with which these comparisons are presented and the corresponding possibility of *instrumentation bias* [6] (considering software as a type of instrument requiring proper "calibration" for accurate measurements). In this editorial we propose an initial set of guidelines for minimal reporting of algorithmic usage to minimize such bias understanding that the discussion will hopefully elicit a much more comprehensive response from the larger community.

Although previous articles have discussed similar concerns, oftentimes within a much larger context (e.g. fMRI reporting [5]), we are particularly interested in comparative evaluations of software and provide additional discussion through experience and actual examples found within the literature. It is hoped that this commentary serves to raise awareness to both authors and reviewers to such problematic issues (not unlike other recent articles detailing additional potential sources of methodological bias [4, 8, 7]).

[1]Corresponding author: PO Box 801339, Charlottesville, VA 22908; T: 434-924-7730; email address: ntustison@virginia.edu.

## 2. Guidelines

*2.1. Parameters should be included (perhaps the precise command line call)*

Perhaps the most blatantly egregious sin is omitting

*2.2. Hassle the developers for a set of default parameters*

*2.3. Scripts used to invoke the algorithms should be posted*

*2.4. Comparisons should be performed on publicly available data*

*2.5. Corollary to the above: Resulting images should be publicly posted*

*2.6. Scripts used to create the plots should also be posted*

*2.7. When possible, consultation should be made with the original authors*

*2.8. Provide details of where the algorithm was obtained*

*2.9. Provide system details of where the algorithm was run*

*2.10. Avoid confirmation bias: Don't code up your own version particularly if one is already available*

*2.11. Co-authors should verify findings*

*2.12. Properly contextualize comparisons*

Accuracy is important but other considerations should be included such as biological plausibility. Torsten's CURT algorithm could produce an "accurate" registration but not one which is biologically plausible.

## 3. Conclusion

## 4. Acknowledgments

## 5. Notes/Thoughts

- Several kinds of bias have cropped up in the neuroimaging community recently:
  - methodological bias: Kreikesgorte
  - statistical bias: Vul2009, Tustison2012
  - What about the dead salmon fmri study—multiple comparisons correction?
  - The single subject VBM study (assuming the single subject is representative of the population).
  - overview: 9 circles of scientific hell —neuroskeptic
  - Types of measurement bias: [6]
    - Instrument bias (code as a type of instrument which needs to be calibrated correctly).
    - Expectation bias (i.e. confirmation bias)

  We are concerned with the latter two.

- How does one avoid confirmation bias? If I'm comparing my algorithm to somebody else's algorithm, I'm going to be naturally predisposed to confirmatory evidence that my algorithm is better while neglecting (most likely not with intent of doing so) evidence that disconfirms my original belief (that my algorithm is better). Richard Feynman quote (Cargo Cult Science):

  > The first principle [of science] is that you must not fool yourself–and you are the easiest person to fool. So you have to be very careful about that. After you've not fooled yourself, it's easy not to fool other scientists. You just have to be honest in a conventional way after that.

- One particular aspect of the previous item is that it needs to be highlighted if somebody codes up the algorithm to be compared themselves. Writing bug-free code is difficult. Look at how many bug fixes are provided by the ITK community on a daily basis.

- Also, is simply saying that one algorithm is better than another sufficient? Should there be some theoretical explanation for it. For example, when I wrote my DMFFD image registration paper, the comparison wasn't IRTK vs. DMFFD but rather FFD vs. DMFFD and there were theoretical reasons why performance should be better with the latter. Also, I insisted to the reviewer on not using IRTK for comparison since there are also so many other performance-related issues (type of interpolation, gradient step, metric, metric implementation, etc.) which would confound comparisons. Also, it probably helped that the theme had a more specific, more verifiable focus, i.e. DMFFD produces a more efficient energy minimizer since it acts as a preconditioner in the standard gradient descent optimization.

- Comparisons are best made with publicly available data and results of the comparisons should be made available.

- Operating system needs to be defined. Freesurfer results varied with MacOSx.

- Ideally, the authors of the new algorithm would work with the authors of the compared algorithm. Given human nature, this type of cooperation is not guaranteed. However, a minimum set of information is needed.

- Arno Klein as an example of somebody who did it right. Takes a lot of work.

- 

- The source of the package software needs to be defined. For example, N4 has been instantiated in Slicer, ANTs, and c3d (also might be implemented in one of Styner's NITRC projects). However, each might use different default parameters and have other tweaks which effects performance. In Vladimir Fonov's github repository containing various processing scripts for MINC, one can see from the history how N4 was used but then the users switched back to N3MNI (due to performance issues?). However, they used N4 out of the c3d package which the original authors (N.T./B.A.) haven't touched in three years so the parameters aren't optimal (shrink factors = 4, spline distance = 100 with 3 levels: $100 \times 50 \times 50$). Changing these parameters (which are crucial to performance) isn't accessible to the user from c3d.

- Collaborators should help verify findings by repeating a (representative subset of a) study independently, from scratch, preferably on another computer perhaps even with a different operating system. A scientist can do this him/herself as well. Personal experience suggests that this procedure will often uncover coding errors sometimes simply by forcing one to reread code or scripts again or clarify documentation.

- wiki link— most studies ignore precision. it is critical in longitudinal studies. precision is impacted by parameter choices as well as computational architecture (Freesurfer paper)

- parameters impact accuracy - e.g. clarkson's paper vs our recent tuning of DiReCT parameters w.r.t. expected thickness values.

- Biological plausibility is important—a more accurate method may produce less biologically plausible results. Cross-referencing with other fields is critical e.g. patterns of MRI atrophy should match pathology distribution. There is an art to this, of course.

- Prediction—an algorithm that yields a "more significant" result is not always better. might mean "finds more effected areas" or smaller p-values. Testing prediction complements studies of significance and cumulative distribution functions.

- Verification is supported by open science technology for sharing data and code.
- Great resources for data-sharing:
  - figshare
  - MIDAS
  - slicedrop
  - R
- Great resources for code-sharing
  - github
  - NITRC
  - ITK

## References


[1] Ashburner, J., Aug 2012. Spm: a history. Neuroimage 62 (2), 791–800.

[2] Cox, R. W., Aug 2012. Afni: what a long strange trip it's been. Neuroimage 62 (2), 743–7.

[3] Jenkinson, M., Beckmann, C. F., Behrens, T. E. J., Woolrich, M. W., Smith, S. M., Aug 2012. Fsl. Neuroimage 62 (2), 782–90.

[4] Kriegeskorte, N., Simmons, W. K., Bellgowan, P. S. F., Baker, C. I., May 2009. Circular analysis in systems neuroscience: the dangers of double dipping. Nat Neurosci 12 (5), 535–40.

[5] Poldrack, R. A., Fletcher, P. C., Henson, R. N., Worsley, K. J., Brett, M., Nichols, T. E., Apr 2008. Guidelines for reporting an fmri study. Neuroimage 40 (2), 409–14.

[6] Sackett, D. L., 1979. Bias in analytic research. J Chronic Dis 32 (1-2), 51–63.

[7] Tustison, N. J., Avants, B. B., Cook, P. A., Kim, J., Whyte, J., Gee, J. C., Stone, J. R., Nov 2012. Logical circularity in voxel-based analysis: Normalization strategy may induce statistical bias. Hum Brain Mapp.

[8] Vul, E., Pashler, H., Aug 2012. Voodoo and circularity errors. Neuroimage 62 (2), 945–8.