frontiers in
NEUROSCIENCE

OPINION ARTICLE
published: 09 September 2013
doi: 10.3389/fnins.2013.00162

Instrumentation bias in the use and evaluation of scientific software: recommendations for reproducible practices in the computational sciences

Nicholas J. Tustison[1]*, Hans J. Johnson[2], Torsten Rohlfing[3], Arno Klein[4], Satrajit S. Ghosh[5], Luis Ibanez[6] and Brian B. Avants[7]

[1] *Department of Radiology and Medical Imaging, University of Virginia, Charlottesville, VA, USA*
[2] *Department of Psychiatry, University of Iowa, Iowa City, IA, USA*
[3] *SRI International, Menlo Park, CA, USA*
[4] *Department of Psychiatry and Behavioral Science, Stony Brook University School of Medicine, Stony Brook, NY, USA*
[5] *Massachusetts Institute of Technology, McGovern Institute for Brain Research, Cambridge, MA, USA*
[6] *Kitware Inc., Clifton Park, NY, USA*
[7] *Penn Image Computing and Science Laboratory, Department of Radiology, University of Pennsylvania, Philadelphia, PA, USA*
**Correspondence: ntustison@virginia.edu*

Edited by:
Pedro Antonio Valdes-Sosa, Cuban Neuroscience Center, USA

Keywords: best practices, open science, comparative evaluations, confirmation bias, reproducibility

1. INTRODUCTION

> By honest I don't mean that you only tell what's true. But you make clear the entire situation. You make clear all the information that is required for somebody else who is intelligent to make up their mind.
>
> Richard Feynman

The neuroscience community significantly benefits from the proliferation of imaging-related analysis software packages. Established packages such as SPM (Ashburner, 2012), the FMRIB Software Library (FSL) (Jenkinson et al., 2012), Freesurfer (Fischl, 2012), Slicer (Fedorov et al., 2012), and the AFNI toolkit (Cox, 2012) aid neuroimaging researchers around the world in performing complex analyses as part of ongoing neuroscience research. In conjunction with distributing robust software tools, neuroimaging packages also continue to incorporate algorithmic innovation for improvement in analysis tools.

As fellow scientists who actively participate in neuroscience research through our contributions to the Insight Toolkit[1] (e.g., Johnson et al., 2007; Ibanez et al., 2009; Tustison and Avants, 2012) and other packages such as MindBoggle,[2] Nipype[3] (Gorgolewski et al., 2011), and the Advanced Normalization Tools (ANTs),[4] (Avants et al., 2010, 2011) we notice an increasing number of publications that intend a fair comparison of algorithms which, in principle, is a good thing. Our concern is the lack of detail with which these comparisons are often presented and the corresponding possibility of *instrumentation bias* (Sackett, 1979) where "defects in the calibration or maintenance of measurement instruments may lead to systematic deviations from true values" (considering software as a type of instrument requiring proper "calibration" and "maintenance" for accurate measurements). Based on our experience (including our own mistakes), we propose a preliminary set of guidelines that seek to minimize such bias with the understanding that the discussion will require a more comprehensive response from the larger neuroscience community. Our intent is to raise awareness in both authors and reviewers to issues that arise when comparing quantitative algorithms. Although herein we focus largely on image registration, these recommendations are relevant for other application areas in biologically-focused computational image analysis, and for reproducible computational science in general. This commentary complements recent papers that highlight statistical bias (Kriegeskorte et al., 2009; Vul and Pashler, 2012), bias induced by registration metrics (Tustison et al., 2012), and registration strategy (Yushkevich et al., 2010) and guideline papers for software development (Prlic and Procter, 2012).

[1] http://www.itk.org
[2] http://www.mindboggle.info
[3] http://nipy.org/nipype

[4] http://stnava.github.io/ANTs/

2. GUIDELINES

A comparative analysis paper's longevity and impact on future scientific explorations is directly related to the completeness of the evaluation. A complete evaluation requires preparation (before any experiment is performed) and effort to publish its details and results. Here, we suggest general guidelines for both of these steps most of which derive from basic scientific principles of clarity and reproducibility.

2.1. DESIGNING THE EVALUATION STUDY

The very idea that one (e.g., registration) algorithm could perform better than all other algorithms on all types of data is fundamentally flawed. Indeed, the "No Free Lunch Theorem" provides bounds on solution quality. That is, it specifically demonstrates that "improvement of performance in problem-solving hinges on using prior information to match procedures to problems" (Wolpert and Macready, 1997). Therefore, the first thing that authors of new algorithms should do is identify how their methods differ with respect to other available techniques in terms of the use of prior knowledge. Furthermore, the author must consider if it is possible to incorporate prior knowledge across existing methods.

2.1.1. Demand that the algorithm developers provide default parameters for the comparative context being investigated

Expert knowledge of a specific program and/or algorithm is most likely found with the original developers who would be in a position to provide optimal parameterization. Relevant parameter files and sample scripts that detail command line calls should accompany an algorithm to aid in its proper use, evaluation, and comparison. For example, the developers of the image registration program elastix (Klein et al., 2010) provide an assortment of parameter files on a designated wiki page[5] listed in tabular format complete with short description (including applied modality and object of interest) and any publications which used that specific parameter file. Another example is the National Alliance for Medical Image Computing registration use case inventory[6] where each listed case comprises a test dataset, a guided step-by-step tutorial, the solution, and a custom Registration Parameter Presets file with optimized registration parameters.

2.1.2. Do not implement your own version of an algorithm, particularly if one is available from the original authors. If you must re-implement, consider making your implementation available

Much is left unstated in published manuscripts where novel algorithmic concepts are presented. Ideally, the authors provide an instantiation of the code to accompany the manuscript. As observed in Kovacevic (2006), however, this is often not the case (even in terms of pseudocode). As a result, comparative evaluations are sometimes carried out using code developed not by the original authors but by the group doing the comparison. For example, in Clarkson et al. (2011), the authors compared three algorithms for estimating cortical thickness. Two of the algorithms were coded by the authors of the study while the third was used "off the shelf." Thus, a natural question to ask is whether the performance difference is due to the algorithm itself, implementation quality, and/or the parameter tuning. None of these are addressed by Clarkson et al. (2011) which may decrease the publication's usefulness.

2.1.3. Perform comparisons on publicly available data

For reasons of reproducibility and transparency, evaluations should be performed using publicly available data sets. Given the rather large number of such institutional efforts including NIREP,[7] IXI,[8] NKI,[9] OASIS,[10] Kirby,[11] LONI,[12] and others, evaluations should include (if not be exhausted by) comparisons using such data. While evaluation on private cohorts might exclude such possibilities, such evaluations should be extensively motivated in the introduction and/or discussion. For example, if a particular algorithm with general application is found to perform better on a private cohort of Parkinson's disease subject data, reasons for performance disparity should be offered and supplemented with analysis on public data.

2.2. PUBLISHING THE EVALUATION

2.2.1. Include parameters

In Klein et al. (2009), 14 non-linear registration algorithms were compared using four publicly available, labeled brain MRI data sets. As part of the study, the respective algorithms' authors were given an opportunity to tune the parameters to ensure good performance which were then distributed on Prof. Klein's website.[13] In contrast, not specifying parameters leaves one susceptible to criticisms of confirmation and/or instrumentation bias. For example, in a recent paper (Haegelen et al., 2013),[14] the authors compared their ANIMAL registration algorithm with SyN (Avants et al., 2011) and determined that "registration with ANIMAL was better than with SyN for the left thalamus" in a cohort of Parkinson's disease patients. The difference in the authors' experience and investment between the two algorithms could bias algorithmic performance assessment. However, inclusion of parameter settings for ANIMAL and SyN would permit independent verification by reviewers or readers of the article.

2.2.2. Provide details as to the source of the algorithm

Origin should be provided for any code or package used during the evaluation. For example, N4 (Tustison et al., 2010) is a well-known inhomogeneity correction algorithm for MRI first made available as a tech report (Tustison and Gee, 2009). However, since its inclusion in the Insight Toolkit, different programs have been made available. N4 is also available in ANTs (the only version directly maintained by the original authors), as a module in Slicer,[15] a wrapper of the Slicer module in Nipype,[16] a module in `c3d`,[17] and as a plugin in the BRAINS suite.[18] While each version is dependent on the original source code, there could exist subtle variations which can affect performance. As one specific example, the `c3d` implementation hard-codes certain parameter values with no access to modify them by the user.

2.2.3. Co-authors should verify findings

Although different journals have varying guidelines for determining co-authorship, there is at least an implied sense of responsibility for an article's contents assumed by each of the co-authors. Strategies taken by journal editorial boards are used to reduce undeserving authorship attribution such as requiring the listing of the specific contributions of each co-author. Additional proposals have included signed statements of responsibility for the contents of an article (Anonymous, 2007). We suggest that at least one co-author independently verify a subset of the results by running the data processing and analysis on their own computational platform. The point of this exercise is to verify not only

[5] http://elastix.bigr.nl/wiki/index.php/Parameter_file_database

[6] http://www.na-mic.org/Wiki/index.php/Projects:RegistrationDocumentation:UseCaseInventory

[7] http://www.nirep.org

[8] http://www.brain-development.org

[9] http://fcon_1000.projects.nitrc.org/indi/pro/nki.html

[10] http://www.oasis-brains.org

[11] http://mri.kennedykrieger.org/databases.html

[12] http://www.loni.ucla.edu/Research/Databases/

[13] http://mindboggle.info/papers/evaluation_NeuroImage2009.php

[14] Similar issues can be found in Wu et al. (2013).

[15] http://www.slicer.org/slicerWiki/index.php/Documentation/4.2/Modules/N4ITKBiasField Correction

[16] http://www.mit.edu/~satra/nipype-nightly/interfaces/generated/nipype.interfaces.slicer.filtering.n4itkbiasfieldcorrection.html

[17] http://www.itksnap.org8/pmwiki/pmwiki.php?n=Convert3D.Documentation

[18] http://www.nitrc.org/plugins/mwiki/index.php/brains:N4ITK

reproducibility but also that the process can be explained in sufficient detail.

2.2.4. Provide computational platform details of the evaluation

A recent article (Gronenschild et al., 2012) pointed out significant differences in FreeSurfer output that varied with release version and with operating system. While the former is to be expected given upgrades and bug fixes which occur between releases, the latter underscores both the need for consistency in study processing as well as the reporting of computational details for reproducibility.

2.2.5. Supply pre- and post-processing steps

In addition to disclosure of all parameters associated with the methodologies to be compared, all processing steps from the raw to the final processed images in the workflow need to be specified. Tools like Nipype (Gorgolewski et al., 2011) capture this provenance information in a formal and rigorous way, but at a minimum the shell scripts or screen shots of the parameter choices should be made available. Justification for any deviation of steps between algorithms needs to be provided.

2.2.6. Post the resulting data online

The current publishing paradigm limits the quantity of results that can be posted. There are only so many pages allowed for a particular publication and displaying every slice of every processed image, for example, is not feasible. This results in possible selection bias where results provided in the manuscript are selected by the authors for demonstrating the effect postulated at the onset of the study. Thus, differences in performance assessment tend to be exaggerated based strictly on visual representations in the paper. Publication simply in print (or as figures in a PDF file) and its limitations in terms of dynamic range or spatial resolution also severely limits the ability of reviewers and readers to perform more sophisticated evaluation beyond simple visual inspection.

Alternatively (or additionally), online resources such as the the LONI Segmentation Validation Engine (Shattuck et al., 2009)[19] can be used to evaluate individual algorithms for brain segmentation on publicly available data sets and compare with previously posted results. A top ranking outcome provides significant external validation for publishing newly proposed methodologies (e.g., Eskildsen et al., 2012).

[19] http://sve.loni.ucla.edu

2.2.7. Put comparisons and observed performance differences into context

In addition to algorithmic and study specifics, it is important to discuss potential limitations concerning qualitative and/or quantitative assessment metrics. In Rohlfing (2012), the author pointed out deficiencies in using standard overlap measures and image similarity metrics in quantifying performance of image registration methods. Other issues, such as biological plausibility of the resulting transforms, need to also be considered. Also important for inclusion is discussion of the possible reasons for performance disparity. If one algorithm outperforms another, reporting of those findings would be much more significant if the authors discuss possible reasons for relative performance levels.

3. CONCLUSION

Considering that computational sciences permeate neuroimaging research, certain safeguards should be in place to prevent (or at least minimize) potential biases and errors that can unknowingly affect study outcomes. There is no vetting agency for ensuring that analysis programs used for research are reasonably error-free. In addition, these software packages are simply "black boxes" to many researchers who are not formally trained to debug code, and who, in most cases, have only a very superficial understanding of the algorithms that they apply. And even to those of us who are trained to debug code, understanding someone else's code, perhaps implemented in an unfamiliar programming language and different coding style, is oftentimes very difficult. To this end, algorithmic comparisons are a very good way of evaluating general performance. We hope that the guidelines proposed in this editorial help the community in future comparative assessments and avoid errors in scientific computing that may otherwise lead to publication of invalid results (Merali, 2010).

REFERENCES

Anonymous. (2007). Who is accountable? *Nature* 450:1. doi: 10.1038/450001a

Ashburner, J. (2012). SPM: a history. *Neuroimage* 62, 791–800. doi: 10.1016/j.neuroimage.2011.10.025

Avants, B. B., Tustison, N. J., Song, G., Cook, P. A., Klein, A., and Gee, J. C. (2011). A reproducible evaluation of ANTs similarity metric performance in brain image registration. *Neuroimage* 54, 2033–2044. doi: 10.1016/j.neuroimage.2010.09.025

Avants, B. B., Yushkevich, P., Pluta, J., Minkoff, D., Korczykowski, M., Detre, J., et al. (2010). The optimal template effect in hippocampus studies of diseased populations. *Neuroimage* 49, 2457–2466. doi: 10.1016/j.neuroimage.2009.09.062

Clarkson, M. J., Cardoso, M. J., Ridgway, G. R., Modat, M., Leung, K. K., Rohrer, J. D., et al. (2011). A comparison of voxel and surface based cortical thickness estimation methods. *Neuroimage* 57, 856–865. doi: 10.1016/j.neuroimage.2011.05.053

Cox, R. W. (2012). AFNI: what a long strange trip it's been. *Neuroimage* 62, 743–747. doi: 10.1016/j.neuroimage.2011.08.056

Eskildsen, S. F., Coupé, P., Fonov, V., Manjón, J. V., Leung, K. K., Guizard, N., et al. (2012). BEaST: brain extraction based on nonlocal segmentation technique. *Neuroimage* 59, 2362–2373. doi: 10.1016/j.neuroimage.2011.09.012

Fedorov, A., Beichel, R., Kalpathy-Cramer, J., Finet, J., Fillion-Robin, J.-C., Pujol, S., et al. (2012). 3D slicer as an image computing platform for the quantitative imaging network. *Magn. Reson. Imaging* 30, 1323–1341. doi: 10.1016/j.mri.2012.05.001

Fischl, B. (2012). FreeSurfer. *Neuroimage* 62, 774–781. doi: 10.1016/j.neuroimage.2012.01.021

Gorgolewski, K., Burns, C. D., Madison, C., Clark, D., Halchenko, Y. O., Waskom, M. L., et al. (2011). Nipype: a flexible, lightweight and extensible neuroimaging data processing framework in python. *Front. Neuroinform.* 5:13. doi: 10.3389/fninf.2011.00013

Gronenschild, E. H. B. M., Habets, P., Jacobs, H. I. L., Mengelers, R., Rozendaal, N., van Os, J., et al. (2012). The effects of FreeSurfer version, workstation type, and Macintosh operating system version on anatomical volume and cortical thickness measurements. *PLoS ONE* 7:e38234. doi: 10.1371/journal.pone.0038234

Haegelen, C., Coupé, P., Fonov, V., Guizard, N., Jannin, P., Morandi, X., et al. (2013). Automated segmentation of basal ganglia and deep brain structures in MRI of Parkinson's disease. *Int. J. Comput. Assist. Radiol. Surg.* 8, 99–110. doi: 10.1007/s11548-012-0675-8

Ibanez, L., Audette, M., Yeo, B. T., and Goland, P. (2009). Spherical demons registration of spherical surfaces. *Insight J.* Available online at: http://hdl.handle.net/10380/3117

Jenkinson, M., Beckmann, C. F., Behrens, T. E. J., Woolrich, M. W., and Smith, S. M. (2012). FSL. *Neuroimage* 62, 782–790. doi: 10.1016/j.neuroimage.2011.09.015

Johnson, H. J., Harris, G., and Williams, K. (2007). BRAINSFit: mutual information registrations of whole-brain 3D images, using the insight toolkit. *Insight J.* Available online at: http://hdl.handle.net/1926/1291

Klein, A., Andersson, J., Ardekani, B. A., Ashburner, J., Avants, B., Chiang, M.-C., et al. (2009). Evaluation of 14 nonlinear deformation algorithms applied to human brain MRI registration. *Neuroimage* 46, 786–802. doi: 10.1016/j.neuroimage.2008.12.037

Klein, S., Staring, M., Murphy, K., Viergever, M. A., and Pluim, J. P. W. (2010). elastix: a toolbox for intensity-based medical image registration. *IEEE Trans. Med. Imaging* 29, 196–205. doi: 10.1109/TMI.2009.2035616

Kovacevic, J. (2006). From the editor in chief. *IEEE Trans. Image Process.* 15, 12.

Kriegeskorte, N., Simmons, W. K., Bellgowan, P. S. F., and Baker, C. I. (2009). Circular analysis in systems neuroscience: the dangers of double dipping. *Nat. Neurosci.* 12, 535–540. doi: 10.1038/nn.2303

Merali, Z. (2010). Computational science: error, why scientific programming does not compute. *Nature* 467, 775–777. doi: 10.1038/467775a

Prlic, A., and Procter, J. B. (2012). Ten simple rules for the open development of scientific software. *PLoS Comput. Biol.* 8:e1002802. doi: 10.1371/journal.pcbi.1002802

Rohlfing, T. (2012). Image similarity and tissue overlaps as surrogates for image registration accuracy: widely used but unreliable. *IEEE Trans. Med. Imaging* 31, 153–163. doi: 10.1109/TMI.2011.2163944

Sackett, D. L. (1979). Bias in analytic research. *J. Chronic. Dis.* 32, 51–63. doi: 10.1016/0021-9681(79)90012-2

Shattuck, D. W., Prasad, G., Mirza, M., Narr, K. L., and Toga, A. W. (2009). Online resource for validation of brain segmentation methods. *Neuroimage* 45, 431–439. doi: 10.1016/j.neuroimage.2008.10.066

Tustison, N. J., and Avants, B. B. (2012). The TVDMFFDVR algorithm. *Insight J.* Available online at: http://hdl.handle.net/10380/3334

Tustison, N. J., Avants, B. B., Cook, P. A., Kim, J., Whyte, J., Gee, J. C., et al. (2012). Logical circularity in voxel-based analysis: normalization strategy may induce statistical bias. *Hum. Brain Mapp*. doi: 10.1002/hbm.22211. [Epub ahead of print].

Tustison, N. J., Avants, B. B., Cook, P. A., Zheng, Y., Egan, A., Yushkevich, P. A., et al. (2010). N4ITK: improved N3 bias correction. *IEEE Trans. Med. Imaging* 29, 1310–1320. doi: 10.1109/TMI.2010.2046908

Tustison, N. J., and Gee, J. C. (2009). N4ITK: Nick's N3 ITK implementation for MRI bias field correction. Technical report, University of Pennsylvania. Available online at: http://www.insight-journal.org/browse/publication/640

Vul, E., and Pashler, H. (2012). Voodoo and circularity errors. *Neuroimage* 62, 945–948. doi: 10.1016/j.neuroimage.2012.01.027

Wolpert, D., and Macready, W. G. (1997). No free lunch theorems for optimization. *IEEE Trans. Evol. Comput.* 1, 67–82. doi: 10.1109/4235.585893

Wu, G., Kim, M., Wang, Q., and Shen, D. (2013). S-HAMMER: hierarchical attribute-guided, symmetric diffeomorphic registration for MR brain images. *Hum. Brain Mapp*. doi: 10.1002/hbm.22233. [Epub ahead of print].

Yushkevich, P. A., Avants, B. B., Das, S. R., Pluta, J., Altinay, M., Craige, C., et al. (2010). Bias in estimation of hippocampal atrophy using deformation-based morphometry arises from asymmetric global normalization: an illustration in ADNI 3 T MRI data. *Neuroimage* 50, 434–445. doi: 10.1016/j.neuroimage.2009.12.007

Received: 03 July 2013; accepted: 20 August 2013; published online: 09 September 2013.

Citation: Tustison NJ, Johnson HJ, Rohlfing T, Klein A, Ghosh SS, Ibanez L and Avants BB (2013) Instrumentation bias in the use and evaluation of scientific software: recommendations for reproducible practices in the computational sciences. Front. Neurosci. 7:162. doi: 10.3389/fnins.2013.00162

This article was submitted to Brain Imaging Methods, a section of the journal Frontiers in Neuroscience.