Title

A very simple, re-executable neuroimaging publication

Authors
Ghosh, Satrajit S
Poline, Jean-Baptiste
Keator, David B
et al.

Publication Date
2017-07-24

DOI
10.12688/f1000research.10783.1

Peer reviewed

METHOD ARTICLE

REVISED A very simple, re-executable neuroimaging publication [version 2; referees: 1 approved, 3 approved with reservations]

Satrajit S. Ghosh [1,2], Jean-Baptiste Poline[3], David B. Keator[4],
Yaroslav O. Halchenko[5], Adam G. Thomas [6], Daniel A. Kessler [7],
David N. Kennedy [8]

[1]McGovern Institute for Brain Research, Massachusetts Institute of Technology: MIT, Cambridge, MA, USA
[2]Department of Otology and Laryngology, Harvard Medical School, Boston, MA, USA
[3]Henry Wheeler Jr. Brain Imaging Center, Helen Wills Neuroscience Institute, University of California, Berkeley, Berkeley, CA, USA
[4]Department of Psychiatry and Human Behavior, Department of Computer Science, Department of Neurology, University of California, Irvine, Irvine, CA, USA
[5]Department of Psychological and Brain Sciences, Dartmouth College, Hanover, NH, USA
[6]Data Science and Sharing Team, National Institute of Mental Health, Intramural Research Programs, Bethesda, MD, USA
[7]Department of Psychiatry and Department of Radiology, University of Michigan, Ann Arbor, MI, USA
[8]Eunice K. Shriver Center and Department of Psychiatry, University of Massachusetts Medical School, Worcester, MA, USA

v2 **First published:** 10 Feb 2017, **6**:124 (doi: 10.12688/f1000research.10783.1)
Latest published: 15 Jun 2017, **6**:124 (doi: 10.12688/f1000research.10783.2)

Abstract

Reproducible research is a key element of the scientific process.
Re-executability of neuroimaging workflows that lead to the conclusions arrived at in the literature has not yet been sufficiently addressed and adopted by the neuroimaging community. In this paper, we document a set of procedures, which include supplemental additions to a manuscript, that unambiguously define the data, workflow, execution environment and results of a neuroimaging analysis, in order to generate a verifiable re-executable publication.
Re-executability provides a starting point for examination of the generalizability and reproducibility of a given finding.

This article is included in the INCF gateway.

Open Peer Review

Referee Status: ? ? ✓ ?

	Invited Referees			
	1	**2**	**3**	**4**
REVISED **version 2** published 15 Jun 2017	? report		✓ report	? report
version 1 published 10 Feb 2017	? report	? report		

1 **Konrad Hinsen**, Rue Charles Sadron, France
Synchrotron SOLEIL, France

2 **Allan J. MacKenzie-Graham**, University of California, Los Angeles, USA

3 **Chao-Gan Yan**, Institute of Psychology, China

4 **Ze Wang**, Temple University, USA

Discuss this article

Comments (0)

Corresponding author: David N. Kennedy (David.Kennedy@umassmed.edu)

Author roles: Ghosh SS: Conceptualization, Formal Analysis, Methodology, Software, Validation, Writing – Original Draft Preparation, Writing – Review & Editing; **Poline JB**: Conceptualization, Methodology, Writing – Original Draft Preparation, Writing – Review & Editing; **Keator DB**: Data Curation, Methodology, Resources, Writing – Original Draft Preparation, Writing – Review & Editing; **Halchenko YO**: Conceptualization, Formal Analysis, Methodology, Resources, Software, Writing – Original Draft Preparation, Writing – Review & Editing; **Thomas AG**: Investigation, Validation, Writing – Review & Editing; **Kessler DA**: Investigation, Validation, Writing – Review & Editing; **Kennedy DN**: Conceptualization, Data Curation, Funding Acquisition, Project Administration, Supervision, Validation, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

How to cite this article: Ghosh SS, Poline JB, Keator DB *et al.* **A very simple, re-executable neuroimaging publication [version 2; referees: 1 approved, 3 approved with reservations]** *F1000Research* 2017, **6**:124 (doi: 10.12688/f1000research.10783.2)

Grant information: This work was supported by: NIH-NIBIB P41 EB019936 (ReproNim), NIH-NIBIB R01 EB020740 (Nipype), and NIH-NIMH R01 MH083320 (CANDIShare). J-BP was also partially supported by NIH NIH-NIDA 5U24 DA039832 (NIF).
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

First published: 10 Feb 2017, **6**:124 (doi: 10.12688/f1000research.10783.1)

REVISED **Amendments from Version 1**

In response to reviewer comments, several parts were revised in the manuscript:

- We now take more care in providing complete descriptions of the OS and software versions that are being used as 'reference' and test runs.
- Included a number of additional references.
- We have updated to using a Docker version of the workflow as the 'reference' run, in order to ease exact re-execution.
- Updated the github content and instructions to the potential users to reflect these changes, and provide instructions as to how to create this specific docker container.
- Added additional discussion on the licensing terms that are conveyed with the FSL-based tools we used.
- Discuss the numeric similarity thresholding procedure, and expand upon the operating system-based differences noted between the 'reference' run and a MacOS run.
- Updated Table 1 to reflect this additional information (correlation coefficients between OS, percentiled range of voxel difference between OS runs)
- Expand upon the Conda environment provisioning description.

See referee reports

Introduction

The quest for more reproducibility and replicability in neuroscience research spans many types of problems. True reproducibility requires the observation of a 'similar result' through the execution of a subsequent independent, yet similar, analysis on similar data. However, what constitutes 'similar', and how to appropriately annotate and integrate a lack of replication in specific studies remains a problem for the community and the literature that we generate.

The reproducibility problem

A number of studies have brought the reproducibility of science into question (Prinz *et al.*, 2011). Numerous factors are critical to understand reproducibility, including: sample size, and its related issues of power and generalizability (Button *et al.*, 2013; Ioannidis, 2005); P-hacking, trying various statistical approaches in order to find analyses that reach significance (Simmons *et al.*, 2011; Simonsohn *et al.*, 2014); completeness of methods description, the written text of a publication cannot completely describe an analytic method in its entirety. Coupled with this is the publication bias that arises from only publishing results from the positive ("significant") tail of the distribution of findings. This contributes to a growing literature of findings that do not properly 'self-correct' through an equivalent publication of negative findings (that indicate a lack of replication). Such corrective aggregation is needed to balance the inevitable false positives that result from the millions of experiments that are performed each year.

But before even digging too deeply into the exceedingly complex topic of reproducibility, there already is great concern that a typical neuroimaging publication, the basic building block that our scientific knowledge enterprise is built upon, is rarely even re-executable, even by the original investigators. The general framework for a publication is the following: take some specified "Data", apply a specified "Analysis", and generate a set of "Results". From the Results, claims are then made and discussed. In the context of this paper, we consider "Analysis" to include the software, workflow and execution environment, and use the following definitions of reproducibility:

> **Re-executability (publication-level replication):** The exact same data, operated on by the exact same analysis should yield the exact same result. This is currently a problem since publications, in order to maintain readability, do not typically provide a complete specification of the analysis method or access to the exact data.
>
> **Generalizability:** We can divide generalizability into three variations:
>
> > ***Generalization Variation 1:*** Exact Same Data + ***Nominally 'Similar' Analyses*** should yield a 'Similar' Result (i.e. FreeSurfer subcortical volumes compared to FSL FIRST)
> >
> > ***Generalization Variation 2: Nominally 'Similar' Data*** + Exact Same Analysis should yield a 'Similar' Result (i.e. the cohort of kids with autism I am using compared to the cohort you are using)
> >
> > ***Generalized Reproducibility:*** Nominally 'Similar' Data + Nominally 'Similar' Analyses should yield a 'Similar' Result

Since we do not really characterize data, analysis, and results very exhaustively in the current literature, this lack of provenance (Mackenzie-Graham *et al.*, 2008) permits the concept of 'similar' to have lots of wiggle room for interpretation (both to enhance similarity and to discount differences, as desired by the interests of the author).

In this paper, we look more closely at the re-executability necessary for publication-level replication. The technology exists, in many cases, to make neuroimaging publications that are fully re-executable. Re-executability of an initial publication is a crucial step in the goal of overall reproducibility of a given research finding. There are already examples of re-executable individual articles (e.g. Waskom *et al.*, 2014), as well as journals that propose to publish reproducible and open research (e.g. https://rescience.github.io). Here, we propose a formal template for a reproducible brain imaging publication and provide an example on fully open data from the NITRC Image Repository. The key elements to publication re-executability are definition of and access to: 1) the data; 2) the processing workflow; 3) the execution environment; and 4) the complete results. In this report, we use existing technologies (i.e., NITRC (http://nitrc.org), NIDM (http://nidm.nidash.org), Nipype (http://nipy.org/nipype), NeuroDebian (http://neuro.debian.net)) to generate a re-executable publication for a very simple analysis problem, which can form an essential template to guide future progress in enhancing re-executability of workflows in neuroimaging publications. Specifically, we explore the issue of exact re-execution (identical execution environment) and re-execution of identical workflow and data in 'similar' execution environments (Glatard *et al.*, 2015; Gronenschild *et al.*, 2012).

Methods

Overview

We envision a ‘publication’ with four supplementary files, the: 1) data file, 2) workflow file, 3) execution environment specification, and 4) results. The task the author would like to enable, for an interested reader, will be to facilitate the use of the first three specifications and easily be able to run them, and confirm (or deny) the similarity of the results from an independent re-execution compared to those published.

For the purpose of this report, we wanted an easy to execute query run on completely open, publically available data. We also wanted to use a relatively simple workflow that could be run in a standard computational environment and have it operate on a tractable number of subjects. We selected a workflow and sample size such that the overall processing could be accomplished in a few hours. The complete workflow and results can be found in the Github repository (doi, 10.5281/zenodo.800758; Ghosh *et al.*, 2017).

The data. The dataset for this exercise was created by a query as an unregistered guest user of the NITRC Image Repository (NITRC-IR; RRID:SCR_004162; Kennedy *et al.*, 2016). We queried the NITRC-IR search page (http://www.nitrc.org/ir/app/template/Index.vm; 1-Jan-2017) on the ‘MR’ tab with the following specification: age, 10–15 years old; Field Strength, 3. This query returned 24 subjects, which included subject identifier, age, handedness, gender, acquisition site, and field strength. We then selected the ‘mprage_anonymized’ scan type and ‘NIfTI’ file format in order to access the URLs (uniform resource locators) for the T1-weighted structural image data of these 24 subjects. The subjects had the following characteristics: age=13.5 +/- 1.4 years; 16 males, 8 females; 8 right handed, 1 left and 15 unknown. All of these datasets were from the 1000 Functional Connectomes project (Biswal *et al.*, 2010), and included 9 subjects from the Ann Arbor sub-cohort, and 15 from the New York sub-cohort. We captured this data in tabular form (Supplementary File 1). Following the recommendations of the Joint Declaration of Data Citation Principles (Starr *et al.*, 2015), we used the Image Attribution Framework (Honor *et al.*, 2016) to create a unique identifier for this data collection (image collection: doi, 10.18116/C6C592; Kennedy, 2017). Data collection identifiers are useful in order to track and attribute future reuse of the dataset and maintain the credit and attribution connection to the constituent images of the collection which may, in general, come from heterogeneous sources. Representative images from this collection are shown in Figure 1.

The workflow. For this example, we use a simple workflow designed to generate subcortical structural volumes. We used the following tools from the FMRIB software library, version 5.0.9 (FSL, RRID:SCR_002823; Jenkinson *et al.*, 2012), conformation of the data to FSL standard space (fslreorient2std), brain extraction (BET), tissue classification (FAST), and subcortical segmentation (FIRST).

This workflow is represented in Nipype (RRID:SCR_002502; Gorgolewski *et al.*, 2011) to facilitate workflow execution and provenance tracking. The workflow is available in the GitHub repository. The workflow also includes an initial step that accesses the contents of Supplementary Table 1, which are pulled from a Googles Docs spreadsheet (https://docs.google.com/spreadsheets/d/11an55u9t2TAf0EV2pHN0vOd8Ww2Gie-tHp9xGULh_dA/edit?usp=sharing) to copy the specific data files to the system, and a step that extracts the volumes (in terms of number of voxels and absolute volume) of the resultant structures. The code for these additional steps is included in the GitHub repository as well. In this workflow, the following regions are assessed: brain and background (as determined from the masks generated by BET, the brain extraction tool), gray matter, white matter and CSF (from the output of FAST), and left and right accumbens, amygdala, caudate, hippocampus, pallidum, putamen, and thalamus-proper (from the output of FIRST) See Figure 2 for the workflow diagram.

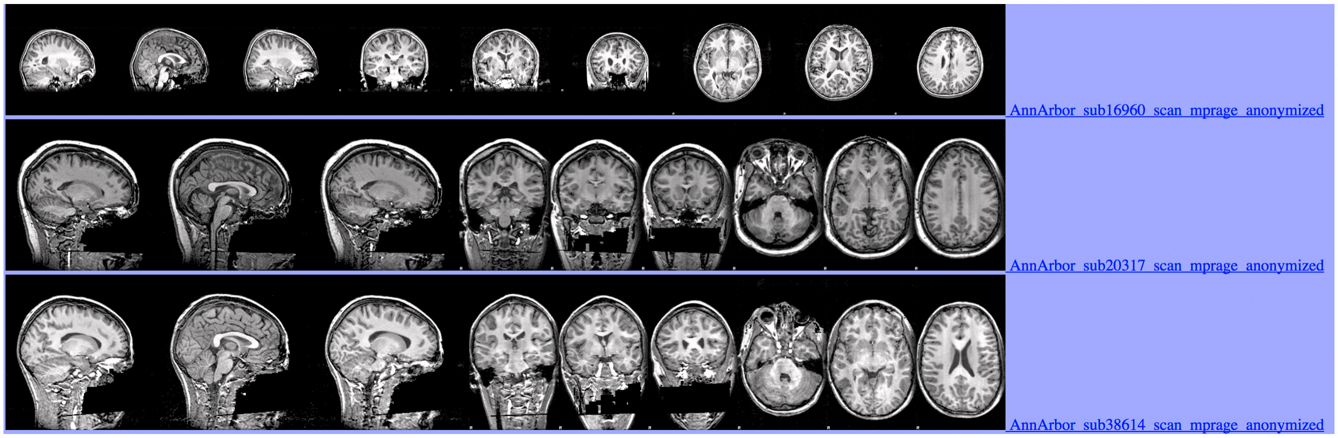
AnnArbor_sub16960_scan_mprage_anonymized
AnnArbor_sub20317_scan_mprage_anonymized
AnnArbor_sub38614_scan_mprage_anonymized

Figure 1. Example images from a subset of three of the subject image datasets used.

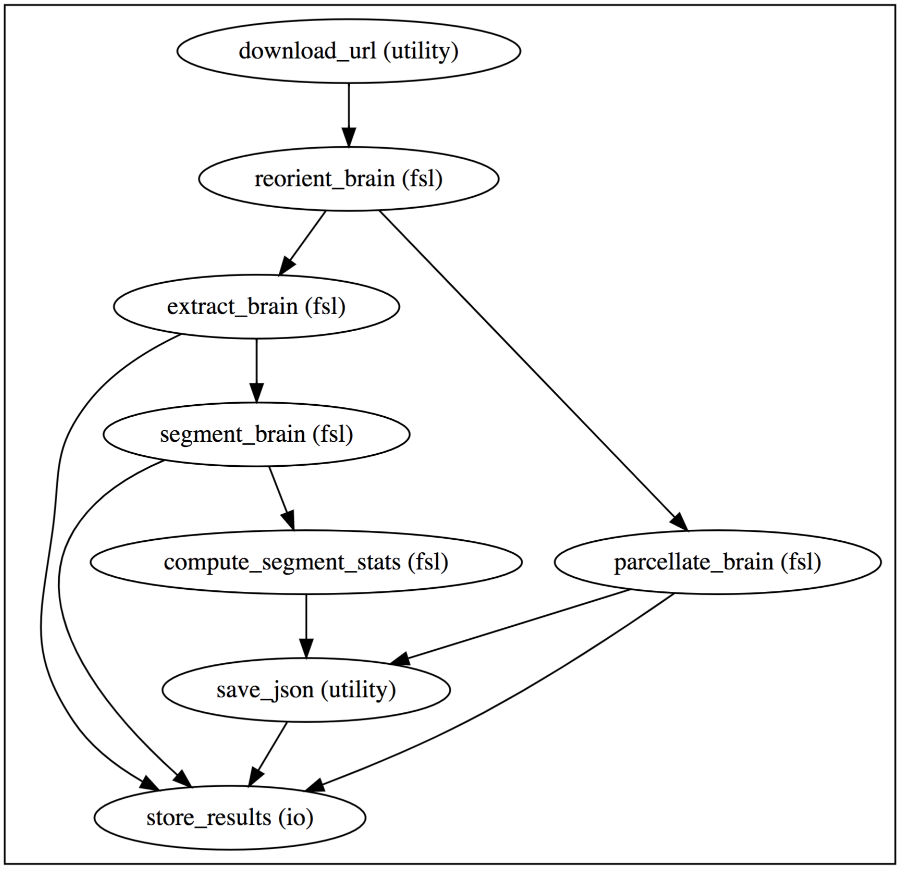

Figure 2. Workflow diagram. The sequence and dependence of processing events used in this example re-executable publication.

The execution environment. In order to utilize a computational environment that is, in principle, accessible to the other users in configuration identical to the one used to carry out this analysis, we created a Docker (https://www.docker.com/) container to encapsulate the specific computation environment and analysis pipeline components: https://hub.docker.com/r/repronim/simple_workflow/tags/. A Docker container permits efficient environment and software delivery for easy deployment on most common operating systems (Linux, Windows, Mac). The build instructions for the Docker container are provided on the GitHub repository, and uses Debian 8.7 as the base operating system.

Setting up the software environment on a different machine. In addition to a Docker container, one can re-execute the workflow on a different machine or cluster than the one used originally. General instructions for setting up the needed software environment on GNU/Linux and MacOS systems is provided in the README.md file in the GitHub repository. We assume FSL is installed and accessible on the command line (FSL can be found at https://fsl.fmrib.ox.ac.uk). In order to establish a precise overall environment, we use Conda (https://conda.io/), a cross-platform package manager and handles user installations for many packages into a controlled environment. Unlike many operating system package managers (e.g., yum, apt), Conda does not require root privileges. This allows individuals to replicate isolated virtual environments easily without requiring system administrator help. Conda uses standard PATH variables to isolate the environments. Coupled with Anaconda cloud and conda-forge, Conda is capable of installing versioned dependencies of Python and other packages. In this way, a Python 2.7.12 environment can be set up and the Nipype workflow re-executed with a few shell commands, as noted in the README.md.

One can also use the NITRC Computational Environment (NITRC-CE, RRID:SCR_002171). The NITRC-CE is built upon NeuroDebian (RRID:SCR_004401; Hanke & Halchenko, 2011), and comes with FSL (version 5.0.9-3~nd14.04+1) pre-installed on an Ubuntu 14.04 operating system. We run the computational environment on the Amazon Web Services (AWS) elastic cloud computing (EC2) environment. With EC2, the user can select properties of their virtual machine (number of cores, memory, etc.) in order to scale the power of the system to their specific needs. For this paper, we used the NITRC-CE v0.42, with the following specific identifier (AMI ID): ami-ce11f2ae.

The reference run

We performed the analysis (the above described workflow applied to the above described data, using the described computational system) with the Docker container provided and stored these results in our GitHub repository as the 'reference run', representing the official result that we are publishing for this analysis.

Generating the reference run. In order to run the analysis, we executed the following steps:

1) Download the Docker image

```
> docker pull repronim/simple_workflow:1.1.0
```

2) Run the Docker image as follows to perform the analysis:

```
> docker run -it --rm -v \
$PWD/output:/opt/repronim/simple_workflow/scripts/output\
  repronim/simple_workflow:1.1.0 run_demo_workflow.py\
  --key 11an55u9t2TAf0EV2pHN0vOd8Ww2Gie-tHp9xGULh_dA
```

Exact re-execution. In principle, any user could run the analysis steps, as described above, to obtain an exact replication of the reference results. The similarity of this result and the reference result can be verified by running the following command:

```
> python check_output.py
```

This program will compare the new results to the archived reference results and report on any differences, allowing for a numeric tolerance of 1e-6. If differences are found, a comma separated values (CSV) file is generated that quantifies these differences. The threshold in the 'check_output.py' script is simply selected to catch the presence of any numerical difference between the test run and the reference run. We discuss the implications of any differences, if found, below.

Re-execution on other systems. While the reference analysis was run using the provided Docker container, this analysis workflow can be run, locally or remotely, on many different operating systems. In general, the exact results of this workflow depends on the exact operating system, hardware, and the software versions. Execution of the above commands can be accomplished on any other Mac OS X or GNU/Linux distribution, as long as FSL is installed. In these cases, the results of the 'python check_output.py' command may indicate some numeric differences in the resulting volumes. In order to demonstrate these potential differences, we ran this identical workflow on the Mac OS X 10.12.4, CentOS 7.3, and NITRC Computational Environment on AWS.

Continuous integration

In addition to the reference run, the code for the project is housed in the Github repository. This allows integration with external services, such as CircleCI (http://circleci.com), which can re-execute the computation every single time a change is accepted into the code repository. Currently, the continuous integration testing runs on amd64 Debian (8.7) and uses FSL (5.0.9) from NeuroDebian. This re-execution generates results that are compared with the reference run, allowing us to evaluate a similar analysis automatically.

Results

Exact versions of data, code, environment details, and output

The specific versions of data used in this publication are available from NITRC. The code, environment details, and reference output are all available from the GitHub repository. The results of the reference run are stored in the expected_output folder of the GitHub repository at https://github.com/ReproNim/simple_workflow/tree/1.1.0/expected_output. By sharing the results of this reference run, as well as the data workflow, and a program to compare results from different runs, we can enable others to verify that they can arrive at the exact same result (if they use the exact same execution environment), or how close they come to the reference results if they utilize a different computational system (that may differ in terms of operating system, software versions, etc.).

Comparison of reference run and execution on other environments

When the workflow is re-executed in the same fashion (using the supplied Docker container) there is no observed difference in the output, regardless of Linux or Mac base platform. We also compared the execution of the reference run and re-execution natively (i.e. not via Docker container) in a separate MacOS environment. Table 1 indicates the numerical differences found when running natively on this MacOS example. Re-execution of the native analysis on the CentOS 7.3 and NITRC-CE (Ubuntu 14.04) on AWS provides an identical result to the reference run.

Discussion

Re-executability is an important first step in the establishment of a more comprehensive framework of reproducible computing. In order to properly compare the results of multiple papers, the underlying details of processing are essential to know to interpret the causes of 'similarity' and 'dissimilarity' between findings. By explicitly including linkage between a publication, and its data, workflow, execution environment and results, we can enhance the ability of the community to examine the issues related to reproducibility of specific findings.

In this publication, we are not looking at the causes of operating system dependence of neuroimaging results, but rather to emphasize the presence of this source of analysis variation, and examine ways to reduce this source of variance. Detailed results of neuroimaging analyses have been shown to be dependent on the exact details of the processing, specific computational operating system and software version (Glatard *et al.*, 2015; Gronenschild *et al.*, 2012). In this work, we replicate the observation that, despite an exact match on the data and workflow, the results of analysis can differ between execution in different operating systems. The implications of these differences are complex. On the one hand, the correlation of the volumetric results within each individual subject, and in aggregate across the population is very high (0.918 - 1.000). On the other

Table 1. Summary volumetric results from the simple workflow for the 24 subjects. Results are shown from the reference run (Docker) and a comparison run executed on a Mac OS X (10.12.4) system. The mean differences between these two systems are also summarized.

		Reference Run (Docker)		Mac OSX (10.12.4)			Difference				
Hemisphere	**Region**	**Mean Volume (mm3)**	**STD**	**Mean Volume (mm3)**	**STD**	**Correlation**	**Mean Volume (mm3)**	**STD**	**Range**	**Mean Absolute Volume**	**Percent of Reference**
Left	Accumbens	471.2	178.8	468.4	181.0	0.980	2.8	35.7	[-72.0, 75.6]	25.8	5.5
	Amygdala	840.2	257.9	849.2	262.9	0.963	-9.1	70.8	[-303.0, 84.0]	32.1	3.8
	Caudate	3842.4	650.4	3840.5	630.7	0.980	2.0	130.9	[-320.6, 518.4]	56.1	1.5
	Hippocampus	3276.2	794.9	3264.5	783.6	0.997	11.7	62.0	[-147.0, 185.0]	44.2	1.3
	Pallidum	1683.5	316.0	1668.8	313.6	0.995	14.7	30.7	[-16.8, 99.0]	20.2	1.2
	Putamen	4881.1	965.4	4890.2	952.7	0.998	-9.2	59.5	[-145.0, 144.0]	45.4	0.9
	Thalamus	8088.6	1240.8	8108.8	1239.0	0.998	-20.2	71.4	[-135.0, 194.0]	54.5	0.7
Right	Accumbens	389.1	147.6	402.2	148.5	0.966	-13.1	38.5	[-145.5, 32.0]	24.1	6.2
	Amygdala	882.3	297.8	897.6	290.7	0.918	-15.3	119.3	[-464.0, 221.0]	68.2	7.7
	Caudate	3781.0	762.1	3780.3	767.7	0.999	0.7	27.1	[-60.1, 48.0]	20.6	0.5
	Hippocampus	3433.4	796.9	3453.3	793.2	0.997	-19.9	60.6	[-190.0, 78.0]	42.4	1.2
	Pallidum	1694.5	293.8	1695.2	289.5	0.997	-0.7	21.8	[-64.0, 60.9]	12.9	0.8
	Putamen	4965.3	1017.3	4942.8	1008.0	0.993	22.5	118.0	[-201.6, 397.0]	78.4	1.6
	Thalamus. Proper	7723.6	1120.6	7724.3	1129.4	0.999	-0.7	52.9	[-145.0, 118.0]	33.6	0.4
Total	CSF	189856.5	26936.5	190209.3	26668.1	0.999	-352.8	1025.2	[-4274.2, 20.0]	359.7	0.2
	Gray Matter	684781.8	86048.8	684111.7	86173.4	1.000	670.1	2009.7	[-11.3, 7113.4]	676.3	0.1
	White Matter	513866.1	52348.3	513802.8	52341.8	1.000	63.3	467.3	[-541.2, 2192.4]	125.3	0.0
	Brain	1388504.4	132850.8	1388123.8	132962.5	1.000	380.6	1511.6	[-534.0, 6956.6]	425.1	0.0

hand, we see, per structure, a range of volumetric differences that reveals a large span of percentage of structure differences, differences that are, not surprisingly, dependent upon the overall size of the structure itself. The extremes of this distribution of the average difference (provided in Table 1) can be as large as 7.7% for the right amygdala. Sources of volumetric variance in this range can be troubling as biological changes on this order of volumetric difference can otherwise be the types of changes that studies are designed to observe. While in this case, the volumetric differences are not numerically large, it illustrates the general nature of this overall concern.

Publications can be made re-executable relatively simply by including links to the data, workflow, and execution environment. A re-executable publication with shared results is thus ***verifiable,*** by both the authors and others, increasing the trust in the results. The current simple example shows a simple volumetric workflow on a small dataset in order to demonstrate the way in which this could work in the real world. We felt it important to document this on a small problem (in terms of data and analysis complexity) in order to encourage others to actually verify these results, which is a practice we would like to see become more routine and feasible in the future. While this example approach is 'simple' in the context of what it accomplishes, it is still a rather complex and *ad hoc* procedure to follow. As such, it provides a roadmap for improvement, simplification, and standardization of the ways that these descriptive procedures can be handled.

Progress in simplifying this simple example can be expected in the near future on many fronts. Software deployments that are coupled with specific execution environments (such as Docker, Vagrant, Singularity, or other virtual or container machine instances) are now being deployed for common neuroimaging applications. In addition, more standardized data representations (such as BIDS, Gorgolewski *et al.*, 2016; NIDM, Keator *et al.*, 2013; BDBags, http://bd2k.ini.usc.edu/tools/bdbag/) will simplify how experimental data is assembled for sharing and use in specific software applications. Data distributions with clear versioning of the data, such as DataLad (http://datalad.org), will unify versioned access to data resources and sharing of derived results. While the workflow in this case is specified using Nipype, extensions to LONI Pipeline, shell scripting, and other workflow specifications is easily envisioned. Tools necessary to capture local execution environments (such as ReproZip, http://reprozip.org) will help users to share the software environment of their workflows in conjunction with their publications more easily.

Conclusion

We have demonstrated a simple example of a fully re-executable publication to take publically available neuroimaging data and compute some volumetric results. This is accomplished by augmenting the publication with four 'supplementary' files that include exact specification of 1) data, 2) workflow, 3) execution environment, and 4) results. This provides a roadmap to enhance the reproducibility of neuroimaging publications, by providing a basis for verifying the re-executability of individual publications and providing a more structured platform to examine the generalizability of the findings across changes in data, workflow details and execution environments. We expect these types of publication considerations to advance to a point where it can be relatively simple and routine to provide such supplementary materials for neuroimaging publications.

Software and data availability

Workflow and results are available on GitHub at: https://github.com/ReproNim/simple_workflow.

Archived workflow and results as at time of publication: doi, 10.5281/zenodo.800758 (Ghosh *et al.*, 2017).

License: Apache License, Version 2.0.

The data used in this publication are available at NITRC-IR (project, fcon_1000; image collection: doi, 10.18116/C6C592 - Kennedy, 2017) and referenced in Supplementary File 1. These data were originally gathered from the NITRC-IR, 1000 Functional Connectomes project, Ann Arbor and New York sub-projects.

Consent

The data used is anonymized and publicly available at NITRC-IR. Consent for the data sharing was obtained by each of the sharing institutions.

Author contributions

DNK, SSG, YOH and JBP conceived the study. SG designed the workflow, DNK generated and executed the data query, YOH designed the execution environment, DBK designed the data model. DNK, SSG, J-BP, DAK and AGT executed the re-execution experiments. DNK prepared the first draft of the manuscript. All authors were involved in the revision of the draft manuscript and have agreed to the final content.

Competing interests

No competing interests were disclosed.

Grant information

This work was supported by: NIH-NIBIB P41 EB019936 (ReproNim), NIH-NIBIB R01 EB020740 (Nipype), and NIH-NIMH R01 MH083320 (CANDIShare). J-BP was also partially supported by NIH NIH-NIDA 5U24 DA039832 (NIF).

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Acknowledgments

This work was conceived for and initially developed at OHBM Hackathon 2016 (http://brainhack.org/categories/ohbm-hackathon-2016/). We are exceedingly grateful to Cameron Craddock and the rest of the organizers of this event, and the Organization for Human Brain Mapping for support of their Open Science Special Interest Group (http://www.humanbrainmapping.org/i4a/pages/index.cfm?pageID=3712).

Supplementary material

Supplementary File 1: Data specification file. This file contains the basic demographics of the subjects (Subject, Age, Hand, Gender, and Acquisition Site) as well as the URL to the imaging data, as hosted at NITRC-IR (project, fcon_1000).

Click here to access the data.

References

Biswal BB, Mennes M, Zuo XN, *et al.*: **Toward discovery science of human brain function.** *Proc Natl Acad Sci U S A.* 2010; **107**(10): 4734–4739.
PubMed Abstract | Publisher Full Text | Free Full Text

Button KS, Ioannidis JP, Mokrysz C, *et al.*: **Power failure: why small sample size undermines the reliability of neuroscience.** *Nat Rev Neurosci.* 2013; **14**(5): 365–376.
PubMed Abstract | Publisher Full Text

Ghosh SS, Halchenko Y, Poline JB, *et al.*: **ReproNim - Simple Paper v1.0.0 [Data set].** *Zenodo.* 2017.
Data Source

Glatard T, Lewis LB, Ferreira da Silva R, *et al.*: **Reproducibility of neuroimaging analyses across operating systems.** *Front Neuroinform.* 2015; **9**: 12.
PubMed Abstract | Publisher Full Text | Free Full Text

Gorgolewski K, Burns CD, Madison C, *et al.*: **Nipype: a flexible, lightweight and extensible neuroimaging data processing framework in python.** *Front Neuroinform.* 2011; **5**: 13.
PubMed Abstract | Publisher Full Text | Free Full Text

Gorgolewski KJ, Auer T, Calhoun VD, *et al.*: **The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments.** *Sci Data.* 2016; **3**: 160044.
PubMed Abstract | Publisher Full Text | Free Full Text

Gronenschild EH, Habets P, Jacobs HI, *et al.*: **The effects of FreeSurfer version, workstation type, and Macintosh operating system version on anatomical volume and cortical thickness measurements.** *PLoS One.* 2012; **7**(6): e38234.
PubMed Abstract | Publisher Full Text | Free Full Text

Hanke M, Halchenko YO: **Neuroscience Runs on GNU/Linux.** *Front Neuroinform.* 2011; **5**: 8.
PubMed Abstract | Publisher Full Text | Free Full Text

Honor LB, Haselgrove C, Frazier JA, *et al.*: **Data Citation in Neuroimaging: Proposed Best Practices for Data Identification and Attribution.** *Front Neuroinform.* 2016; **10**: 34.
PubMed Abstract | Publisher Full Text | Free Full Text

Ioannidis JP: **Why most published research findings are false.** *PLoS Med.* 2005; **2**(8): e124.
PubMed Abstract | Publisher Full Text | Free Full Text

Jenkinson M, Beckmann CF, Behrens TE, *et al.*: **FSL.** *Neuroimage.* 2012; **62**(2): 782–790.
PubMed Abstract | Publisher Full Text

Keator DB, Helmer K, Steffener J, *et al.*: **Towards structured sharing of raw and derived neuroimaging data across existing resources.** *Neuroimage.* 2013; **82**: 647–61.
PubMed Abstract | Publisher Full Text | Free Full Text

Kennedy DN: **ReproNim Simple Workflow test dataset.** *ReproNim.* 2017.
Data Source

Kennedy DN, Haselgrove C, Riehl J, *et al.*: **The NITRC image repository.** *Neuroimage.* 2016; **124**(Pt B): 1069–1073.
PubMed Abstract | Publisher Full Text | Free Full Text

Mackenzie-Graham AJ, Van Horn JD, Woods RP, *et al.*: **Provenance in neuroimaging.** *Neuroimage.* 2008; **42**(1): 178–195.
PubMed Abstract | Publisher Full Text | Free Full Text

Prinz F, Schlange T, Asadullah K: **Believe it or not: how much can we rely on published data on potential drug targets?** *Nat Rev Drug Discov.* 2011; **10**(9): 712.
PubMed Abstract | Publisher Full Text

Simmons JP, Nelson LD, Simonsohn U: **False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant.** *Psychol Sci.* 2011; **22**(11): 1359–1366.
PubMed Abstract | Publisher Full Text

Simonsohn U, Nelson LD, Simmons JP: **P-curve: a key to the file-drawer.** *J Exp Psychol Gen.* 2014; **143**(2): 534–547.
PubMed Abstract | Publisher Full Text

Starr J, Castro E, Crosas M, *et al.*: **Achieving human and machine accessibility of cited data in scholarly publications.** *PeerJ Comput Sci.* 2015; **1**: pii: e1.
PubMed Abstract | Publisher Full Text | Free Full Text

Waskom ML, Kumaran D, Gordon AM, *et al.*: **Frontoparietal representations of task context support the flexible control of goal-directed cognition.** *J Neurosci.* 2014; **34**(32): 10743–55.
PubMed Abstract | Publisher Full Text | Free Full Text

Open Peer Review

Current Referee Status: ? ? ✓ ?

Version 2

Referee Report 24 July 2017

doi:10.5256/f1000research.12766.r24057

?

Ze Wang
Department of Radiology, Temple University, Philadelphia, PA, USA

The authors showed a framework to pack up the data, software, and analysis methods, specially for neuroimaging study. The results are reproducible which is the major aim of this paper. While I do think this paper is of interest to the general neuroimaging field, we should also notice that complete replication needs the data, software, and analysis procedures. The major hurdle however is often the data since data release may be sensitive and prohibited by regulations. How would the framework described in this paper be helpful for that case should be discussed. Is it possible to have software securely access the data but not leak the data (meaning prohibiting downloading, transfering out of the database etc)?

Is the rationale for developing the new method (or application) clearly explained?
Yes

Is the description of the method technically sound?
Yes

Are sufficient details provided to allow replication of the method development and its use by others?
Yes

If any results are presented, are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions about the method and its performance adequately supported by the findings presented in the article?
Yes

Competing Interests: No competing interests were disclosed.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Referee Report 17 July 2017

doi:10.5256/f1000research.12766.r24058

✓ **Chao-Gan Yan**

CAS Key Laboratory of Behavioral Science, Institute of Psychology, Bejing, China

This is an exciting endeavor to address the reproducibility crisis by providing re-executable neuroimaging publication. This study is well designed, analyzed and written, with the analysis can be easily replicated. I only have some minor comments.

1. One may would like to follow this paper to try to make their publication re-executable. As the first step is sharing data, the authors may would like to provide some guidance to share data.

2. As the paper entitled “very simple”, the analysis here is structural image analysis without any statistical group comparison. I wholehearted look forward to the authors' future progress on functional imaging. The flexibility of functional imaging is much more than structural imaging, thus the authors may would like to provide some discussion on that.

3. Group comparison and thresholding (with multiple comparison correction) are challenging to be reproduced, thus recently have been questioned a lot (including Eklund *et al.*, 2016 PNAS [1]). The authors' re-executable strategy may have some potential to prevent lots of p-hacking practices, thus deserve some discussion.

References

1. Eklund A, Nichols TE, Knutsson H: Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates.*Proc Natl Acad Sci U S A*. 2016; **113** (28): 7900-5 PubMed Abstract | Publisher Full Text

Is the rationale for developing the new method (or application) clearly explained?

Yes

Is the description of the method technically sound?

Yes

Are sufficient details provided to allow replication of the method development and its use by others?

Yes

If any results are presented, are all the source data underlying the results available to ensure full reproducibility?

Yes

Are the conclusions about the method and its performance adequately supported by the findings presented in the article?

Yes

Competing Interests: No competing interests were disclosed.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Referee Report 26 June 2017

doi:10.5256/f1000research.12766.r23503

?

Konrad Hinsen [1,2]

[1] Centre de Biophysique Moléculaire (UPR4301 CNRS), Rue Charles Sadron, Orléans, France

[2] Division Expériences, Synchrotron SOLEIL, Saint Aubin, France

The revision of the manuscript is based on an extensive rewrite of the underlying software environment. Platform independence has been enhanced significantly by the use of a Docker container for producing the reference results. The instructions for the software environment have been improved as well. Moreover, using the authors' Docker image does not require me to accept any license conditions. My main criticism of this work has been resolved: I can easily reproduce the authors' results, and I feel confident that others can do it as well.

The manuscript itself has also improved significantly, and most issues raised by the two reviewers have been addressed satisfactorily.

The only aspect which I still consider unsatisfactory is the discussion of numerical discrepancies. The manuscript still mentions a "numeric tolerance of 1e-6", without saying if this is an absolute or a relative tolerance. An inspection of the Python script check_output.py reveals that this tolerance value is not used explicitly there. The reference and actual values are compared using the NumPy function allclose(), which, according to its documentation, applies an absolute tolerance of 1.e-8 when called as in the script. The authors should (1) use an explicit threshold in their script, rather than relying on a default value that may change in future versions of NumPy and (2) quote this same value in the manuscript.

Furthermore, the only explanation given for the choice of threshold is "The threshold in the ‘check_output.py’ script is simply selected to catch the presence of any numerical difference between the test run and the reference run." Catching *any* numerical difference would imply a threshold of zero. The purpose of a threshold is to distinguish between "small" differences that are most likely due to platform-dependent roundoff from "significant" differences whose cause needs to be investigated. The choice of threshold therefore depends both on the numerical algorithms (potential for roundoff discrepancies) and on the scientific interpretation of the results (how big must a change be to influence the conclusions?), which is what should be explained in the article.

Competing Interests: No competing interests were disclosed.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Version 1

Referee Report 06 March 2017

doi:10.5256/f1000research.11627.r20114

?

Allan J. MacKenzie-Graham

Division of Brain Mapping, Department of Neurology, University of California, Los Angeles, Los Angeles, CA, USA

I found the paper to be clear and straightforward, easy to read and understand. I find the use of a NITRC-CE virtual machine as an execution environment in combination with Nipype an excellent mechanism for facilitating re-executability and documentation of a series of processing steps. Combined with the use of GitHub to keep track of the elements, I believe that this is a remarkably good and fairly easily implemented solution. This encapsulation is excellent for re-execution of an analysis or for applying exactly the same analysis to a novel set of data, however, as the authors state, it only begins to address reproducibility and generalizability across multiple execution environments.

In this context, it is not clear to me what the processing environment was on the Mac OS comparison test, specifically what version of the OS and what version of the FSL tools were used in each case. Second, Ubuntu 12.04 is used within NITRC-CE v0.42, but the Mac OS comparison appears to be done against a machine running Ubuntu 16.04. I assume that 16.04 is a typo, but it should be corrected - and if it is not a typo, then a justification for a change in OS version should be given. In previous work by myself[1] and others[2], we showed that differences in software version and compilation settings can lead to measurably different results. A statement expressing the software versions used and that the software was compiled using similar settings (e.g. same level of optimization, same architecture was used - x86 or x64, etc.) would be reassuring. I realize that this similarity in environment is implied by the mechanism used to install the software, but it should be stated explicitly.

Overall, this is an excellent manuscript and implementation for sharing re-executable neuroimaging results. The methods and reporting can readily be repeated and replicated, supporting the main thrust of the paper.

References

1. Mackenzie-Graham AJ, Van Horn JD, Woods RP, Crawford KL, Toga AW: Provenance in neuroimaging.*Neuroimage*. 2008; **42** (1): 178-95 PubMed Abstract | Publisher Full Text
2. Gronenschild EH, Habets P, Jacobs HI, Mengelers R, Rozendaal N, van Os J, Marcelis M: The effects of FreeSurfer version, workstation type, and Macintosh operating system version on anatomical volume and cortical thickness measurements.*PLoS One*. 2012; **7** (6): e38234 PubMed Abstract | Publisher Full Text

Competing Interests: No competing interests were disclosed.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 31 May 2017

David Kennedy, University of Massachusetts Medical School, USA

We thank the reviewer for their thoughtful review and helpful comments. We have revised the manuscript and design of this manuscript to meet many of the concerns raised, and we believe that this has resulted in an improved presentation. We have also reworked the repository and the way the experiment can be reproduced and extended.

Reviewer Comment: In this context, it is not clear to me what the processing environment was on the Mac OS comparison test, specifically what version of the OS and what version of the FSL tools were used in each case.
Response: In the various 'comparison runs' that we present, we now take more care in providing complete descriptions of the OS and software versions that are being used.

Reviewer Comment: Second, Ubuntu 12.04 is used within NITRC-CE v0.42, but the Mac OS comparison appears to be done against a machine running Ubuntu 16.04. I assume that 16.04 is a typo, but it should be corrected - and if it is not a typo, then a justification for a change in OS version should be given. In previous work by myself[1] and others[2], we showed that differences in software version and compilation settings can lead to measurably different results. A statement expressing the software versions used and that the software was compiled using similar settings (e.g. same level of optimization, same architecture was used - x86 or x64, etc.) would be reassuring. I realize that this similarity in environment is implied by the mechanism used to install the software, but it should be stated explicitly.
Response: As above, there is clearly some ambiguity as to how we presented the 'comparison runs' in the original manuscript that we try to clarify in this version. In the original, the NITRC-CE AWS Ubuntu 12.04 FSL 5.0.9 run was the 'reference' and we compared to runs of this workflow (using FSL 5.0.9 in all cases) on local MAC OS 10.10.4, and Ubuntu 16.04 platforms. In the current version, we add a Docker version of the of the workflow (FSL 5.0.9, Debian jessie (8.7)), and enhance the description of the comparison runs. Finally, thanks for the additional references which we now also include.

Competing Interests: None

Referee Report 21 February 2017

doi:10.5256/f1000research.11627.r20292

?

Konrad Hinsen [1,2]
[1] Centre de Biophysique Moléculaire (UPR4301 CNRS), Rue Charles Sadron, Orléans, France
[2] Division Expériences, Synchrotron SOLEIL, Saint Aubin, France

This article aims to demonstrate how a neuroimaging study can be published in such a way that readers can re-execute the complete workflow with (relative) ease. Although the concrete study used as an example is probably of little scientific interest, it could serve as a template and guideline for other researchers using the same software tools.

My main issue with this paper is that I was unable to re-execute the workflow, in spite of the authors' efforts to document the process. A detailed technical explanation is provided at the end of this review. In addition to technical obstacles that could in principle be overcome, the main issue is the use of the proprietary software package FSL, whose licence can be interpreted as prohibiting its use in the context of a review for F1000Research. It is not clear to me if using FSL via NeuroDebian would somehow solve this problem, because I did not succeed in obtaining the NeuroDebian docker image.

Another issue of the interpretation of numerical tolerances. The article says that the authors made comparisons "allowing for a numeric tolerance of 1e-6", but no explanation is given for this particular choice, nor is it said if this is an absolute or a relative error criterion. Table 1 shows numerical results obtained on different platforms, but provides no guide for their interpretation. How big would a difference have to to influence the scientific interpretation of the results?

Finally, readers wishing to take this example as a starting point for preparing their own studies reproducibly would benefit from a more extensive discussion of the technical choices and of the work required for actually composing, rather than simply consulting, the authors' code repository. For example, it is not stated explicitly anywhere that the workflow takes the form of a Python script (run_demo_workflow.py). The use of conda in re-creating an environment with precise versions of each software package is also not generally known and deserves more explanation.

Attempting to re-execute the example workflow

I tried to re-execute the workflow on a MacBookPro running under macOS 10.11.6, using both procedures explained in the authors' README.md.

1. "Within your current environment".

The first obstacle is "Make sure FSL is available in your environment and accessible from the command line." Where do I get FSL? Which version(s) are acceptable? Please provide at the very least a link to the software's Web page (https://fsl.fmrib.ox.ac.uk/).

I decided not to install FSL on my computer because I am not willing to accept the licence. It excludes commercial use, and in particular "use of the Software to provide any service to an external organisation for which payment is received". F1000Research offers reviewers a reduction on future article page charges, which could be interpreted as a form of payment. Beyond this specific legal issue, I also consider it unreasonable to request reviewers to register as users of proprietary software, providing personal data for marketing purposes in the process.

I did, however, continue the setup process to check it for completeness. The instruction "If you already have a `conda` environment, please follow the detailed steps below." lacks some precision: where exactly do I have to start if I already have a conda environment? The right answer is "at 'conda config --add channels conda-forge'", which I think is not obvious.

2. "Within Docker"

Running the Simple_Prep_docker under macOS ends with the error message

```
readlink: illegal option -- f
usage: readlink [-n] [file ...]
```

I modified the script, replacing "readlink" by "greadlink" from Homebrew's coreutils package. The next error message then is sed: -i may not be used with stdin

There are two uses of "sed -i" in the script, but for neither one it is obvious under which conditions it would erroneously act on stdin. I decided to give up at this point. Considering the use of apt-get in the script, it probably requires Debian or Ubuntu Linux anyway.

I am not sure the authors can do much to address this issue, given that writing platform-independent shell scripts is difficult to impossible, but they should at least say clearly in the installation instructions for which platforms they have actually tested the installation.

Competing Interests: No competing interests were disclosed.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 31 May 2017

David Kennedy, University of Massachusetts Medical School, USA

We thank the reviewer for their thoughtful review and helpful comments. We have revised the manuscript and design of this manuscript to meet many of the concerns raised, and we believe that this has resulted in an improved presentation. We have also reworked the repository and the way the experiment can be reproduced and extended.

Reviewer Comment: My main issue with this paper is that I was unable to re-execute the workflow, in spite of the authors' efforts to document the process. A detailed technical explanation is provided at the end of this review.
Response: We are sorry that this did not work as expected in your case. As further detailed in our response to your detailed technical explaination below, and in the github 'issue' (https://github.com/ReproNim/simple_workflow/issues/15), we hope that we have successfully ammended the procedure to make the system even more re-executable by including the Docker re-execution option.

Reviewer Comment: In addition to technical obstacles that could in principle be overcome, the main issue is the use of the proprietary software package FSL, whose licence can be interpreted as prohibiting its use in the context of a review for F1000Research. It is not clear to me if using FSL via NeuroDebian would somehow solve this problem, because I did not succeed in obtaining the NeuroDebian docker image.
Response: We agree that licensing issues are very important to consider. Many of the users in the neuroimaging community have FSL locally, and have consented to the FSL licensing terms; hence re-execution of this workflow incurs no additional licensing issues. Users using a NITRC-CE AWS instance are explicitly presented a page that that details the licensing terms of the software installed on that instance and use of the instance is with the acknowledgement of the licensing terms. Our initial Docker did not present these licensing terms, but in our new README for the repository and Docker, we have include a notice for commercial use.

Reviewer Comment: Another issue of the interpretation of numerical tolerances. The article says that the authors made comparisons "allowing for a numeric tolerance of 1e-6", but no explanation is given for this particular choice, nor is it said if this is an absolute or a relative error criterion. Table 1 shows numerical results obtained on different platforms, but provides no guide for their interpretation. How big would a difference have to to influence the scientific interpretation of the results?
Response: We now provide more information on this issue. The threshold we apply in the 'check_output.py' script is simply selected to catch the presence of any numerical difference between the test run and the reference run. The default in the version of numpy used is in the present container is 1e-5. The biological interpretation of these differences is multi-faceted. On the one hand, the correlation of the volumetric results within each individual subject, and in aggregate across the population is very high (0.918 - 1.000). On the other hand, we see, per structure, a range of volumetric differences that reveals a large span of percentage of structure differences, differences that are, not surprisingly, dependent upon the overall size of the structure itself. The extremes of this distribution of the average difference (provided in Table 1) can be as large as 7.7% for the right amygdala. Sources of volumetric variance in this range can be troubling as biological changes on this order of volumetric difference can otherwise be the types of changes that studies are designed to observe.

Reviewer Comment: Finally, readers wishing to take this example as a starting point for preparing their own studies reproducibly would benefit from a more extensive discussion of the technical choices and of the work required for actually composing, rather than simply consulting, the authors' code repository. For example, it is not stated explicitly anywhere that the workflow takes the form of a Python script (run_demo_workflow.py). The use of conda in re-creating an environment with precise versions of each software package is also not generally known and deserves more explanation.
Response: We now add more discussion of the topic of approaches that others can take to generate more re-executable workflows, and the various challenges in representing the execution environment. Specifically, conda (https://conda.io/) is a cross-platform package manager and handles user installations for many packages into a controlled environment. Unlike many operating system package managers (e.g., yum, apt), Conda does not require root privileges. This allows individuals to replicate isolated virtual environments easily without requiring system administrator help. Conda uses standard PATH variables to isolate the environments. Coupled with Anaconda cloud and conda-forge, Conda is capable of installing versioned dependencies of Python and other packages.

Reviewer Comment: Reviewer Attempting to re-execute the example workflow
The first obstacle is "Make sure FSL is available in your environment and accessible from the command line." Where do I get FSL? Which version(s) are acceptable? Please provide at the very least a link to the software's Web page (https://fsl.fmrib.ox.ac.uk/).
Response: Links are now provided, and we remind the reader of this page that 5.0.9 is the version of the reference run. Again, half of the point of this exercise is to provide the opportunity to see what happens, numerically, IF the user is already using a different environment or version.

Reviewer Comment: I decided not to install FSL on my computer because I am not willing to accept the licence. It excludes commercial use, and in particular "use of the Software to provide any service to an external organisation for which payment is received". F1000Research offers reviewers a reduction on future article page charges, which could be interpreted as a form of payment. Beyond this specific legal issue, I also consider it unreasonable to request reviewers to register as users of proprietary software, providing personal data for marketing purposes in the process.
Response: For these various reasons, we have now elected to also include a Docker version of this workflow that precludes the need to locally download specific software to one's local computer. We have also included a commercial use statement from the FSL developers in the README. While most members of the neuroimaging community have access to FSL, in the future, we hope to move to FOSS versions of imaging tools to make such testing accessible to a broader community.

Reviewer Comment: I did, however, continue the setup process to check it for completeness. The instruction "If you already have a `conda` environment, please follow the detailed steps below." lacks some precision: where exactly do I have to start if I already have a conda environment? The right answer is "at 'conda config --add channels conda-forge'", which I think is not obvious.
Response: We have updated the script to test for existence of software and create a standalone environment that does not interfere with any user environment directly. The user can activate this environment if the user so chooses. Since FSL download on non-Debian systems requires going to a site, we simply recommend that people do so on their own. The same script is also used in the Docker container. We have updated this in the README for the repository.

Reviewer Comment: 2. "Within Docker"
etc.
I am not sure the authors can do much to address this issue, given that writing platform-independent shell scripts is difficult to impossible, but they should at least say clearly in the installation instructions for which platforms they have actually tested the installation.
Response: We thank the reviewer for alerting us to this compatibility issue. As we have discussed in the github issue, as a workaround, we first have pre-generated a Dockerfile so that reviewer could run the analysis and validate our findings. We have recently finalized our changes to the script to make it compatible with both Linux and OSX so it should now run natively on reviewers infrastructure without issues. We very much appreciate the time and effort the reviewer took to help finalize this script.

Competing Interests: None