A Very Simple Re-Executable Neuroimaging Publication

Authors:
Satrajit S. Ghosh[1,2] (0000-0002-5312-6729), Jean-Baptiste Poline[3] (0000-0002-9794-749X), David B. Keator[4] (0000-0001-5281-5576), Yaroslav O. Halchenko[5] (0000-0003-3456-2493), Adam G. Thomas[6] (0000-0002-2850-1419), Daniel A. Kessler[7] (0000-0003-2052-025X), David N. Kennedy[8] (0000-0002-9377-0797)

[1]McGovern Institute for Brain Research, MIT, Cambridge, Massachusetts 02139, USA.
[2]Department of Otology and Laryngology, Harvard Medical School, Boston, Massachusetts 02114, USA.
[3]Henry Wheeler Jr. Brain Imaging Center, Helen Wills Neuroscience Institute, University of California, Berkeley, California 94720, USA.
[4]Department of Psychiatry and Human Behavior, Department of Computer Science, Department of Neurology, University of California, Irvine, California 92697, USA.
[5]Department of Psychological and Brain Sciences, Dartmouth College, Hanover, New Hampshire 03755, USA.
[6]Data Science and Sharing Team, NIMH IRP, Bethesda, MD, USA.
[7]Department of Psychiatry and Department of Radiology, University of Michigan, Ann Arbor, Michigan 48109, USA.
[8]Eunice K. Shriver Center and Department of Psychiatry, University of Massachusetts Medical School, Worcester, Massachusetts, 01605, USA.

Title:
A Very Simple Re-Executable Neuroimaging Publication

Abstract:
Reproducible research is a key element of the scientific process. Re-executability of the 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 that 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.

Keywords:

Neuroimaging analysis, re-executable publication, reproducibility

Main Body:

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 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, Schlange, & Asadullah, 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, Nelson, & Simonsohn, 2011; Simonsohn, Nelson, & Simmons, 2014); completeness of methods description, the written text of a publication can not 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 does 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 our 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 to: take some specified “Data”, apply a specified “Analysis”, 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

Because we do not really characterize data, analysis, and results very exhaustively in the current literature, the concept of 'similar' has 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 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 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 at the NITRC Image Repository. The key elements to publication re-executability is 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 that 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).

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 ran 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 entire publication can be found in the http://github.com/ReproNim/simple_workflow repository.

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, Haselgrove, Riehl, Preuss, & Buccigrossi, 2016). We queried (on 1-Jan-2017) for publically available imaging data of subjects aged 10-15 years old, with a MRI field strength of 3 Tesla. This query returned 24 subjects, 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 URL's (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 (http://doi.org/10.18116/C6C592). We captured this data in tabular form, and stored in a publicly accessible Google Drive spreadsheet available at: https://docs.google.com/spreadsheets/d/11an55u9t2TAf0EV2pHN0vOd8Ww2Gie-tHp9xGULh_dA/edit?usp=sharing. Representative images from this collection are shown in Figure 1.

[Insert Figure 1 about here…]

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 (FSL, RRID:SCR_002823, (Jenkinson, Beckmann, Behrens, Woolrich, & Smith, 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, (K. Gorgolewski et al., 2011)) to facilitate workflow execution and provenance tracking. The workflow is available at https://github.com/ReproNim/simple_workflow/tree/e4063fa95cb494da496565ec27c4ffe8a4901c45. The workflow also includes an initial step that reads the worksheet in a public Google drive directory, identified above, 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. 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).

[Insert Figure 2 here]

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 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 12.04 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 Version 42, with the following specific identifier (AMI ID): ami-ce11f2ae.

Setting up the software environment on a different machine

To re-execute a workflow on a different machine or cluster than the one used originally, the first step is to set up the software environment. A README.md file in the GitHub repository describes how to set up this environment on GNU/Linux and MacOS systems. We assume FSL is installed and accessible on the command line. A Python environment can be set up and the Nipype workflow re-executed with a few shell commands as noted in the README.md.

The Reference Run

We performed the analysis (the above described workflow applied to the above described data, using the described computational system) 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) Launch an instance of NITRC-CE version v0.42 from AWS (we selected a 16 core c4.8xlarge instance type)
2) Execute the following commands on this system to install the workflow, configure the environment and run the workflow:

```
> curl -Ok
https://raw.githubusercontent.com/ReproNim/simple_workflow/e4063fa95cb494da496565ec
27c4ffe8a4901c45/Simple_Prep.sh
> source Simple_Prep.sh
> cd simple_workflow
> source activate bh_demo
> python run_demo_workflow.py --key  \
      11an55u9t2TAf0EV2pHN0vOd8Ww2Gie-tHp9xGULh_dA
```

The Details within the Simple_Prep.sh Script: In order to run this workflow we need both FSL tools and a Python environment to run the Nipype workflow. We achieve the specification of the Python environment using conda, a package manager that can be run without administrative privileges across different platforms. An environment specification file ensures that specific versions of Python and other libraries are installed and used. The setup script then downloads the Simple Workflow repository and creates and activates the specifically versioned Python environment for Nipype.

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 on the computational environment:

```
> 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.

Re-execution on other systems

While the Reference Analysis was run using NITRC-CE (Ubuntu 12.04) running on AWS, 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 and an Ubuntu 16.04 platforms.

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) that 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 (Wheezy) 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/b0504592edafb8d4c6336a2497c216db5909ddf6/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 reference run is re-executed in the same environment there is no observed difference in the output. We also compared the execution of the reference run and re-execution in a separate MacOS environment. Table 1 indicates the numeric differences found in these alternate system example runs.

[Insert Table 1 here…]

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, specifically computational operating system and software version (Glatard et al., 2015). In this work, we replicate the observation that, despite an exact match on the data and workflow, the results of analysis differ (if even only very slightly) between execution in different operating systems. While in this case, the volumetric differences are not 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 the ways to improve, simplify, and standardize 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 (K. J. Gorgolewski et al., 2016), NIDM (K. J. Gorgolewski et al., 2016), BDBags (http://bd2k.ini.usc.edu/tools/bdbag/), etc.) will simplify how experimental data is assembled for sharing and used 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 4 ‘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.

Data (and Software) Availability:

The data used in this publication is available at DOI and referenced in Google doc. It is originally served from the NITRC-IR, 1000 Functional Connectomes project, Ann Arbor and New York sub-projects. The software is available on GitHub at: https://github.com/ReproNim/simple_workflow.

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

Author Contributions (if more than one author):
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:
The authors have no competing interests related to the work reported in this publication.

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

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).

References

Biswal, B. B., Mennes, M., Zuo, X.-N., Gohel, S., Kelly, C., Smith, S. M., … Milham, M. P. (2010). Toward discovery science of human brain function. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(10), 4734–4739. https://doi.org/10.1073/pnas.0911855107

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365–376. https://doi.org/10.1038/nrn3475

Glatard, T., Lewis, L. B., Ferreira da Silva, R., Adalat, R., Beck, N., Lepage, C., … others. (2015). Reproducibility of neuroimaging analyses across operating systems. *Frontiers in Neuroinformatics*, *9*, 12.

Gorgolewski, K., Burns, C. D., Madison, C., Clark, D., Halchenko, Y. O., Waskom, M. L., & Ghosh, S. S. (2011). Nipype: a flexible, lightweight and extensible neuroimaging data processing framework in python. *Frontiers in Neuroinformatics*, *5*, 13. https://doi.org/10.3389/fninf.2011.00013

Gorgolewski, K. J., Auer, T., Calhoun, V. D., Craddock, R. C., Das, S., Duff, E. P., … Poldrack, R. A. (2016). The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. *Scientific Data*, *3*, 160044. https://doi.org/10.1038/sdata.2016.44

Hanke, M., & Halchenko, Y. O. (2011). Neuroscience Runs on GNU/Linux. *Frontiers in Neuroinformatics*, *5*, 8. https://doi.org/10.3389/fninf.2011.00008

Ioannidis, J. P. A. (2005). Why Most Published Research Findings Are False. *PLOS Medicine*, *2*(8), e124. https://doi.org/10.1371/journal.pmed.0020124

Jenkinson, M., Beckmann, C. F., Behrens, T. E. J., Woolrich, M. W., & Smith, S. M. (2012). FSL. *NeuroImage*, *62*(2), 782–790. https://doi.org/10.1016/j.neuroimage.2011.09.015

Kennedy, D. N., Haselgrove, C., Riehl, J., Preuss, N., & Buccigrossi, R. (2016). The NITRC image repository. *NeuroImage*, *124*(Pt B), 1069–1073. https://doi.org/10.1016/j.neuroimage.2015.05.074

Prinz, F., Schlange, T., & Asadullah, K. (2011). Believe it or not: how much can we rely on published data on potential drug targets? *Nature Reviews Drug Discovery*, *10*(9), 712–712. https://doi.org/10.1038/nrd3439-c1

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychological Science*, *22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve: a key to the file-drawer. *Journal of Experimental Psychology. General*, *143*(2), 534–547. https://doi.org/10.1037/a0033242

Figures and Tables:

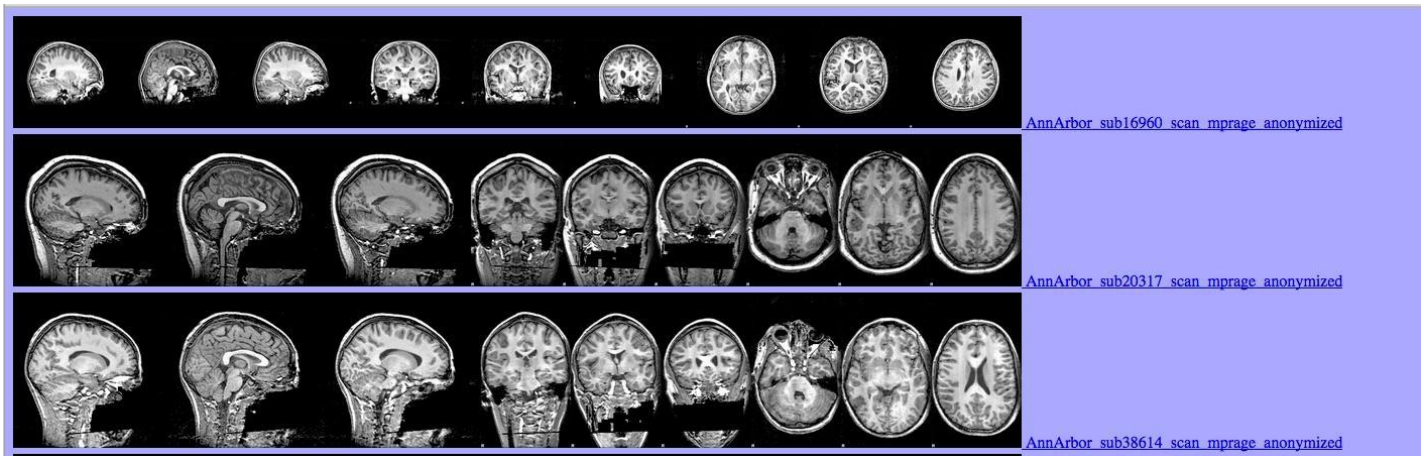

Figure 1: Example images from a subset of three of the subjects image datasets used.

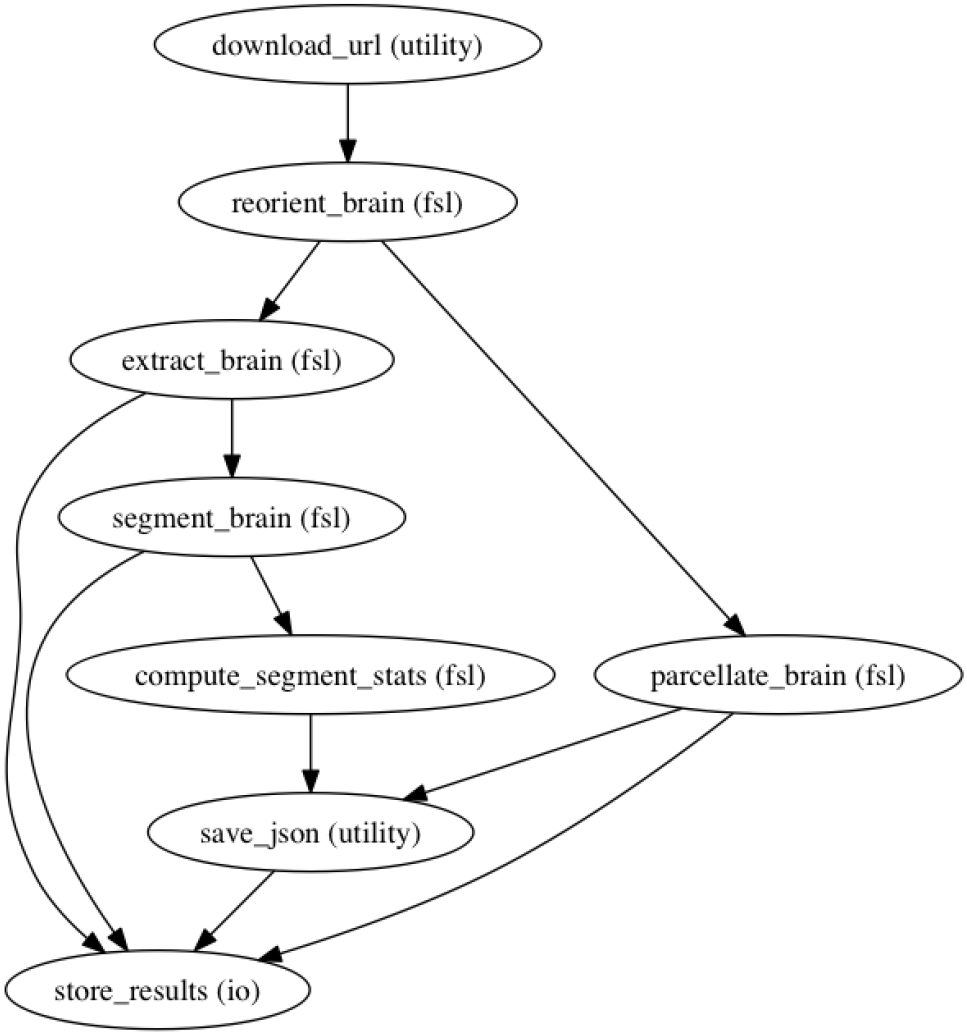

Figure 2. Workflow diagram. The sequence and dependence of processing events used in this example re-executable publication.

Table 1

Hemisphere	Region	Reference Run (AWS Ubuntu 12.04) Mean Volume (mm3)	Reference Run (AWS Ubuntu 12.04) STD	Mac OSX (10.10.4) Mean Volume (mm3)	Mac OSX (10.10.4) STD	Difference Mean Volume (mm3)	Difference STD	Mean Absolute Volume	Percent of Reference
Left	Accumbens	423.0	176.0	420.3	177.1	2.7	34.6	24.0	5.7
	Amygdala	755.5	267.0	765.3	276.3	-9.8	69.7	30.3	4.0
	Caudate	3467.6	741.6	3466.9	735.7	0.7	108.1	48.7	1.4
	Hippocampus	3008.5	1047.2	2996.0	1029.7	12.5	57.6	40.6	1.3
	Pallidum	1533.9	396.8	1519.7	391.7	14.2	30.2	19.1	1.2
	Putamen	4459.3	1245.8	4465.4	1225.5	-6.2	62.3	44.1	1.0
	Thalamus	7290.9	1498.1	7312.4	1516.9	-21.5	70.6	52.7	0.7
Right	Accumbens	350.0	145.6	363.3	149.1	-13.3	42.3	23.4	6.7
	Amygdala	796.9	305.2	809.9	299.4	-13.0	116.9	64.4	8.1
	Caudate	3433.3	912.9	3433.7	921.6	-0.4	27.6	19.7	0.6
	Hippocampus	3132.7	986.3	3153.3	995.4	-20.6	60.9	41.4	1.3
	Pallidum	1541.4	389.4	1541.4	382.0	0.0	23.4	12.8	0.8
	Putamen	4549.5	1382.8	4523.1	1348.5	26.4	114.3	73.7	1.6
	Thalamus.Proper	6959.6	1347.3	6962.0	1367.0	-2.4	54.7	33.5	0.5
Total	CSF	173256.8	42784.0	173518.5	42496.2	-261.7	749.4	267.9	0.2
	Gray Matter	628272.3	158697.3	627784.0	159151.0	488.4	1442.8	493.8	0.1
	White Matter	467644.0	102424.3	467601.5	102476.0	42.5	337.7	93.2	0.0
	Brain	1269173.2	290423.7	1268904.0	290726.9	269.2	1081.5	306.3	0.0

Table 1: Summary volumetric results from the simple workflow for the 24 subjects. Results are shown from the Reference Run (AWS Ubuntu 12.04) and a comparison run executed on a Mac OS X (10.10.4) system. The mean differences between these two systems are also summarized.