Characterizing and Optimizing Generalizability and Sensitivity in Neuroimaging Through Repeated Analysis

Gregory Kiar
April 6th, 2018

Outline

This proposal follows the template established in Appendix A of the:
McGill University Department of Biological and Biomedical Engineering Guidelines

Summary of research proposal

With an explosion in data collection across nearly all scientific disciplines, we are entering the big-data era of scientific exploration. From the Sloan Digital Sky Survey [1] in cosmology, to the UK BioBank [2], Human Connectome Project [3], and the Consortium of Reproducibility and Reliability (CoRR) [4] in Neuroscience, initiatives are being launched to federate this data-tsunami, many with particular emphasis and commitment towards open data-sharing.

As access to data has increased drastically, and the availability of diverse and customizable processing tools has grown similarly, it has become apparent that a lack of reproducibility in data analysis is becoming a plague in many disciplines of science [5]. While on occasion this is the result of p-hacking (i.e. the modification of analyses in search for significant results), it is often due to much more innocent means such as software bugs [25]. The growth of open science and increased availability of open datasets also poses a risk for further failure to adequately reproduce and generalize findings due to a much higher degree of variation in data quality and acquisition conditions. Unless addressed, science will continue to fail to do justice to the extraordinary datasets we possess. Computational sciences have an opportunity to develop more rigorous practices and tools which lower the barrier for producing reproducible analyses.

The main hypothesis I wish to test in my work is that optimizing pipelines or analysis tools to provide accurate results on small, largely homogenous cohorts of data (as is currently commonplace in neuroimaging), results in claims which are overfit to their data, and resultantly tool settings which may perform inadequately on large, heterogeneous datasets. The core outcome of this project will be the ability to characterize the generalizability and sensitivity of neuroimaging analyses, and provide an accessible method for researchers to optimize their scientific designs with respect to the selection of tools, hyperparameters, and datasets.

Statistical procedures and metrics must be developed which enable testing the generalizability and sensitivity of neuroimaging experiments to perturbations. The development of computational infrastructures which enable accessible data discovery, deployment of analyses, provenance recording, sample perturbation, and redistribution of results are necessary to enable the development of models described above. Thus, my contributions will be:

- the creation of infrastructures for accessible and reproducible neuroimaging, enabling
- the development of methods and models to characterize pipelines, ultimately informing
- the a priori optimization of context-specific pipeline selection in neuroimaging.

The successful completion of my proposed project has the potential to transform the process by which scientists and tool developers quantify their analyses, and will lead to the production of more richly described, trustworthy scientific studies and tools, and more effective use of open datasets. I will describe the progress and accomplishments to date and the plan for future work in the sections that follow.

Summary of progress

Dependencies

In order for the computational tools proposed above to be achievable and sustainable, it is imperative that they sit atop the shoulders of existing and emerging standards. In particular, the two areas of standardization relevant to these projects are: tool and data representation.

The *Boutiques* specification documents the command-line execution instructions of tools, and provides an interface for validating their description, execution, the existence of expected outputs, and more. I have contributed to this project, including leading the development of a Python package [6]. The advent of Boutiques in the context of pipeline deployment is that it enables clear documentation and repeatable execution of tools across platforms, making it an important backbone of the proposed projects. The Boutiques paper, for which I am the second author, was recently accepted to Gigascience [7].

In the realm of data organization, the *Brain Imaging Data Structure (BIDS)* [8] has experienced tremendous adoption in the neuroimaging community. This standard prescribes an organization of folders and files on disk, and serves as a low-barrier solution to unambiguously and accessibly share datasets. From this, along with other I worked on the development of the BIDS app initiative [9] which prescribes specific command-line arguments and inputs for tools to easily be run on BIDS datasets. Many neuroimaging tools have adopted and contributed to the development of this standard, including a pipeline for structural connectome estimation, which I developed: ndmg [10]. This standard enables both the bulk querying of metadata and rapid deployment of pipelines across disparate datasets.

Planned Projects

To have the largest impact, the proposed infrastructures should enable both the *execution* of neuroimaging pipelines, and the *discovery and aggregation* of datasets. First, I have built Clowdr [11]: an execution environment for pipelines based on Boutiques, which allows users to develop tools locally, deploy them on clouds (i.e. Amazon) or clusters (i.e. Compute Canada), and easily monitor and share the results through the web. Second, I have begun development of Apine: a web-query engine for exploring BIDS-organized datasets. It can perform cross-study searches of arbitrary dataset qualities captured in BIDS.

Both of these projects are open source, were accepted for posters and software demonstrations at the Organization for Human Brain Mapping (OHBM) 2018 conference, and will be expanded upon in the following section.

I have yet to undertake work on the statistical portion of my proposed project, but will expand upon previous work of mine and others in this area in the following section.

Research proposal

Overview

As was summarized on Page 1, the eventual objective of my work is to develop a simple mechanism for evaluating and characterizing neuroimaging pipelines leading to more informed pipeline development, selection, and more powerful scientific discoveries. This project must be undertaken in several steps since some of the technical requirements required to efficiently perform the aforementioned analysis are lacking.

The first chapter in my thesis will cover the development of these tools, Clowdr and Apine, which increase the accessibility to both perform complex analyses in high performance computing environments and discover broad and diverse datasets for analysis, respectively.

The second chapter of my thesis will concern the development of a paradigm for neuroimaging analysis that enables the characterization of tools and analyses, so that their sensitivity and generalizability may be evaluated. This will include the use of the platforms developed in Chapter 1, and target commonly used tools to demonstrate the bias-variance trade-off as it is embodied in this context, and provide researchers with a method of performing this analysis themselves during tool development.

The third and final chapter of my thesis will explore the effect of optimizing pipelines for specific contexts based on the characterizations obtained in Chapter 2, and demonstrate the impact that tool selection and tuning have on scientific claims across a variety of neuroimaging contexts, including the processing of i) small homogeneous datasets, ii) large homogeneous datasets, iii) small heterogeneous datasets, and iv) large heterogeneous datasets. This work will establish a principled method for context-aware a priori pipeline selection in neuroimaging.

Throughout this proposal I will regularly reference definitions of repeatability, replicability, and reproducibility. In the interest of clarity, I have define what each of these terms means in the lexicons I have adopted in the Glossary following this document. In particular, I adopt the meanings defined by the *Association for Computing Machinery* [12]. In the context of inference, I adopt the Goodman et al. [13] terminology. A succinct guide to these and other terminologies is available in [14].

Chapter 1: Accessible and Repeatable Scientific Computing

The backbone of many neuroimaging analyses and claims relies on two key components: the data being processed, and the tool doing the processing. External conditions, such as where the tools were being deployed, the format of the data, which hyperparameters were used, are all certainly of importance, but rarely are sufficiently recorded or documented in order for studies to be adequately reproduced, let alone meta-analyzed or staged alongside similar findings [15].

While documenting the above and other similar features may help in the replicability of pipelines themselves, the datasets being processed often suffer similarly. Here we will explore the case in which this metadata is recorded sufficiently for replication.

Given that researchers produced a replicable study, we still suffer from a clear understanding of what role each design decision, data point, or tool setting played on the resulting picture. There currently exists no known methods (other than ad-hoc) which easily enables studying the effect of perturbations on findings, by allowing iterative modification of both the data being processed and the tool settings.

The two tools I propose to develop each enhance the reproducibility and ability to perform perturbations on analyses, *Clowdr* from the perspective of pipelines, and *Apine* regarding datasets. Clowdr will serve as a micro-environment for deploying and recording executions and scientific analyses. Apine will serve as a data discovery tool and enable the querying of disparate datasets, which can both increase researchers' ability to identify and access openly-used datasets that meet their desiderata, but also dramatically increase sample size for their experiments.

Clowdr

The primary objective of Clowdr is to increase the accessibility, reproducibility, and permutability of scientific analyses. With respect to accessibility, Clowdr enables users to develop and execute tools locally, and seamlessly transition to executing these same tools either on high performance computing clusters or commercial clouds, while monitoring their progress and sharing the results. This paradigm will shorten the feedback and development process for scientific software developers, and easily enable meta-analyses, such as hyperparameter sweeps and optimizations with a low barrier to entry.

There exist many platforms which aid in scientific computing, including CBRAIN [16], BrainCODE [17], OpenNeuro [18], and LONI [19] in the neuroimaging space alone. Each of these tools enables neuroscientists to interact with datasets and established computational tools from the comfort of their web browsers, and have been an incredible asset for hundreds of scientists. These platforms often rely on an underlying structure for representing how tools are defined and will be executed; in the CBRAIN case, this is the Boutiques [7] command-line descriptive framework. These definitions enforce standards on the tools being executed, bringing clarity and consistency to their execution. While Clowdr adopts this approach and the Boutiques framework for standardizing tool representation, the key difference with these science-as-a-service infrastructures is that the tools available to users are restricted to "production-level" workflows and are not conducive to rapid iteration or tool development. Additionally, the data derivatives produced in each of these platforms are tied to the infrastructure and require both data ingest and extraction for subsequent statistical evaluation.

Clowdr operates as a “microservice,” meaning that no persistent server must be running, and that it can be very easily used with minimal installation or configuration. It has been tested and developed to run locally (Mac OSX, Linux), on clusters (Compute Canada; SLURM environment), and commercial clouds (Amazon Web Services). The ability to flexibly deploy and modify executions through Clowdr allows permutation and perturbation tests to be performed accessibly on tools and data, and perform hyperparameter tuning. This tool has the potential to lower the barrier to entry for testing in scientific tool development, and result in higher quality and better characterized pipelines, and ultimately more impactful scientific claims.

I have developed Clowdr for special consideration to neuroimaging, as it has underlying support for data organizations such as BIDS, which, when paired with Boutiques, dramatically reduce the burden on scientists when specifying the executions they wish to perform. Because of this, the combination of Clowdr and Boutiques have become somewhat of a native execution environment for BIDS apps, which immediately opens the door to using it with a variety of tools such as FSL [20], FreeSurfer [21], AFNI [22], SPM [23], MRtrix [24], ndmg [10], and others.

The Clowdr python package is open-source on Github, https://github.com/clowdr/clowdr, hosted on the Python Package Index, https://pypi.python.org/pypi/clowdr, and has been accepted for publication as a poster and software demonstration at the OHBM 2018 annual meeting. While development will continue on this project, the minimum viable product satisfying all of the necessary desiderata as highlighted above has already been accomplished, and further work will go into extending this tool.

Apine

The goal of Apine is to increase the findability and query-ability of datasets, and enable that to easily translate to data selection when performing experiments. While initiatives like BIDS have led to dramatically more accessible open datasets, and tools such as pyBIDS or Datalad lend to investigating and indexing these datasets from a command-line, there is currently no lightweight microservice for exploring collections of disparate publicly available datasets in the web. In the context of both my work and neuroimaging more broadly, the ability to identify datasets which are compatible with an analyses enables better characterization of outcomes, reduces the likelihood of overfitting to smaller collections of data, and has potential to increase the statistical power of derived claims.

Currently, the Apine project, a part of the International Neuroinformatics Coordinating Facility (INCF) organization, operates on collections of JavaScript Object Notation (JSON) files which have been generated to summarize BIDS datasets, and launches a lightweight server with a RESTful endpoint for users to query. These queries are intended to be flexible, and users can provide restrictions; for instance, a researcher may be interested only in datasets that contain functional and anatomical imaging modalities, where the anatomical sessions collected include MRI sequences such as T1w and Inplane T2w images, and the functional sessions consist of at least two runs (this query is shown below in Listing 1). These types of queries being available

through web browsers enables powerful interfaces to be developed, which dramatically increase the findability and searchability of datasets.

```
/dataset?datasetID&modality=func,anat&filename_key=inplaneT2,T1w,run-02
```

Listing 1: Apine query of a collection of BIDS datasets for functional and anatomical modalities, inplaneT2w and T1w anatommical scans, and at least two runs of the functional task

While Apine exists currently as a proof-of-concept, the next steps for development include the extension to include a broader range of query criteria, such as participant-specific, session-specific, or other types of metadata. Additionally, as the tools surrounding BIDS metadata management have become more mature, Apine will adopt alternative back-ends for datasets, including either the pyBIDS package or Datalad.

Development on Apine has been largely limited to a series of collaborative workshops with members of INCF, key developers of the BIDS specification, and other leaders in metadata management, provenance, and querying in neuroimaging. These include but are not limited to: Jean-Baptiste Poline, Satra Ghosh, Chris Gorgolewski, Yarick Halchenko, and Michael Hanke. Continuing to foster these collaborations will enable this project to both excel and be sustained as these data standards evolve.

As mentioned above, Apine is open-source on Github, https://github.com/INCF/apine, and has been accepted for publication as a poster and software demonstration at the OHBM 2018 annual meeting.

Chapter 2: Characterization of pipeline generalizability and sensitivity

Given that data are now more easily discoverable, and pipelines easily deployable and modifiable, we have the ability to process an abundance of data across a range of operating points of our tools. Extending this, we can characterize our pipelines and evaluate contexts and hyperparameters for which our tools are more sensitive to variation, or stable/generalizable. The central research question I would like to answer here, is whether we can meaningfully characterize the performance of tools to get a sense of the generalizability and stability of their derivatives a priori to performing experiments on new data.

Not unlike the receiver-operating characteristics curve that is commonplace in predictive modeling, characterizing performance based upon two theoretically different objectives, the sensitivity to variation and the robustness or stability to noise, has potential to provide a rich description of the quality of a tool being used and provide insight into the settings that should be chosen based on the nature of the question being asked (see illustration in Figure 1).

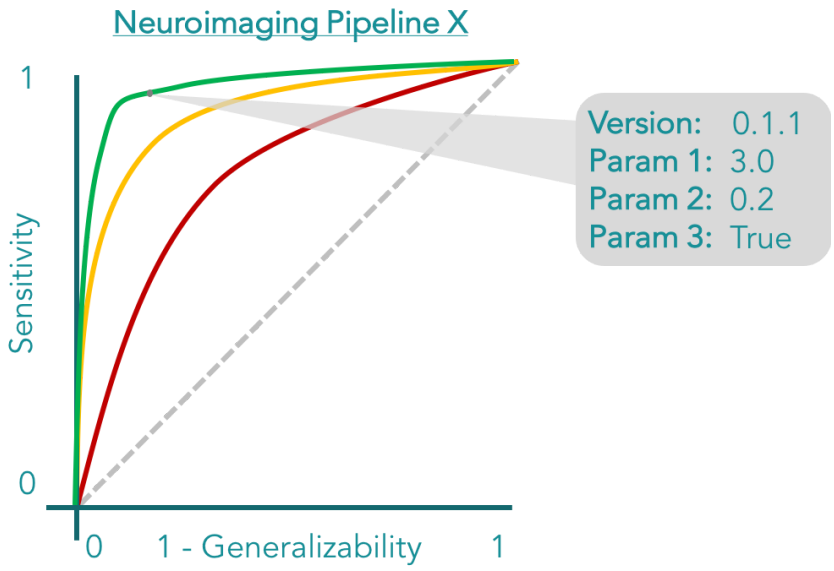

Figure 1: illustration of a potential characterization curve, weighing the sensitivity against the generalizability of analyses.

As awareness of the importance of reproducibility has continued to increase in the neuroimaging community, several groups have shone light on tools that have produced unstable results, calling the derived claims into question. In some cases, these differences were due to software bugs which have been since fixed [25]; in other cases, there are simply presently-unexplained differences between software stacks performing ideally similar operations [26].

The NPAIRS [27] project tackled a problem similar to what I propose in the context of functional MRI statistical parametric maps (SPMs). They developed a reproducibility coefficient, and using a cross-correlation method they term “split-half resampling” use it to evaluate the stability of SPMs derived from subsets of their dataset. While I admire the approach taken by the authors, there are several obvious limitations I’d like to address in my work. First, I’d like to more generally address these questions in MRI, rather than specifically functional MRI maps. Second, I plan to address the reproducibility of *claims derived from data*, rather than just the *derived data itself*. This distinction is important, as ultimately the models I develop to characterize stability or robustness of claims will be addressing and characterizing the learned outcomes, as opposed to specific derivatives, which makes this method more generally applicable. Where Strother et al. [27] evaluated the “*results reproducibility*,” as defined by Goodman et al. [13], I seek to evaluate the “*inferential reproducibility*” of claims drawn from derived data of the tools in question. In balancing sensitivity or accuracy with reproducibility as described above, ground-truth classifications will be used where possible (i.e. age, sex, scanner manufacturer, etc.) to inform inference tasks and balance optimizing the result stability with performance.

The successful completion of this project and the platforms developed in Chapter 1 synergistically enable tool developers and researchers to run a variety of "vibration-testing," i.e. data perturbation, permutation, sample selection, hyperparameter selection, etc., resulting in a) less overfitting of tools, b) better understanding of the impact of hyperparameter selection in pipelines, and importantly, c) more generalizable scientific discoveries.

I anticipate this project will take approximately 1.5 years, and I will commence development during the summer of 2018. Initially, I will perform a detailed literature review on performance metrics of both processing pipelines and machine learning algorithms, to evaluate the state of the art in terms of specific models which can be leveraged here. Simultaneously, I will use Clowdr to process data from the Consortium of Reliability and Reproducibility (CoRR) [28] dataset with various BIDS apps. I will use ndmg, Freesurfer, and C-PAC to process diffusion, structural, and functional data in this dataset, respectively. From the processed data, I will extract common features (i.e. tract length, cortical thickness, and region-wise correlations, respectively). I will iterate in this process with varieties of data permutations (i.e. number of samples, which samples) and perturbations (i.e. 1-voxel vibrations, Gaussian blurring, salt-and-pepper noise) and hyperparameter settings, which will be determined both through literature review of common selections, and discussions with tool developers. I will then perform bootstrapped classification across available phenotypic data, obtaining a performance profile for each operating point. I will then fit the models discovered or developed as a result of my literature review to these profiles, and attempt to characterize the performance.

I believe that this is an impactful and valid contribution for me to undertake as I am uniquely qualified to perform all of the required computations and analyses in a streamlined, robust, and provenance-preserving manner necessary to properly carry out this large experiment. Additionally, as the developer of the ndmg pipeline, a close collaborator with the lead developer of C-PAC, Cameron Craddock, and member of the Evans' lab with considerable experience in structural MR image analysis, I have the expertise available to design and carryout realistic experiments on the derived data products. The choice of CoRR dataset was because there are subsets of this collection which are largely homogenous, while others are much more heterogeneous, in terms of geographical location, participant demographics, and scan quality. This makes it a useful dataset for developing a model for generalizability. My experience in pipeline optimization and development of statistical evaluation methods for data reliability [29] also puts me in a uniquely qualified position to undertake the task of broadly classifying the generalizability and accuracy of arbitrary neuroimaging pipelines.

Chapter 3: Pipeline optimization in varying inference contexts

With a better understanding of the behaviour of tools being deployed on our data, we can now attempt to use this knowledge to optimize pipelines for new datasets a priori. This has potential to dramatically increase the generalizability of subsequent claims, an importantly remove the possibility p-hacking. By leveraging the understanding our tools performance at various operating points, and metadata regarding the size and degree of homogeneity about the dataset

we wish to process, we can make informed decisions about our analytical designs towards answer puzzling scientific questions. Here I wish to investigate whether specific regions on the sensitivity/generalizability curves created in Chapter 2 should be chosen targeted based on the type of data being processed, or type of analysis being performed (see illustration in Figure 2).

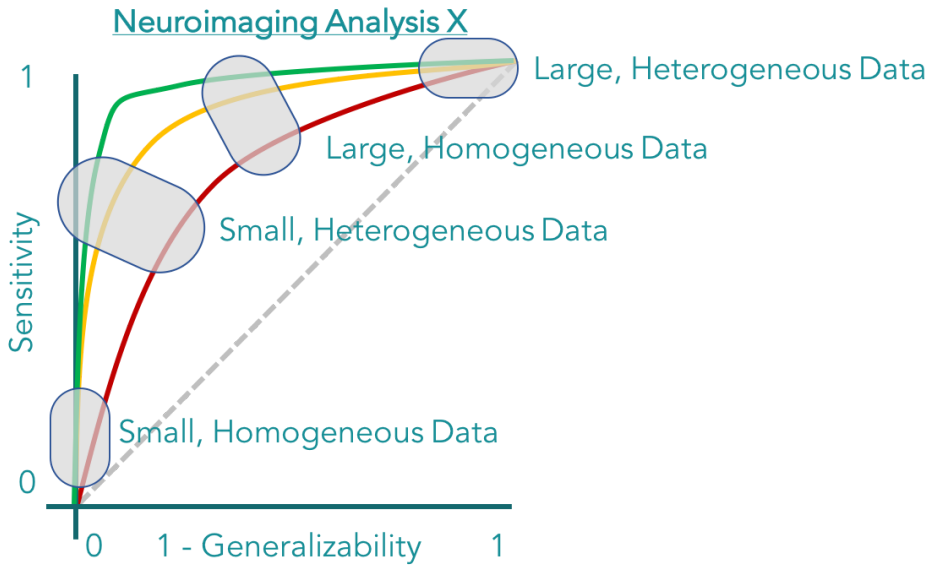

Figure 2: illustration of a potential characterization curve, with particular emphasis on identified regions of interest based on various processing contexts, such as dataset size and degree of homogeneity

While the number of publicly available datasets growing daily, an important admission is that the quality and consistency of these data are not uniform. Data are collected from magnets of different strengths, with varying collection protocols, on participants of different demographics which may have tendencies to move more or less during acquisition (i.e. children vs adults). However, tools are often still tuned and evaluated on small homogeneous cohorts of data which may not adequately reflect the performance of the tool using these same settings on another dataset. In the case of datasets such as ABIDE [30], which is commonly regarded as noisy, a tool developed or tuned to provide highly sensitive results on the Kirby21 [31] dataset may provide an elevated variance not reflective of the cohorts represented in the ABIDE data, rather the processing choices (see [32] in contrast with [33]). Additionally, if pipelines are tuned on the data from which they will be deriving claims, the results may end up overfit and tread dangerously close to “p-hacking.”

Size	Homogenous	Example Dataset
Small	No	Subset of CoRR [28]
Small	Yes	Kirby21 [31]

Large	No	ABIDE [30]
Large	Yes	HCP [34]

Table 1: Examples of datasets based on size and type categorizations.

Given the response curves created for each tool in Chapter 2, and four coarse dataset types (summarized in Table 1), we have the opportunity to develop a justified a priori approach for hyperparameter selection. This meaningfully informs tool selection and experimental design, such that processing choice is no longer a “black box” process.

To accomplish this, I plan to use the four exemplar datasets listed in Table 1 and process them using the characterized pipelines in Chapter 2. In each case, I will choose several parameter settings over the sensitivity/generalizability curve for each tool, and evaluate the quality of claims made from the derived data (i.e. p-value/power, effect size, relative-variance). Importantly, the curves used to identify operating points will be developed on a different dataset than that being tested, to avoid overfitting of this technique. I expect that this work will suggest that particular optimizations are suitable in different contexts. For example, small homogeneous datasets may be better suited for more highly generalizable operating points, relative to larger datasets. The datasets chosen in Table 1 are preliminary and each exemplify qualities that make them unique from one another, and fit well within their categorizations.

Leveraging the same skills demonstrated in Chapters 1 and 2, I believe that I am qualified to undertake this project, and that it is an intriguing and powerful area of exploration in neuroimaging. The question posed in this chapter is also of relatively low risk, as successful results would provide a benchmark upon which these features of generalizability and sensitivity can be further explored, and null results suggest the need for new models to characterize the desiderata in question. This leaves room for expansion and extension, such as attempting to incorporate quality control metrics of the derivatives themselves, or tackle this problem for specific domains independently.

Conclusion

Throughout my Ph.D. I will undertake challenges in developing computational infrastructure and models for balancing generalizability and sensitivity in neuroimage analysis, and propose a new method for pipeline and parameter selection in neuroimaging. As an advocate for open and sustainable science, all of the tools and models I develop throughout my degree will be publicly available and made as general-purpose as possible, so that they may be used and extended by others both in neuroimaging and other disciplines of science. I believe that the successful completion of my proposed project has potential to increase the quality of neuroimaging research, allow me to develop important skills in statistical modelling and medical image analysis, and serve as a springboard for many related areas of exploration which I can continue to pursue throughout my academic career.

References

[1] York. Astronomical J. 2000 Sep; 120(3):1579.
[2] Sudlow. PLoS med. 2015 Mar; 12(3):e1001779.
[3] Van Essen. Neuroimage 80. 2013; 62-79.
[4] Zuo. Scientific data 1 2014; 140049.
[5] Baker. Nature News. 2016; 533(7604):452.
[6] Glatard. Zenodo (2017); 10.5281/zenodo.877168.
[7] Glatard. GigaScience (2018).
[8] Gorgolewski. Scientific Data 3 (2016): 160044.
[9] Gorgolewski. PLoS CB 13, no. 3 (2017): e1005209.
[10] Kiar. Zenodo. (2016);10.5281/zenodo.60206.
[11] Kiar. Zenodo. (2018); 10.5281/zenodo.1205654.
[12] Association for Computing Machinery. 2016.
[13] Goodman. Sci. Transl. Med. (2016): 8:341ps12.
[14] Plesser. Front. in Neuroinf. 11-76 (2018): 1662-5196.
[15] Crepinsek. Applied Soft Computing 19 (2014): 161-170.
[16] Sherif. Front. in neuroinf. 8 (2014): 54.
[17] Jeanson, Francis. Frontiers event abstract (2014).
[18] Poldrack. NeuroImage 144 (2017): 259-261.
[19] Rex. Neuroimage 19, no. 3 (2003): 1033-1048.
[20] Smith. Neuroimage 23 (2004): S208-S219.
[21] Fischl. Neuroim. 2012 Aug 15;62(2):774-81.
[22] Cox. Comp. and Biomed. res 29, no. 3 (1996): 162-173.
[23] Eickhoff. Neuroimage 25, no. 4 (2005): 1325-1335.
[24] Tournier. Int. J. of Imag Sys.Tech. 22, 1 (2012): 53-66.
[25] Eklund. PNAS (2016): 201602413.
[26] Bowring. bioRxiv (2018): 285585.
[27] Strother. NeuroImage 15, no. 4 (2002): 747-771.
[28] Zuo. Scientific data 1 (2014): 140049.
[29] Kiar. bioRxiv (2018): 188706.
[30] Di Martino.Molecular psych 19, no. 6 (2014): 659.
[31] Landman. NeuroImage. (2010) NIHMS/PMC:252138
[32] Khundrakpam. Cereb. Cortex 27, 3 (2017): 1721-1731.
[33] Haar. Cerebral Cortex 26, no. 4 (2014): 1440-1452.
[34] Van Essen. Neuroimage 80 (2013): 62-79.

Glossary

ACM [12] Repeatability:

"Same team, running the same experimental setup, gets the same results"

ACM [12] Replicability:

"Different team, running the same experimental setup, gets the same results"

ACM [12] Reproducibility:

"Different team, running a similar experimental setup, gets the same results"

Goodman et al. [13] Results Reproducibility:

"Obtain the same results from an independent study with procedures as closely matched to the original study as possible."

Goodman et al. [13] Inferential Reproducibility:

"Draw the same conclusions from either an independent replication of a study or a reanalysis of the original study."