The Journal of Open Source Software

BIDSonym: a BIDS App for the pseudo-anonymization of neuroimaging datasets

Peer Herholz[1], **Rita Marie Ludwig**[2], **and Jean-Baptiste Poline**[1]

1 NeuroDataScience-Origami lab, McConnell Brain Imaging Centre, The Neuro (Montreal Neurological Institute-Hospital), Faculty of Medicine, McGill University, Montreal, Quebec, Canada **2** University of Oregon, Eugene, United States

DOI: 10.21105/joss.03169

Software

- Review
- Repository
- Archive

Editor: Olivia Guest

Reviewers:

- @deep-introspection
- @chrisgorgo
- @neuromusic

Submitted: 11 March 2021
Published: 07 June 2021

Statement of Need

Due to the evolution of research incentives, technicaladvancements and the development of new standards (Eickhoff et al., 2016; Gorgolewski et al., 2016; Nichols et al., 2017; Poldrack et al., 2013; Poldrack & Gorgolewski, 2017, 2014), increasingly greater amounts of neuroimaging data are being shared either publicly or made available through data user agreements. These datasets originate from small samples of participants collected by individual research groups, as well as from "Big Data" samples including thousands of participants collected by large research consortia (UK Biobank (Sudlow et al., 2015), HCP (Van Essen et al., 2013), ABIDE (Di Martino et al., 2014), ADNI (Mueller et al., 2005), etc.). While data sharing is important and beneficial (Eickhoff et al., 2016; Nichols et al., 2017; Poldrack & Gorgolewski, 2014; Poline et al., 2012), privacy of participant data must be protected (Bannier et al., 2020; Brakewood & Poldrack, 2013). To that end, Ethic Review Boards and data sharing platforms typically require that uploaded datasets are provided in anonymized or pseudo-anonymized form, limiting participant reidentification. However, the (pseudo-) anonymization process is deceptively complex; attempts at ensuring data privacy must take into consideration all dataset components, including imaging modalities, as well as national legal and ethical frameworks. Several algorithms have been developed to (pseudo-) anonymize imaging datasets but they offer limited solutions. Some are attached to specific software, some are limited to specific computing environments; most miss an in-depth assessment and treatment of the metadata attached to the dataset, or lack the capacity to automatize (pseudo-) anonymization across large datasets. BIDSonym was created to address these points in one simple, flexible, and general tool that offers users an array of automated (pseudo-)anonymization options to augment participant privacy in neuroimaging datasets. There are two components of neuroimaging datasets that arguably pose the largest risk to maintaining participant privacy: the structural images and accompanying metadata (e.g. metadata text files or information embedded in image file headers). Structural images contain visible identifiable participant information via facial features like the eyes, nose, and mouth, and privacy is usually addressed through a process called "defacing" within which all or a subset of these features are removed from the final structural data files. The metadata text files may additionally contain identifiable participant data through the recording of acquisition time and location, and personal details such as date of birth, height, and weight. Here, privacy is maintained by removing or blurring this information from the final dataset. BIDSonym addresses both vulnerabilities in neuroimaging datasets, obviating the need for multiple steps within a data sharing pipeline to ensure participant privacy.

Summary

In concordance with the BIDS-App template (Gorgolewski et al., 2017), BIDSonym operates as a command line tool written in Python (Rossum, 1995) and is intended to run in its containerized version (either using Docker (https://www.docker.com) or Singularity (https://sylabs.io),

providing all necessary software dependencies. However, it is also available as a Python package via PyPi (https://pypi.org) to facilitate reuse in a development environment. BIDSonym expects BIDS datasets (Gorgolewski et al., 2016) and provides three core functionalities as depicted in Figure 1: defacing of structural (i.e. T1 and T2 weighted images, adaptation of potentially sensitive metadata information, and evaluation of (pseudo-) anonymization results.

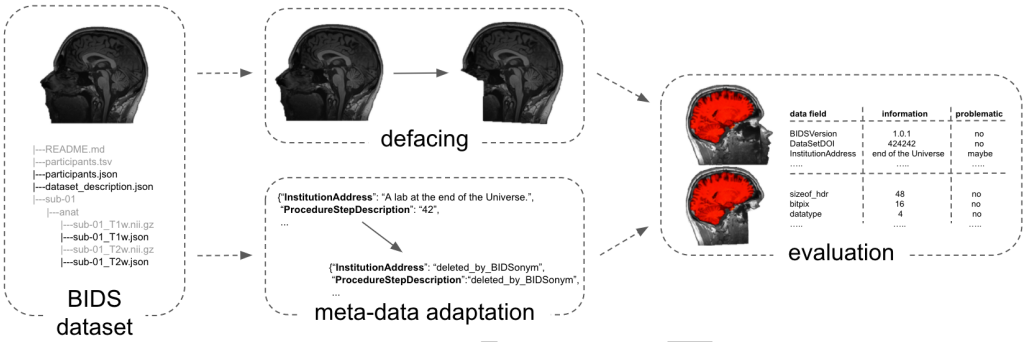

Figure 1: Overview of BIDSonym's functionality | Providing a dataset in BIDS as input, structural images are defaced, meta-data fields adapted as requested and the performance of the defacing, as well as all meta-data fields (in both the json sidecar files and image headers) evaluated.

Focusing on the first aspect, BIDSonym supports a multitude of commonly used defacing algorithms and tools, including pydeface (Gulban et al., 2019), mri_deface (Bischoff-Grethe et al., 2007), quickshear (Schimke & Hale, 2011), and mridefacer (Hanke & Halchenko, 2018). Based on the chosen tool, facial features of the structural images of either specified participants or the whole data set are then removed. Figure 2 provides an example for the different algorithms and tools available for defacing a structural T1 weighted image. Furthermore, structural images from other modalities (e.g. T2 weighted) can also be defaced in which case the defaced T1 weighted structural image will be utilized as a deface mask. In order to account for possible errors during the defacing process, BIDSonym moves the non-defaced original structural images into a distinct directory before defacing, allowing users to test multiple defacing settings and/or options. Following BIDS (Gorgolewski et al., 2016), those files are moved to a different directory ('/sourcedata') and a description identifier ('*_desc-nondeid*') is added to the filenames.

A comparable behavior is implemented with regard to metadata information in BIDSonym's second core functionality. The information in the metadata files that accompany neuroimaging data and in the headers of neuroimaging data files will be gathered and listed within a tabular file ('*.tsv'). If specified, the extracted information will then be queried for potentially sensitive information (e.g. name, date or place of birth, etc.) and marked accordingly. Additionally, users can specify that certain information should be deleted from metadata files, in which case they will be moved and renamed as described for the neuroimaging data.

BIDSonym's third core functionality implements quality control assessments of the defacing results and of the information present in the data. Concerning the first, this includes (interactive) plots that allow to evaluate if the applied algorithm and settings were too stringent (e.g. removing voxels belonging to the brain). Regarding the second, information present in the meta-data files and image headers are gathered in tabular format within respective files ('*.tsv'). Each table contains all key-value pairs present in a given file and aims to provide an assessment of potentially sensitive information.

The Journal of Open Source Software

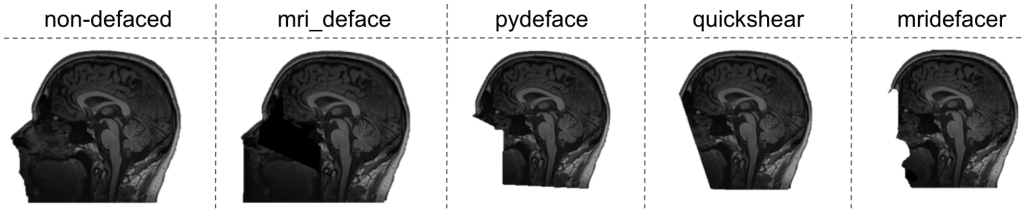

Figure 2: Defacing examples | Results of the different algorithms and tools (columns) included in BIDSonym, displayed in comparison to the corresponding original structural image (most left).

More information on BIDSonym's workflow, the corresponding processing steps and outcomes, as well as installation instructions can be found in the respective documentation (https://peerherholz.github.io/BIDSonym) and GitHub repository (https://github.com/PeerHerholz/BIDSonym). BIDSonym provides a straightforward and flexible way to pseudo-anonymize neuroimaging datasets by a variety of means, operating on both small and large datasets through its implementation following the BIDS-App template (Gorgolewski et al., 2017). BIDSonym depends on the nibabel (Brett et al., 2020), nipype (Gorgolewski et al., 2011), nilearn (Abraham et al., 2014), pybids (Yarkoni et al., 2019) and pandas (McKinney, 2010) python packages (all are well maintained and tested) and is licensed under the BSD-3 license (https://opensource.org/licenses/BSD-3-Clause). As data sharing becomes more widely adopted, BIDSonym fills an important gap for the neuroimaging community.

Acknowledgements

P.H. and J.B.P. were supported in parts by funding from the Canada First Research Excellence Fund, awarded to McGill University for the Healthy Brains for Healthy Lives initiative, the National Institutes of Health (NIH) NIH-NIBIB P41 EB019936 (ReproNim), as well as the National Institute of Mental Health of the NIH under Award Number R01MH096906. P.H. was additionally supported by research scholar award from Brain Canada, in partnership with Health Canada, for the Canadian Open Neuroscience Platform initiative. This project originated as part of Neurohackademy which is funded by the National Institute of Mental Health through a grant to Ariel Rokem and Tal Yarkoni (R25MH112480). Finally, all the contributors listed in the project's Zenodo and GitHub repository have contributed code and intellectual labor to further improve BIDSonym. The same holds true for users that reported issues and continue to do so.

References

Abraham, A., Pedregosa, F., Eickenberg, M., Gervais, P., Mueller, A., Kossaifi, J., Gramfort, A., Thirion, B., & Varoquaux, G. (2014). Machine learning for neuroimaging with scikit-learn. *Frontiers in Neuroinformatics*, *8*. https://doi.org/10.3389/fninf.2014.00014

Bannier, E., Barker, G., Borghesani, V., Broeckx, N., Clement, P., Vaya, M. de la I., Emblem, K. E., Ghosh, S., Glerean, E., Gorgolewski, K. J., Havu, M., Halchenko, Y. O., Herholz, P., Hespel, A., Heunis, S., Hu, Y., Chuan-Peng, H., Huijser, D., Jancalek, R., … Zhu, H. (2020). *The Open Brain Consent: Informing research participants and obtaining consent to share brain imaging data*. PsyArXiv. https://doi.org/10.31234/osf.io/f6mnp

Bischoff-Grethe, A., Ozyurt, I. B., Busa, E., Quinn, B. T., Fennema-Notestine, C., Clark, C. P., Morris, S., Bondi, M. W., Jernigan, T. L., Dale, A. M., Brown, G. G., & Fischl, B. (2007). A Technique for the Deidentification of Structural Brain MR Images. *Human Brain Mapping*, *28*(9), 892–903. https://doi.org/10.1002/hbm.20312

Brakewood, B., & Poldrack, R. A. (2013). The ethics of secondary data analysis: Considering the application of Belmont principles to the sharing of neuroimaging data. *NeuroImage*, *82*, 671–676. https://doi.org/10.1016/j.neuroimage.2013.02.040

Brett, M., Markiewicz, C. J., Hanke, M., Côté, M.-A., Cipollini, B., McCarthy, P., Jarecka, D., Cheng, C. P., Halchenko, Y. O., Cottaar, M., Ghosh, S., Larson, E., Wassermann, D., Gerhard, S., Lee, G. R., Wang, H.-T., Kastman, E., Kaczmarzyk, J., Guidotti, R., … freec84. (2020). *Nibabel*. Zenodo. https://doi.org/10.5281/zenodo.3757992

Di Martino, A., Yan, C.-G., Li, Q., Denio, E., Castellanos, F. X., Alaerts, K., Anderson, J. S., Assaf, M., Bookheimer, S. Y., Dapretto, M., Deen, B., Delmonte, S., Dinstein, I., Ertl-Wagner, B., Fair, D. A., Gallagher, L., Kennedy, D. P., Keown, C. L., Keysers, C., … Milham, M. P. (2014). The autism brain imaging data exchange: Towards a large-scale evaluation of the intrinsic brain architecture in autism. *Molecular Psychiatry*, *19*(6), 659–667. https://doi.org/10.1038/mp.2013.78

Eickhoff, S., Nichols, T. E., Van Horn, J. D., & Turner, J. A. (2016). Sharing the wealth: Neuroimaging data repositories. *NeuroImage*, *124*(Pt B), 1065–1068. https://doi.org/10.1016/j.neuroimage.2015.10.079

Gorgolewski, K. J., Alfaro-Almagro, F., Auer, T., Bellec, P., Capotă, M., Chakravarty, M. M., Churchill, N. W., Cohen, A. L., Craddock, R. C., Devenyi, G. A., Eklund, A., Esteban, O., Flandin, G., Ghosh, S. S., Guntupalli, J. S., Jenkinson, M., Keshavan, A., Kiar, G., Liem, F., … Poldrack, R. A. (2017). BIDS apps: Improving ease of use, accessibility, and reproducibility of neuroimaging data analysis methods. *PLOS Computational Biology*, *13*(3), e1005209. https://doi.org/10.1371/journal.pcbi.1005209

Gorgolewski, K. J., Auer, T., Calhoun, V. D., Craddock, R. C., Das, S., Duff, E. P., Flandin, G., Ghosh, S. S., Glatard, T., Halchenko, Y. O., Handwerker, D. A., Hanke, M., Keator, D., Li, X., Michael, Z., Maumet, C., Nichols, B. N., Nichols, T. E., Pellman, J., … Poldrack, R. A. (2016). The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. *Scientific Data*, *3*(1), 1–9. https://doi.org/10.1038/sdata.2016.44

Gorgolewski, K. J., 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*. https://doi.org/10.3389/fninf.2011.00013

Gulban, O. F., Nielson, D., Poldrack, R., Lee, J., Gorgolewski, K. J., Sochat, V., & Ghosh, S. (2019). *Pydeface*. http://doi.org/10.5281/zenodo.3524401

Hanke, M., & Halchenko, Y. (2018). *Mridefacer*. https://github.com/mih/mridefacer

McKinney, W. (2010). Data structures for statistical computing in python. *Proceedings of the 9th Python in Science Conference*, *445*, 51–56.

Mueller, S. G., Weiner, M. W., Thal, L. J., Petersen, R. C., Jack, C., Jagust, W., Trojanowski, J. Q., Toga, A. W., & Beckett, L. (2005). The Alzheimer's Disease Neuroimaging Initiative. *Neuroimaging Clinics of North America*, *15*(4), 869–877. https://doi.org/10.1016/j.nic.2005.09.008

Nichols, T. E., Das, S., Eickhoff, S. B., Evans, A. C., Glatard, T., Hanke, M., Kriegeskorte, N., Milham, M. P., Poldrack, R. A., Poline, J.-B., Proal, E., Thirion, B., Van Essen, D. C., White, T., & Yeo, B. T. T. (2017). Best practices in data analysis and sharing in neuroimaging using MRI. *Nature Neuroscience*, *20*(3), 299–303. https://doi.org/10.1038/nn.4500

Poldrack, R. A., Barch, D. M., Mitchell, J., Wager, T., Wagner, A. D., Devlin, J. T., Cumba, C., Koyejo, O., & Milham, M. (2013). Toward open sharing of task-based fMRI data: The

OpenfMRI project. *Frontiers in Neuroinformatics*, *7*. https://doi.org/10.3389/fninf.2013.00012

Poldrack, R. A., & Gorgolewski, K. J. (2017). OpenfMRI: Open sharing of task fMRI data. *NeuroImage*, *144*, 259–261. https://doi.org/10.1016/j.neuroimage.2015.05.073

Poldrack, R. A., & Gorgolewski, K. J. (2014). Making big data open: Data sharing in neuroimaging. *Nature Neuroscience*, *17*(11), 1510–1517. https://doi.org/10.1038/nn.3818

Poline, J.-B., Breeze, J. L., Ghosh, S. S., Gorgolewski, K. J., Halchenko, Y. O., Hanke, M., Helmer, K. G., Marcus, D. S., Poldrack, R. A., Schwartz, Y., Ashburner, J., & Kennedy, D. N. (2012). Data sharing in neuroimaging research. *Frontiers in Neuroinformatics*, *6*. https://doi.org/10.3389/fninf.2012.00009

Rossum, G. (1995). *Python reference manual* [Technical Report]. CWI (Centre for Mathematics; Computer Science).

Schimke, N., & Hale, J. (2011). Quickshear defacing for neuroimages. *Proceedings of the 2nd USENIX Conference on Health Security and Privacy*, 11.

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., Liu, B., Matthews, P., Ong, G., Pell, J., Silman, A., Young, A., Sprosen, T., Peakman, T., & Collins, R. (2015). UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLoS Medicine*, *12*(3). https://doi.org/10.1371/journal.pmed.1001779

Van Essen, D. C., Smith, S. M., Barch, D. M., Behrens, T. E. J., Yacoub, E., & Ugurbil, K. (2013). The WU-Minn Human Connectome Project: An overview. *NeuroImage*, *80*, 62–79. https://doi.org/10.1016/j.neuroimage.2013.05.041

Yarkoni, T., Markiewicz, C. J., Vega, A. de la, Gorgolewski, K. J., Salo, T., Halchenko, Y. O., McNamara, Q., DeStasio, K., Poline, J.-B., Petrov, D., Hayot-Sasson, V., Nielson, D. M., Carlin, J., Kiar, G., Whitaker, K., DuPre, E., Wagner, A., Tirrell, L. S., Jas, M., … Blair, R. (2019). PyBIDS: Python tools for BIDS datasets. *Journal of Open Source Software*, *4*(40), 1294. https://doi.org/10.21105/joss.01294