Protocol Article

Protocol for a machine learning algorithm predicting depressive disorders using the T1w/T2w ratio

David A.A. Baranger [a],*, Yaroslav O. Halchenko [b], Skye Satz [a], Rachel Ragozzino [a], Satish Iyengar [c], Holly A. Swartz [a], Anna Manelis [a],*

[a] Department of Psychiatry, Western Psychiatric Institute and Clinic, University of Pittsburgh Medical Center, University of Pittsburgh, Pittsburgh, PA, USA
[b] Department of Psychological and Brain Sciences, Dartmouth College, NH, USA
[c] Department of Statistics, University of Pittsburgh, Pittsburgh, PA, USA

A B S T R A C T

The T1w/T2w ratio is a novel magnetic resonance imaging (MRI) measure that is thought to be sensitive to cortical myelin. Using this novel measure requires developing novel pipelines for the data quality assurance, data analysis, and validation of the findings in order to apply the T1w/T2w ratio for classification of disorders associated with the changes in the myelin levels. In this article, we provide a detailed description of such a pipeline as well as the reference to the scripts used in our recent report that applied the T1w/T2w ratio and machine learning to classify individuals with depressive disorders from healthy controls.

© 2021 The Authors. Published by Elsevier B.V.
This is an open access article under the CC BY-NC-ND license
(http://creativecommons.org/licenses/by-nc-nd/4.0/)

Specifications table

Subject Area:	Neuroscience
More specific subject area:	Neuroimaging and Psychiatry
Protocol name:	Cortical myelin mapping and machine learning-based prediction
Reagents/tools:	Scanning was performed using a 3T Siemens Prisma scanner with a 64-channel receiver head coil.
Experimental design:	Magnetic resonance imaging (MRI) data were collected following the Human Connectome Project (HCP) 3T imaging protocol. High-resolution T1w images were collected using the MPRAGE sequence with TR=2400ms, resolution=0.8 × 0.8 × 0.8mm, 208 slices, FOV=256, TE=2.22ms, flip angle=8°. High-resolution T2w images were collected using TR=3200ms, resolution=0.8 × 0.8 × 0.8mm, 208 slices, FOV=256, TE=563ms. Field maps were collected in the AP and PA directions using the spin echo sequence (TR=8000, resolution=2 × 2 × 2mm, FOV=210, TE=66ms, flip angle=90°, 72 slices).
Trial registration:	
Ethics:	The study was approved by the University of Pittsburgh Institutional Review Board and participants gave written informed consent (protocol number STUDY20060265)
Value of the Protocol:	Acquisition, preprocessing, and quality assurance of T1w/T2w cortical myelin maps.
	Development of a machine learning analysis pipeline to predict depressive disorders with performance evaluation via permutation.

Methods

Participants

Participant recruitment and demographics has been previously described [1]. In brief, we recruited healthy controls (HC; N=55) and individuals with unipolar depression (UD; N=50) diagnosed with major depressive or persistent depressive disorder using DSM-5 criteria. Participants with a history of head injury, metal in the body, pregnancy, claustrophobia, neurodevelopmental disorders, systemic medical illness, premorbid IQ<85 per the National Adult Reading Test [2], current alcohol/drug abuse, Young Mania Rating Scale scores>10 (YMRS [3]) at scan, or meeting criteria for any psychotic-spectrum disorder were excluded during recruitment. In addition, we excluded from the present analyses participants with brain abnormalities of potential clinical relevance (2 UD); diagnosis change during the course of the study (1 HC was diagnosed with major depressive disorder, and 1 UD was diagnosed with bipolar disorder); scanner and movement related artifacts in T1w or T2w images (4 HC, 7 UD), and myelin maps of insufficient quality (3 HC, 1 UD). A more detailed description of the last two points will be provided below. The final sample included 47 HC and 39 UD.

Neuroimaging data acquisition

The neuroimaging data were collected at the University of Pittsburgh Magnetic Resonance Research Center using a 3T Siemens Prisma scanner with a 64-channel receiver head coil. Neuroimaging data files were named according to the ReproIn convention [4]. DICOM images were converted to BIDS format using *heudiconv* [5] and *dcm2niix* [6]. High-resolution T1w images were collected using the MPRAGE sequence (TR=2400ms, resolution=0.8 × 0.8 × 0.8mm, 208 slices, FOV=256, TE=2.22ms, flip angle=8°). High-resolution T2w images were collected using TR=3200ms, resolution=0.8 × 0.8 × 0.8mm, 208 slices, FOV=256, TE=563ms. Field maps were collected in the AP and PA directions using the spin echo sequence (TR=8000, resolution=2 × 2 × 2mm, FOV=210, TE=66 ms, flip angle=90°, 72 slices). The raw data used [1] are publicly available (https://openneuro.org/datasets/ds003653/versions/1.0.0) [7].

Subject-level preprocessing

Cortical myelin was characterized with the T1w/T2w ratio [8–10] using the PreFreeSurfer, FreeSurfer, and PostFreeSurfer minimal preprocessing pipelines for the human connectome project (HCP) [8]. *Workbench v1.4.2* and *HCPpipelines-4.1.3* were installed system-wide on a workstation with GNU/Linux Debian 10 operating system. Bias field correction in *PreFreeSurfer* used spin echo field maps collected in AP and PA phase encoding directions. MSMSulc [11] in *PostFreeSurfer* was used for registration to standard space. Parcellation of the T1w/T2w ratio maps [10,12] was done using the Glasser Atlas [9] (n=360 regions). Parcellation is a procedure by which points in the brain are mapped to brain regions, or parcels. The mean T1w/T2w ratio of each parcel was computed for each participant. The Glasser Atlas was used as it is the only brain atlas which incorporates the T1w/T2w ratio. Subject-level myelin maps, as well as all code and data used in this protocol, are publicly available (https://github.com/manelis-lab/myelin-paper-NICL2021).

Neuroimaging quality assurance

Quality assurance (QA) was first performed following standard procedures. We additionally provide detail on our customizations to our QA pipelines for T1w, T2w, and T1w/T2w ratio images. T1w and T2w images were visually inspected for noise or movement-induced artifacts [13], including ringing, aliasing, ghosting, blurring, banding, and hyperintensities, as well as for structural abnormalities that may be clinically relevant. Data quality was then examined using *mriqc version 0.15.1* [14]. Background noise enhanced images generated by m*riqc* were visually inspected for the same aforementioned artifacts. As there are no clear guidelines regarding how to use single or combined image quality metrics (IQMs) produced by *mriqc* to identify potential low-quality scans or unacceptable level of noise, we adopted the procedure proposed by the *mriqception* project [15] to identify potential T1w and T2w outlier images. Specifically, the images that passed this initial visual inspection were compared against a large sample of independent deidentified T1w and T2w images whose *mriqc* IQMs were downloaded from the *mriqc* API (https://mriqc.nimh.nih.gov/). The mriqc API collection contains information on over 1.4 million T1 and T2 images. For the purpose of this study, we selected the IQMs from images that were matched to the present study, based on TR, TE, spacing, scanner manufacturer, head coil, and the version of *mriqc* used. This resulted in 1046 T1w comparison images and 619 T2w comparison images. The distribution of IQMs from the present study was compared to the distribution of IQMs from the API. Scans with IQMs beyond the interquartile range (IQR) of the *mriqc* API data (median +/- 1.5 × 75% quartile – 25% quartile) were flagged as potential outliers and were re-inspected. Scans were excluded solely based on IQMs when values fell far outside the expected range (>5 IQRs). One subject (UD) was excluded based on this procedure. A total of 4 HC subjects and 7 UD subjects were excluded for noise or movement-related artifacts. Participants were excluded from analyses even if they failed QC only for T1w or only for T2w images. For the images that passed the *mriqc* API quality assurance, the FreeSurfer generated images were visually inspected for registration errors and artifacts that were not detected previously.

T1w/T2w ratio quality assurance

For the participants whose T1w and T2w images passed the neuroimaging QA steps described above, we evaluated the quality of the T1w/T2w ratio maps. As the T1w/T2w ratio image is a relatively recent addition to the MRI toolkit [10], there is no standard QA pipeline for T1w/T2w ratio maps. Here we detail our procedure for performing QA on T1w/T2w ratio images. First, T1w/T2w ratio images were visually inspected, by comparing each image to the average T1w/T2w ratio map from the HCP [9]. This step allowed us to identify regions with excessively high or low values. Some examples include large regions of apparent low myelin in the occipital cortex due to the transverse sinus interfering with accurate identification of the pial surface, or regions of excessive myelin in the medial frontal cortex that likely appeared due to motion.

Second, beyond gross imaging or anatomical artifacts, some areas in the brain are affected by susceptibility artifacts more than others [16]. These areas include the orbitofrontal cortex
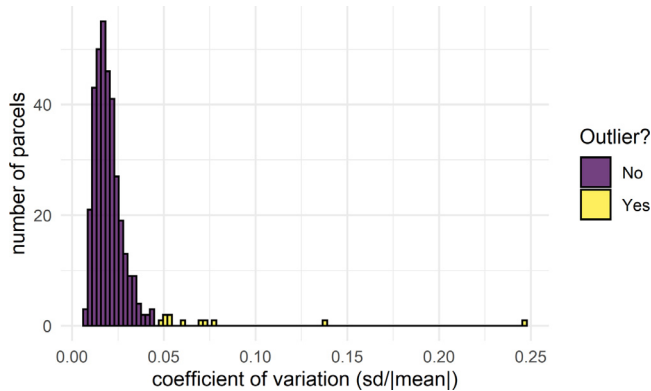
Fig. 1. Results of parcel outlier detection
The coefficient of variation of each parcel (sd/|mean|), colored by whether or not the parcel was determined to be an outlier. sd and |mean| were calculated across all subjects per each parcel.

and the regions in the medial temporal cortex such as hippocampal and entorhinal cortices [13,17]. The presence of susceptibility artifacts varies among subjects due to the differences in the scalp, face, and brain anatomy as well as the placement inside the scanner. Characterizing between-subject variability in the T1w/T2w ratio measures could help identify brain regions affected by susceptibility artifacts the most. We developed an empirical procedure to identify the regions with unusually high variability and remove them from the subsequent analyses. The procedure included calculating the coefficient of variation (sd/mean) to summarize the variability within each parcel [18] (https://github.com/manelis-lab/myelin-paper-NICL2021/blob/master/scripts/ preprocessing/outlier_regions.R). As susceptibility artifacts result in both increased variability in the T1w/T2w ratio and excessively low estimates of the mean T1w/T2w ratio, regions contaminated with susceptibility artifacts will have a very high coefficient of variation, relative to other regions. Rosner's test for outliers [18,19] was used to identify outlier parcels with excessively high variation. Rosner's test assumes that the data, without outliers, are normally distributed. The most extreme values are iteratively tested, identifying up to 10 possible outliers preventing the data from being normal are identified. It may be appropriate to log-transform the coefficient of variation prior to Rosner's test, if the distribution is heavily skewed. This was not done in the present analysis as we were only interested in outliers in the right-hand side of the distribution. If more than 10 possible outliers are suspected, the test needs to be run multiple times. This procedure identified 11 outlier parcels (Fig. 1), including the bilateral hippocampus, entorhinal cortex, presubiculum, piriform cortex, and posterior orbitofrontal cortex complex, and the right subgenual cingulate (bilateral H, EC, PreS, Pir, pOFC, and right 25). The locations for these outlier parcels were, in general, consistent with areas of high susceptibility artifacts [13,17]. These outlier parcels were removed from the subsequent data analyses, thus, leaving 349 parcels per participant in the data set.

Machine learning pipeline

Elastic-net with nested cross-validation

One limitation of using machine learning in neuroimaging data analysis is that there is a larger number of features (e.g., voxels or parcels) relative to the number of participants in the data sets. Elastic net is a regularized regression which combines LASSO and ridge regression (i.e., L1- and L2- norm regularization) [20], which is widely used to select variables in large multivariate analyses [21–23]. Ridge regression reduces the size of overly-large coefficients, while LASSO regression removes variables with small coefficients. In this study, elastic net was used to select parcels whose T1w/T2w ratio was sensitive to distinguishing HC from individuals with depressive disorders. All

variables used in training the elastic net model are publicly available (https://github.com/manelis-lab/myelin-paper-NICL2021). Previous studies that used elastic net for clinical and neuroimaging data analyses often used a whole sample to identify a sparse model. Such approach can result in model bias and overfitting. To mitigate these effects, we used nested cross-validation - a procedure in which a subset of participants is held-out and the entirety of the analytic pipeline is performed in the remaining participants [24]. The resulting model is tested in the held-out participants. This process is repeated multiple times and results in multiple sparse models. Here we provide details on our customizations to the machine learning pipeline.

In the present nested cross-validation analysis, in each iteration of the 'outer loop', two participants (one UD and one HC) were held-out for testing, while the remaining participants were used for model training. During model training, the elastic net alpha (α) parameter was set to 0.5. The optimal lambda (λ) parameter was determined using leave-one-out cross-validation of the training data (i.e., the 'inner loop'). Thus, there were 1833 nested cross-validation models (all possible pairings of UD and HC participants: 39 UD * 47 HC = 1833) and, consequently, 1833 sets of parcels deemed important for UD vs. HC classification (Fig. 2). This analysis allowed us to identify how consistently each parcel was selected across 1833 models. For each parcel, we calculated the proportion of times a given region was selected across all 1833 nested cross-validation models. The less frequently a region is selected, the more strongly its selection depends on the choice of the training and testing samples, indicating that it is a less robust predictor of participant class (e.g., less likely to be generalizable).

We used a linear discriminant analysis (LDA) in the 'outer loop' to verify how well the set of variables selected by elastic net (in the 'inner loop') classified UD vs. HC (Fig. 2). A total of 1833 LDA models were tested as this analysis was run for each repetition of nested cross validation. Specifically, each LDA model was trained using the set of participants and the set of parcels selected by elastic net in that particular 'inner loop' cross-validation model . This LDA model was then used to predicted the class (UD or HC) of the two participants who were held-out for each round of nested cross-validation. This procedure was repeated for all 1833 participant pairs (https://github.com/manelis-lab/myelin-paper-NICL2021/blob/master/scripts/analyses/glmnet_with_LDA_myelin_paper.R). Participant-wise accuracy was computed as the percent of nested models in which the held-out participant was correctly classified. Total model accuracy was computed as the average of the participant-wise accuracies. Model sensitivity and specificity were computed as the average of UD-only and HC-only participant accuracies, respectively (https://github.com/manelis-lab/myelin-paper-NICL2021/blob/master/scripts/analyses/process_glmnet_output.R). The strength of the nested cross validation approach is that it produces a relatively unbiased estimate of model parameter (λ), which would not be the case if the entire training data were used for determining the parameter [25,26]. We note than many other classifiers could be used instead of LDA. The complete output of this pipeline is available (https://github.com/manelis-lab/myelin-paper-NICL2021).

Testing the model performance with permutation analyses

While we believed that the parcels selected by all 1833 elastic net models would likely represent signal, the parcels that were selected by 1 model could reflect noise. Permutation analyses were employed to assess the presence and extent of model bias and overfitting. In each permutation, participant class (UD or HC) in the training data set was permuted (randomized), while the class of the hold-out participants remained unchanged (i.e., true) (Fig. 2). The expected accuracy for permuted data is at chance (in the present case, 50%). Deviations from this expectation are indicative of bias or overfitting. The permutation analyses included the same steps for nested cross-validation procedure described above, except that for each cross-validation fold we conducted 100 repetitions of a permutation procedure, which resulted in a total of 183,300 permutation models (1833*100=183,300). A new randomization seed was set prior to each of these permutations to ensure that permutations would be unique, and the results would be reproducible. Following the processing pipeline for the primary analyses, we calculated the proportion of times each variable was selected in the permutation analyses, reflecting how frequently we can expect each variable to be selected by chance. Variables from the primary analyses were considered to be important for distinguishing control from
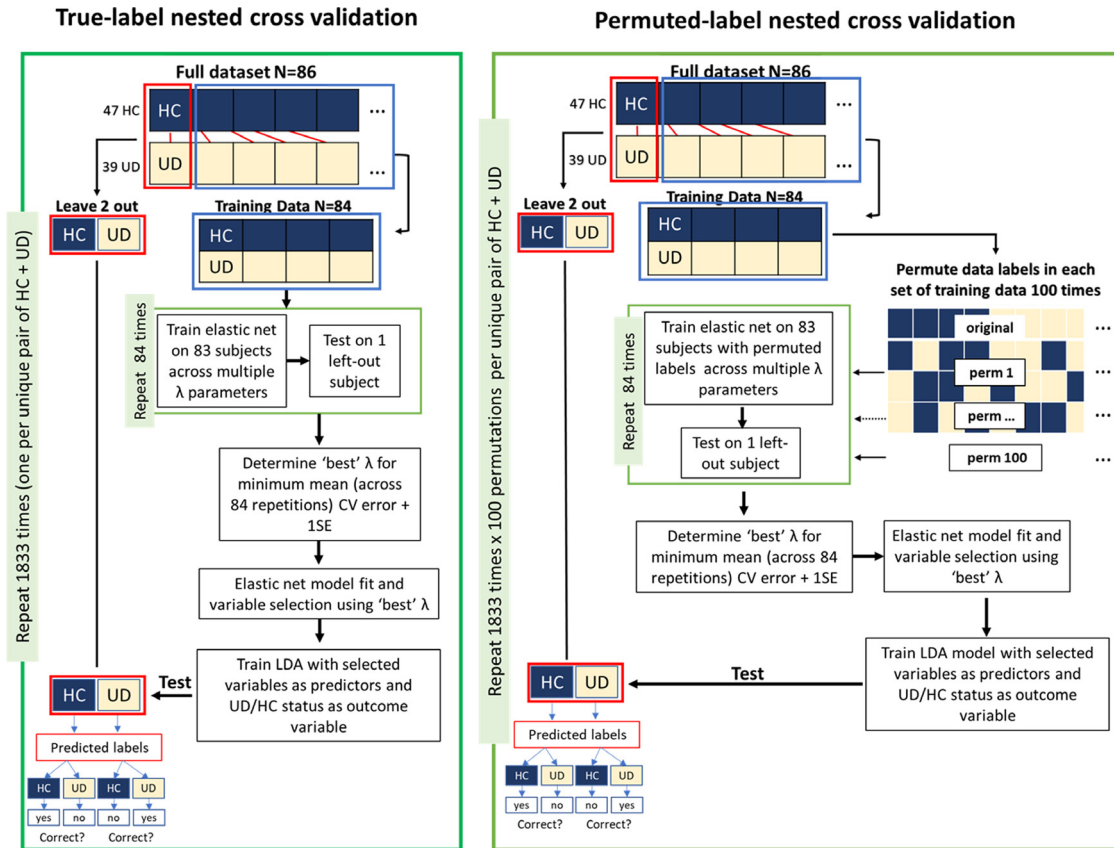
Fig. 2. Diagram of analysis steps.
Figure used with permission from Baranger et al., 2021[1]. Conceptual depiction of analysis steps including: (1) a unique pair of one UD and one HC participant is held-out; (2) an elastic net regression is used to select variables; (3) the retained variables are used an LDA model predicting case/control status; (4) the LDA model is tested on the held-out sample; (5) this process is repeated for each of the $n=1833$ pairs of subjects; (6) for each held-out pair, the training procedure is repeated with 100 unique permutations.

UD if they were selected more than the median value + 3.5 IQR of the selection frequency in permutation analyses. That is, if they were selected more frequently than expected by chance. We applied this threshold, as opposed to computing the significance of each feature via individually permuting them, due to the high computational burden of the latter in the context of nested cross-validation ($1833^*100^*360 = 65{,}988{,}000$ models), In addition, we computed the LDA accuracies for each permutation round as a validation of the variable selection. Accuracy is expected to be higher in the primary analyses than the permutations (i.e., selected variables are truly predictive of participant class). The deviation of permutation accuracy from chance is indicative of the extent to which results are contaminated by bias or overfitting. Deviation from chance would thus indicate that selected variables may not be valid indicators of the outcome, and would additionally indicate that model results may not be generalizable to new samples (https://github.com/manelis-lab/myelin-paper-NICL2021/blob/master/scripts/analyses/permuted_glmnet_with_LDA_myelin_paper.R).

Outcomes

There are three primary outcomes of this protocol. First, we developed a quality assurance pipeline for T1w/T2w ratio images that includes identifying the outlier regions heavily affected by susceptibility artifacts. Second, we developed a nested cross-validation pipeline that runs the feature selection and linear discrimination to distinguish participant classes (UD vs. HC). Third, we verified the selected regions and classification accuracy, using a permutation approach. The raw data are publicly available (https://openneuro.org/datasets/ds003653/versions/1.0.0) [7]. The derivative data as well as the scripts used in this protocol are available (https://github.com/manelis-lab/myelin-paper-NICL2021) [27].

Declaration of Interests

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

D.A.A.B, Y.O.H., S.S., R.R, S.I., and A.M.: declare no conflict of interest.

H.A.S: receives royalties from Wolters Kluwer, royalties and an editorial stipend from APA Press, and honorarium from Novus Medical Education.

References

[1] D.A.A. Baranger, Y.O. Halchenko, S. Satz, et al., Aberrant levels of cortical myelin distinguish individuals with depressive disorders from healthy controls, NeuroImage Clinical 32 (2021) 102790, doi:10.1016/J.NICL.2021.102790.

[2] HE. Nelson, National Adult Reading Test (NART): For the Assessment of Premorbid Intelligence in Patients with Dementia: Test Manual, Nfer-Nelson, 1982.

[3] R.C. Young, J.T. Biggs, V.E. Ziegler, DA. Meyer, A rating scale for mania: Reliability, validity and sensitivity, Br. J. Psych. 133 (1978) 429–435, doi:10.1192/bjp.133.5.429.

[4] M. Visconti di Oleggio Castello, J.E. Dobson, T. Sackett, et al. ReproNim/reproin 0.6.0. January 2020. doi:10.5281/ZENODO.3625000

[5] Y. Halchenko, M. Goncalves, M.V. Castello O. di, et al. nipy/heudiconv: v0.5.4 [0.5.4] - 2019-04-29. April 2019. doi:10.5281/ZENODO.2653784

[6] X. Li, P.S. Morgan, J. Ashburner, J. Smith, C. Rorden, The first step for neuroimaging data analysis: DICOM to NIfTI conversion, J. Neurosci. Methods 264 (2016) 47–56, doi:10.1016/j.jneumeth.2016.03.001.

[7] A. Manelis, D.A.A. Baranger, S. Satz, et al., Cortical myelin measured by the T1w/T2w ratio in individuals with depressive disorders and healthy controls, OpenNeuro (2021), doi:10.18112/openneuro.ds003653.v1.0.0.

[8] M.F. Glasser, S.N. Sotiropoulos, J.A. Wilson, et al., The minimal preprocessing pipelines for the human connectome project, NeuroImage (2013), doi:10.1016/j.neuroimage.2013.04.127.

[9] M.F. Glasser, T.S. Coalson, E.C. Robinson, et al., A multi-modal parcellation of human cerebral cortex, Nature (2016), doi:10.1038/nature18933.

[10] M.F. Glasser, DC. van Essen, Mapping human cortical areas in vivo based on myelin content as revealed by T1- and T2-weighted MRI, J. Neurosci. 31 (32) (2011) 11597–11616, doi:10.1523/JNEUROSCI.2180-11.2011.

[11] E.C. Robinson, K. Garcia, M.F. Glasser, et al., Multimodal surface matching with higher-order smoothness constraints, NeuroImage (2018), doi:10.1016/j.neuroimage.2017.10.037.

[12] R. Shafee, R.L. Buckner, B. Fischl, Gray matter myelination of 1555 human brains using partial volume corrected MRI images, NeuroImage 105 (2015) 473–485, doi:10.1016/j.neuroimage.2014.10.054.

[13] K. Krupa, M. Bekiesińska-Figatowska, Artifacts in Magnetic Resonance Imaging, Polish J. Radiol. 80 (1) (2015) 93, doi:10.12659/PJR.892628.

[14] O. Esteban, D. Birman, M. Schaer, O.O. Koyejo, R.A. Poldrack, KJ. Gorgolewski, MRIQC: advancing the automatic prediction of image quality in MRI from unseen sites, PLoS ONE (2017), doi:10.1371/journal.pone.0184661.

[15] E. Beard MRIQCeption. 2020.

[16] C.A. Olman, L. Davachi, S. Inati, Distortion and signal loss in medial temporal lobe, PLoS ONE 4 (12) (2009), doi:10.1371/journal.pone.0008160.

[17] R. Deichmann, J.A. Gottfried, C. Hutton, R. Turner, Optimized EPI for fMRI studies of the orbitofrontal cortex, NeuroImage 19 (2) (2003) 430–441, doi:10.1016/S1053-8119(03)00073-9.

[18] SP. Millard, EnvStats, an RPackage for Environmental Statistics, Wiley StatsRef: Stat. Ref. Online (2014), doi:10.1002/9781118445112.stat07181.

[19] B. Rosner, On the detection of many outliers, Technometrics (1975), doi:10.1080/00401706.1975.10489305.

[20] H. Zou, T. Hastie, Regularization and variable selection via the elastic net, J. R. Stat. Soc. Ser. B Stat. Methodol. (2005), doi:10.1111/j.1467-9868.2005.00503.x.

[21] A. Manelis, S. Iyengar, H.A. Swartz, ML. Phillips, Prefrontal cortical activation during working memory task anticipation contributes to discrimination between bipolar and unipolar depression, Neuropsychopharmacology (2020), doi:10.1038/s41386-020-0638-7.

[22] H.E. Acuff, A. Versace, M.A. Bertocci, et al., Baseline and follow-up activity and functional connectivity in reward neural circuitries in offspring at risk for bipolar disorder, Neuropsychopharmacology (2019), doi:10.1038/s41386-019-0339-2.

[23] M.A. Bertocci, G. Bebko, A. Versace, et al., Predicting clinical outcome from reward circuitry function and white matter structure in behaviorally and emotionally dysregulated youth, Molecular Psych. (2016), doi:10.1038/mp.2016.5.

[24] G. Varoquaux, P.R. Raamana, D.A. Engemann, A. Hoyos-Idrobo, Y. Schwartz, B. Thirion, Assessing and tuning brain decoders: cross-validation, caveats, and guidelines, NeuroImage 145 (2017) 166–179, doi:10.1016/J.NEUROIMAGE.2016.10.038.

[25] C. Davatzikos, Machine learning in neuroimaging: Progress and challenges, NeuroImage 197 (2019) 652–656, doi:10.1016/j.neuroimage.2018.10.003.

[26] S. Varma, R. Simon, Bias in error estimation when using cross-validation for model selection, BMC Bioinf. 7 (1) (2006) 1–8, doi:10.1186/1471-2105-7-91.

[27] A. Manelis, Y. Halchenko Manelis-lab/myelin-paper-NICL2021: initial public release. August 2021. doi:10.5281/ZENODO.5249850