NeuroImage 265 (2023) 119800

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/neuroimage

Removing the effects of the site in brain imaging machine-learning – Measurement and extendable benchmark

Aleix Solanes[a,b,+], Corentin J Gosling[c,d,m,+], Lydia Fortea[a,e,f], María Ortuño[a], Elisabet Lopez-Soley[a,f,g], Sara Llufriu[a,f,g], Santiago Madero[a,e,f,h], Eloy Martinez-Heras[a,f,g], Edith Pomarol-Clotet[e,i,j], Elisabeth Solana[a,f,g], Eduard Vieta[a,e,f,h], Joaquim Radua[a,e,f,k,l,*]

[a] Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS), Barcelona, Spain
[b] Department of Psychiatry and Forensic Medicine, Autonomous University of Barcelona, Barcelona, Spain
[c] DysCo Lab, Paris Nanterre University, Nanterre, France
[d] Laboratoire de Psychopathologie et Processus de Santé, Université de Paris, Paris, France
[e] Biomedical Network Research Centre on Mental Health (CIBERSAM), Instituto de Salud Carlos III, Madrid, Spain
[f] University of Barcelona, Barcelona, Spain
[g] Center of Neuroimmunology, Laboratory of Advanced Imaging in Neuroimmunological Diseases, Hospital Clinic Barcelona
[h] Barcelona Bipolar Disorders and Depressive Unit, Institute of Neurosciences, Hospital Clinic, Barcelona, Spain
[i] FIDMAG Germanes Hospitalàries Research Foundation, Barcelona, Spain
[j] Benito Menni CASM, Sant Boi de Llobregat, Barcelona, Spain
[k] Department of Psychosis Studies, Institute of Psychiatry, Psychology, and Neuroscience, King's College London, London, United Kingdom
[l] Centre for Psychiatric Research and Education, Department of Clinical Neuroscience, Karolinska Institutet, Stockholm, Sweden
[m] Centre for Innovation in Mental Health (CIMH), School of Psychology, Faculty of Environmental and Life Sciences, University of Southampton, Southampton, UK

ARTICLE INFO

Keywords:
Benchmark
Effects of the site
Machine-learning
Magnetic resonance imaging

ABSTRACT

Multisite machine-learning neuroimaging studies, such as those conducted by the ENIGMA Consortium, need to remove the differences between sites to avoid effects of the site (EoS) that may prevent or fraudulently help the creation of prediction models, leading to impoverished or inflated prediction accuracy. Unfortunately, we have shown earlier that current Methods Aiming to Remove the EoS (MAREoS, e.g., ComBat) cannot remove complex EoS (e.g., including interactions between regions). And complex EoS may bias the accuracy. To overcome this hurdle, groups worldwide are developing novel MAREoS. However, we cannot assess their effectiveness because EoS may either inflate or shrink the accuracy, and MAREoS may both remove the EoS and degrade the data. In this work, we propose a strategy to measure the effectiveness of a MAREoS in removing different types of EoS. FOR MAREOS DEVELOPERS, we provide two multisite MRI datasets with only simple true effects (i.e., detectable by most machine-learning algorithms) and two with only simple EoS (i.e., removable by most MAREoS). First, they should use these datasets to fit machine-learning algorithms after applying the MAREoS. Second, they should use the formulas we provide to calculate the relative accuracy change associated with the MAREoS in each dataset and derive an EoS-removal effectiveness statistic. We also offer similar datasets and formulas for complex true effects and EoS that include first-order interactions. FOR MACHINE-LEARNING RESEARCHERS, we provide an extendable benchmark website to show: a) the types of EoS they should remove for each given machine-learning algorithm and b) the effectiveness of each MAREoS for removing each type of EoS. Relevantly, a MAREoS only able to remove the simple EoS may suffice for simple machine-learning algorithms, whereas more complex algorithms need a MAREoS that can remove more complex EoS. For instance, ComBat removes all simple EoS as needed for predictions based on simple lasso algorithms, but it leaves residual complex EoS that may bias the predictions based on standard support vector machine algorithms.

1. Introduction

Magnetic resonance imaging (MRI) researchers often pool data from different sites to achieve more statistical power to detect true differences (Albajes-Eizagirre et al., 2019). This need for larger sample sizes is also a reality for machine-learning neuroimaging, where small studies may fail to predict or fall into overfitting (Hosseini et al., 2020). However, combining data from different sites is not innocuous. Even if consortiums

* Corresponding author at: IDIBAPS: Institut d'Investigacions Biomediques August Pi i Sunyer, Rosselló 149, 08036 Barcelona, Spain.
+ Joint first authorship

https://doi.org/10.1016/j.neuroimage.2022.119800.
Received 8 December 2021; Received in revised form 17 November 2022; Accepted 5 December 2022
Available online 5 December 2022.

such as ENIGMA use harmonized protocols (Thompson et al., 2014), there are still differences due to varying scanning devices and acquisition sequence parameters. These differences may introduce effects of the site (EoS) that bias the analyses (Solanes et al., 2021).

For instance, imagine we conduct a two-site MRI study to investigate whether we may use baseline MRI to predict the subsequent response to a medication. Imagine also that, by chance, 80% of patients in site A respond to the drug, whereas only 20% in site B. Finally, imagine that site A's MRI device makes the images very bright and site B's device very dark. With these settings, a machine-learning algorithm could predict whether a patient will respond or not, exclusively using the difference in images' brightness between the two MRI devices. In other words, the machine-learning model would predict that patients with bright images will respond, whereas patients with dark images will not. And the machine-learning model would be pretty successful: it would show 80% accuracy! However, this accuracy would be false, inflated, artifactual, exclusively based on an EoS. The balanced accuracy (the average of sensitivity and specificity) separately calculated for each site would be just 50%, like tossing a coin.

Due to the potentially high biases introduced by EoS, researchers worldwide are developing novel Methods Aiming to Remove the EoS (MAREoS). A common and old MAREoS is covarying for the site in the linear model, preferably coded as a random-effects factor (i.e., a mixed-effects analysis) (Favre et al., 2019). Another usual MAREoS is ComBat (Johnson et al., 2007), a batch adjustment method developed for genomics data. Several groups have recently adapted this MAREoS to MRI datasets (Fortin et al., 2018; Radua et al., 2020).

However, we have shown previously that current MAREoS do not entirely remove all differences between sites. Worryingly, these differences may either inflate or shrink the accuracy. In other words, machine-learning algorithms may either use the remaining EoS "fraudulently", thus inflating accuracy rates, or fail to detect true effects due to the noise associated with EoS, thus shrinking accuracy rates (Solanes et al., 2021). While all MAREoS can remove simple additive differences, we are not aware of a MAREoS able to remove complex EoS, such as discrepancies in covariance (i.e., the interaction between brain regions). To avoid reporting biased accuracies, we have provided formulas and an R package to unbiasedly estimate the multisite-corrected accuracy in the presence of residual EoS (Solanes et al., 2021). This package may be helpful to ensure that the EoS do not bias the reported accuracy. However, the goal of the community should be to develop a novel MAREoS able to remove complex EoS entirely.

Unfortunately, MAREoS developers may face a paradox. To our knowledge, there is no straightforward way to measure the EoS-removal effectiveness. For example, in data with EoS and true effects, a MAREoS may yield higher accuracy than another MAREoS for two opposite reasons. It may either reduce the noise associated with EoS (improving the detection of true effects) or fail to remove the EoS (leading to higher accuracy inflation). On the other hand, in data with only EoS and no true effects, a MAREoS may yield a lower accuracy than another MAREoS for two opposite reasons again. It may either remove the EoS better (minimizing the accuracy inflation) or degrade the data more (worsening the detection of true effects).

To overcome this problem, we designed an approach to objectively measure the removal of EoS and the degradation of the data of a given MAREoS. Furthermore, we also provide: a) datasets to conduct these measurements and b) a benchmark website to allow machine-learning researchers readily know the most appropriate MAREoS depending on the situation.

2. Methods

The strategy presented in this paper builds on the study of the change in accuracy associated with a MAREoS. This accuracy change has opposite meanings depending on whether the dataset has only EoS (i.e., no true effects) or only true effects (i.e., no EoS). In datasets with only true effects, an accuracy decrease should only be due to data degradation, a side effect of the MAREoS. Conversely, an accuracy decrease in datasets with only EoS should be due to a correct EoS-removal (plus some potential data degradation). We have noted above that accuracy increases are possible in datasets with both true effects and EoS since the noise associated with EoS may shrink the accuracy (Solanes et al., 2021). However, to simplify the following calculations, we tried to avoid datasets mixing true effects and EoS.

We first describe the datasets and the specific machine-learning algorithm that MAREoS developers should apply to achieve that differences between MAREoS depend only on the MAREoS (while not on the datasets or machine-learning algorithms). Afterward, we present the formulas to measure the effectiveness of a MAREoS. Finally, we show an example with the Johnson-Fortin-Radua version of the ComBat MAREoS (Fortin et al., 2018; Johnson et al., 2007; Radua et al., 2020) (script available at http://enigma.ini.usc.edu/protocols/statistical-protocols/). Readers only interested in the strategy may directly read the section about measuring the EoS-removal effectiveness of a MAREoS.

2.1. Description of the datasets and the machine-learning algorithm

Each simulated dataset includes the baseline MRI data (cortical thickness, cortical surface area, or subcortical volumes) from ~1000 patients from 8 scanner sites, two baseline clinical covariates, and the subsequent responses to a given treatment. The simulated studies would aim to predict the response to the treatment (response vs. no response) from the baseline MRI data (Figure 1). The latter follow normal distributions like those returned by FreeSurfer (Radua et al., 2020) and have linear relationships with two simulated clinical covariates. In this section, we first describe the datasets (along with the machine-learning algorithm) to familiarize developers with them. Afterward, we briefly report how we created the MRI data for interested readers.

Two datasets have only simple EoS (i.e., neither true effects nor complex EoS). The lack of true effects means no relationship between the MRI data and the response. Therefore, machine-learning algorithms should not predict the response. Accuracy should be around 50%, like tossing a coin. However, there are substantial simple differences across sites in response probability and MRI data (e.g., the cortex is systematically measured thicker in some devices). Most machine-learning algorithms may use these simple EoS to "fraudulently" predict the response, inflating the accuracy. These simple EoS should be removable by most MAREoS.

Two other datasets have only simple true effects (i.e., neither EoS nor complex true effects). Thus, there are significant simple relationships between MRI data and the response to treatment (e.g., responders have thicker cortices). Therefore, most machine-learning algorithms should predict the response with >50% accuracy.

To predict the treatment response using the brain imaging data, MAREoS developers should conduct a ten-fold cross-validation using our specific fold distribution. Within each fold, they should fit a lasso algorithm in which the variable to predict is the response to the treatment (coded binarily), and the predictors are the MRI data. For instance, in R, we could use the "glmnet" library (Friedman et al., 2010). Developers may download the specific R scripts from https://www.imardgroup.com/mareos-benchmark/

We chose the simple lasso because we assumed it can only detect simple effects. We reasoned that it is a kind of linear model, and linear models cannot detect complex effects (e.g., interactions or unknown others) unless they are specifically modeled.

We also created datasets with complex true effects or EoS, including first-order interactions between two brain regions or nuclei. To assess the effectiveness of a MAREoS in removing first-order interaction-based complex EoS, we propose using a lasso algorithm with a design matrix that includes the first-order interactions. We chose the lasso with first-order interactions algorithm because, again, we assumed that, being a

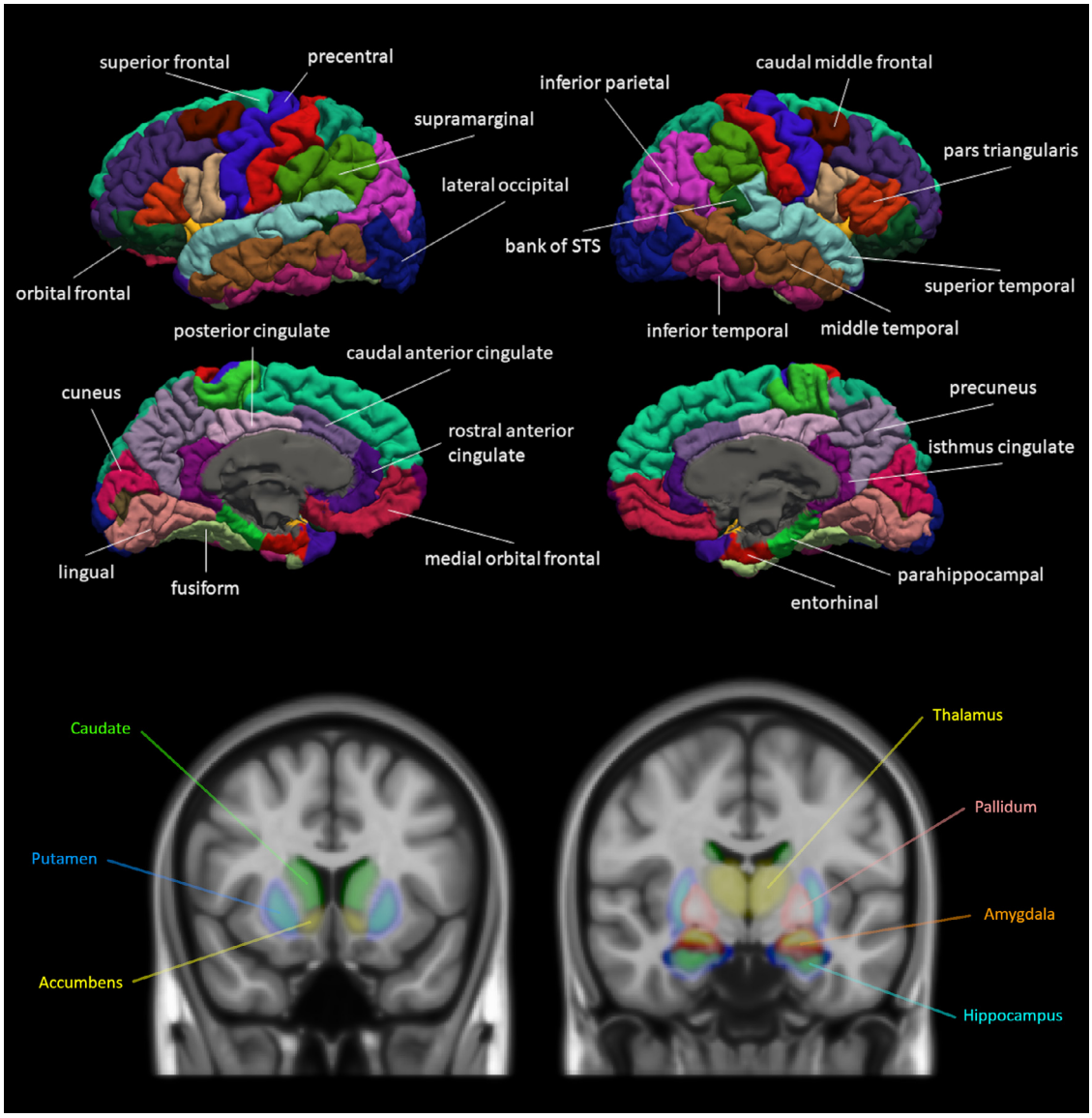

Figure 1. Location of the cortical regions and subcortical nuclei whose thickness, surface area, or volume we provide in the datasets.

linear model, it can only detect simple or first-order interaction-based effects.

We encourage other researchers to describe other complex EoS, create the respective datasets, and add them to the MAREoS benchmark website.

2.2. Creation of the datasets

For the interested reader, we will briefly report how we created each of these MRI datasets. We first generated normally distributed random data for each FreeSurfer region/nucleus, with means and standard deviations similar to real data (Radua et al., 2020). Then, to create simple EoS, we added differences between sites:

$$y^{*}_{r,i,j} = \delta_{r,i} \cdot \left(y_{r,i,j} - y_{r,i,\cdot}\right) + y_{r,i,\cdot} + \gamma_{r,i}$$

where $y_{r,i,j}$ is the cortical thickness, cortical surface area, or subcortical volume of the r^{th} ROI from the j^{th} individual of the i^{th} site, and $\delta_{r,i}$ and $\gamma_{r,i}$ are the multiplicate and additive EoS of the i^{th} site in the r^{th} ROI. We set both $\delta_{r,i}$ and $\gamma_{r,i}$ to follow normal distributions across the regions of a site, and $\delta_{\cdot,i}$ and $\gamma_{\cdot,i}$ to follow normal distributions across the sites. For further information about normally distributed multiplicative and additive effects, please see (Radua et al., 2020). To create interactions between ROIs, we swapped (between patients) the cortical thickness, cortical surface area, or subcortical volume of an ROI of a site to create positive or negative correlations with another ROI. For instance, imagine a site with only five patients where we aim to create a positive correlation between ROIs A and B. If the patient values in ROI A were [12,13,14,15,16], and the patient values in ROI B were [6,9,10,8,7], the correlation would be nearly null (r=0.1). However, after swapping patient values 10 and 7 in ROI B (i.e., [6,9,7,8,10], the correlation would be 0.7. Finally, we added some value to the responders to create true effects. After conducting these transformations, we added the effects of the covariates (adding some value multiplied by the covariate), truncated the resulting values to avoid outliers, and rescaled the data to be like FreeSurfer again.

We created many datasets, but we chose some that effectively only showed EoS or only showed true effects and were varied in features and BAC. To know which only showed EoS or only showed true effects, we used a logistic regression model to predict the response and calculated the accuracy using both standard formulas and the "multisite.accuracy" R package, which corrects for the site (Solanes et al., 2021). We considered "datasets with only EoS" those with ~50% mul-

Table 1
Example of measuring the simple effects of the site (EoS)-removal effectiveness for ComBat using the "simple" datasets.

Algorithm: lasso *without* interactions			
Dataset	Balanced accuracy		Relative accuracy change (RAC)
	Without MAREoS	With ComBat	
"Simple EoS #1"	74%	50%	100%
"Simple EoS #2"	75%	48%	107%
Average $RAC_{[simple\ EoS]}$: [(a)]			100%
"Simple true effects #1"	73%	73%	0%
"Simple true effects #2"	82%	82%	0%
Average $RAC_{[simple\ true\ effects]}$: [(a)]			0%
Simple EoS-removal effectiveness: [(a)]			100%

MAREoS: Method Aiming to Remove the EoS.
[(a)] Average *RAC*s and *EoS-removal effectiveness* are limited to 0-100%. These numbers may differ slightly from those reported at https://www.imardgroup.com/mareos-benchmark/ because the latter are based on a parallel collection of datasets for which the variable "response" is not public.

tisite accuracy – even if they showed high raw accuracy. Similarly, we considered "datasets with only true effects" those that showed similar (high) raw and multisite accuracies (Solanes et al., 2021).

2.3. Measurement of the EoS-removal effectiveness of a MAREoS for simple EoS

As detailed above, each dataset contains baseline multisite MRI data from patients and the subsequent responses to a given treatment. First, separately for each dataset and within a ten-fold cross-validation scheme, the developers must use the training subset to fit and apply a MAREoS to remove the EoS, find and remove the linear effects of two clinical covariates, fit the simple machine-learning algorithm, and use all these models to predict whether patients in the test subset respond to treatment. Second, again separately for each dataset, the developers must calculate the predictions' sensitivity, specificity, and balanced accuracy (BAC). The sensitivity is the percentage of responders correctly predicted to respond. The specificity is the percentage of non-responders correctly predicted not to respond. The BAC is the average of sensitivity and specificity. Relevantly, the developers must calculate the BAC using these basic formulas. In other words, they cannot correct the site with the "multisite.accuracy" R package (Solanes et al., 2021) that we would otherwise recommend. The reason is that we need uncorrected accuracies to measure the EoS-removal effectiveness of a MAREoS. Developers may download the specific R scripts to conduct all these steps from https://www.imardgroup.com/mareos-benchmark/. Afterward, they must perform the following calculations to measure the EoS-removal effectiveness of the MAREoS.

The first calculation, also performed separately for each dataset, is the relative accuracy change (*RAC*):

$$RAC = \frac{[BAC\ without\ MAREoS] - [BAC]}{[BAC\ without\ MAREoS] - 50\%}$$

To illustrate the idea, Table 1 shows the RAC calculations in the "Simple EoS #1" and "Simple true effects #1" datasets using a standard support vector machine (SVM) algorithm and ComBat (see details later). In the dataset with simple EoS, BAC was 74% when we fitted the SVM without applying any MAREoS. Using a MAREoS, the (EoS-inflated) BAC decreased to 50%. Then, $RAC_{[simple\ EoS]}$ in this dataset would be 100% (i.e., the accuracy is 100% closer to 50%). We could naively interpret that the MAREoS has reduced the bias by 100%. However, in the dataset with simple true effects, BAC decreased from 73.05% to 73.02% when using a MAREoS due to an undesirable potential side effect: data degradation. Then, $RAC_{[simple\ true\ effects]}$ in this dataset would be 0.14% (i.e., due to data degradation, the accuracy is 0.14% closer to 50%).

At a theoretical level, it might be interesting to note that the formula of the *RAC* would also work for accuracy increases – possible in datasets with both true effects and EoS. In such datasets, the EoS might prevent the machine-learning algorithm from fully detecting the true effects. For instance, BAC could be 70% before a MAREoS, while 75% after the MAREoS, for what *RAC* would be -25%, now meaning that the accuracy is now 25% farther from 50%. However, this *RAC* would be little informative because we would know neither the amount of EoS removed nor whether there was also data degradation.

Turning to the measurement of the EoS-removal effectiveness, the second calculation consists of adjusting the average $RAC_{[simple\ EoS]}$ (i.e., the naïve bias reduction) with the average $RAC_{[simple\ true\ effects]}$ (i.e., due to data degradation) to derive the EoS-removal effectiveness:

$$\text{EoS} - \text{removal effectiveness} = \frac{\text{mean}\left(RAC_{[\text{simple EoS}]}\right) - \text{mean}\left(RAC_{[\text{simple true effects}]}\right)}{100\% - \text{mean}\left(RAC_{[\text{simple true effects}]}\right)}$$

Limiting the average *RAC*s and the *EoS-removal effectiveness* to 0-100% may be sensible.

Going back to Table 1, if a MAREoS shows $RAC_{[simple\ EoS]}$ = 100% (a naïve 100% reduction in bias) and $RAC_{[simple\ true\ effects]}$ = 0.14% (a 0.14% decrease due to data degradation), then the simple EoS-removal effectiveness was 100%. In the datasets with only simple EoS, we may assume that the MAREoS would first remove simple EoS, reducing the accuracy. And afterward, it would degrade the data leading to (minimally) decreasing the remaining accuracy (Figure 2**, A1**). Or vice versa, we may assume that the MAREoS would first degrade the data, (minimally) decreasing the accuracy. And afterward, it would remove simple EoS, reducing the remaining accuracy (Figure 2**, A2**). Data degradation may seem negligible in this example, but it might be relevant in others.

2.4. Measurement of the EoS-removal effectiveness of a MAREoS for complex EoS

The overall strategy for measuring how well a MAREoS removes complex EoS is the same as for measuring how well a MAREoS removes simple EoS. However, datasets with complex EoS may also include simple true effects or simple EoS, and the MAREoS may remove both simple and complex EoS. Therefore, we may wish to subtract the part of the BAC attributable to simple effects as follows:

$$BAC_{[complex,corrected]} = BAC_{[complex]} - \left(BAC_{[simple]} - 0.5\right)$$

where $BAC_{[complex]}$ is the BAC obtained with the complex machine-learning algorithm (e.g., a lasso with first-order interactions), $BAC_{[simple]}$ is the BAC obtained with the simple machine-learning algorithm (i.e., the lasso *without* interactions), and $BAC_{[complex,corrected]}$ is the $BAC_{[complex]}$ after "subtracting" the simple effects.

One way to see this subtraction from a different perspective is to decompose the accuracy of the complex machine-learning algorithm. Imagine that we have a sample of 100 patients, half responders. Suppose we predict randomly, simply tossing a coin. In that case, we will guess correctly by chance half of the time for what we expect to predict about 50 individuals correctly. Now imagine that a simple machine-learning algorithm correctly predicts 70 individuals. However, we can decompose this number as 50+20, with the 50 corresponding to the number of individuals that we can correctly predict tossing a coin and the 20 corresponding to the extra accuracy provided by the simple effects detected by the simple machine-learning algorithm. Finally, imagine that a complex machine-learning algorithm correctly predicts 85 individuals. Again, we can decompose this number as 50+20+15, with the 20 corresponding to the extra accuracy provided by the simple effects detected by the complex machine-learning algorithm and the 15 corresponding

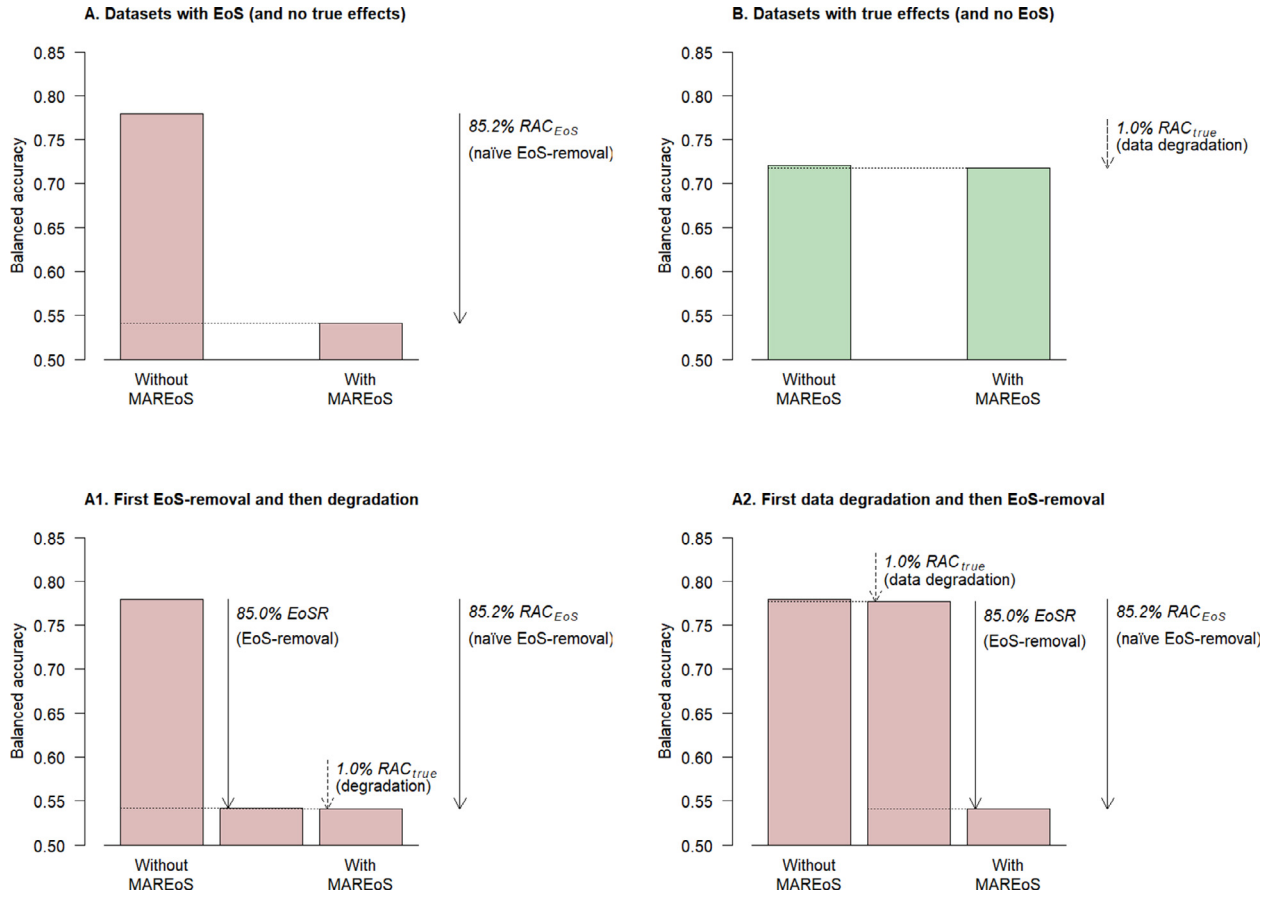

Figure 2. Example of the relative accuracy changes (*RAC*) measurement in datasets with only simple effects of the site (EoS) and only simple true effects, and their relationship with naïve EoS-removal, data degradation, and EoS-removal.

Table 2
Example of measuring the simple effects of the site (EoS)-removal effectiveness for ComBat using the "interaction" datasets.

Algorithm: lasso *without* interactions			
Dataset	Balanced accuracy		Relative accuracy change (RAC)
	Without MAREoS	With ComBat	
"Interaction EoS #1"	63%	50%	100%
"Interaction EoS #2"	66%	50%	100%
Average RAC[simple EoS]: [(a)]			100%
"Interaction true effects #1"	64%	64%	-1%
"Interaction true effects #2"	60%	60%	-4%
Average RAC[simple true effects]: [(a)]			0%
Simple EoS-removal effectiveness: [(a)]			100%

MAREoS: Method Aiming to Remove the EoS. For "Interaction true effects #2", note that the balanced accuracy is indeed 60.0% without MAREoS and 60.4% with ComBat.
[(a)] Average *RAC*s and *EoS-removal effectiveness* are limited to 0-100%.

to the extra accuracy offered by the complex effects beyond the simple effects.

To exemplify the method, we report the calculations for the "Interaction EoS #2" and "Interaction true effects #2" datasets using SVM (Table 6, see details later). In the dataset with complex EoS, $BAC_{[complex]}$ was 87% when we fitted the SVM without applying any MAREoS. However, SVM predictions were likely not exclusively based on complex effects but also simple effects. To quantify these simple effects, we used a simple lasso, which showed a $BAC_{[simple]}$ of 66% (Table 2), i.e., 16% more than tossing a coin. Removing the part of accuracy due to simple effects with the formula above, we estimated that SVM's $BAC_{[complex,corrected]}$ was 87% - 16% = 71%. Using a MAREoS, SVM's $BAC_{[complex]}$ decreased to 66% and lasso's $BAC_{[simple]}$ to 50% (i.e., 0% over tossing a coin, Table 2). We thus estimated that using a MAREoS, SVM's $BAC_{[complex,corrected]}$ was 66% - 0% = 66%. Finally, we could estimate that $RAC_{[complex\ EoS]}$ was 21% using the same formula as for simple effects:

$$RAC_{[complex]} = \frac{[BAC\ without\ MAREoS]_{[complex,corrected]} - [BAC]_{[complex,corrected]}}{[BAC\ without\ MAREoS]_{[complex,corrected]} - 50\%}$$

We did the same calculations for the dataset with complex true effects. Before applying any MAREoS, SVM's $BAC_{[complex]}$ was 67%

(Table 6) and lasso $BAC_{[simple]}$ 60% (i.e., 10% over tossing a coin, Table 2), resulting in SVM's $BAC_{[complex,corrected]}$ = 67% - 10% = 57%. Using a MAREoS, SVM's $BAC_{[complex]}$ was 67% (Table 6), and lasso's $BAC_{[simple]}$ was 60% (i.e., 10% over tossing a coin, Table 2), resulting in SVM's $BAC_{[complex,corrected]}$ = 67% - 10% = 57%. We could then estimate $RAC_{[complex\ true\ effects]}$=0.25%. Finally, if $RAC_{[complex\ EoS]}$ was 21% (a naïve 21% reduction in complex EoS-related bias) and $RAC_{[complex\ true\ effects]}$ = 0.25% (a 0.25% decrease due to complex EoS-related data degradation), then the complex EoS-removal effectiveness was (minimally lower than) 21%.

2.5. Example: ComBat

To exemplify how to measure the EoS-removal effectiveness of a MAREoS, we applied ComBat to the public datasets provided. We then conducted the calculations needed to measure the EoS-removal effectiveness.

First, we downloaded the public datasets from https://www.imardgroup.com/mareos-benchmark/ . Each dataset is a table with the following columns: the identification of the simulated individual, the MRI data (cortical thickness or surface area or subcortical volumes), the site, the values of the two clinical covariates, and the distribution in folds. We also downloaded the Johnson-Fortin-Radua version of the ComBat MAREoS (Fortin et al., 2018; Johnson et al., 2007; Radua et al., 2020) from https://enigma.ini.usc.edu/protocols/statistical-protocols/

The following analyses refer to one dataset. For fold 1, we defined the training subset as individuals in folds 2-10 and the test subset as the set of individuals in fold 1. In the training subset: a) we fitted the ComBat model with the function "combat_fit"; b) we removed the EoS according to the ComBat model with the function "combat_apply"; c) we fitted regressions to estimate the linear effects of the two clinical covariates (a separate linear regression per each brain region); d) we removed the linear effects of the clinical covariates according to these linear regressions; e) and we fitted the lasso algorithm (without interactions when assessing simple effects, or with first-order interactions when assessing complex effects including first-order interactions). Afterward, in the test subset: a) we removed the EoS according to the ComBat model (fitted with the training subset) with function "combat_apply"; b) we removed the linear effects of the clinical covariates according to the linear regressions (fitted with the training subset); c) and we applied the lasso algorithm (fitted with the training subset) to predict the individual responses. After repeating the same procedure for folds 2-10, we had predicted the response in all individuals. We then proceeded to calculate the sensitivity, specificity, and BAC. Finally, we combined the BACs with and without ComBat to calculate the *RAC*. We provide the R scripts to conduct such calculations at https://www.imardgroup.com/mareos-benchmark/

After conducting these analyses for each "Simple" dataset (Table 1), we had a *RAC* for each of the "Simple EoS" datasets (100% and 107%), which we averaged (and limited to 0-100%) to obtain an average $RAC_{[simple\ EoS]}$ of 100%. Similarly, we had a *RAC* for each of the "Simple true effects" datasets (0% and 0%), which we averaged to obtain an average $RAC_{[simple\ true\ effects]}$ of 0%. Finally, we calculated the *EoS-removal effectiveness* (100%). Therefore, we should conclude that ComBat entirely removes the bias related to simple EoS (100%) and has negligible data degradation (0%).

The results were nearly identical when we used a lasso algorithm *without* interactions for the "Interaction" datasets (Table 2). The *RAC*s for the "Interaction EoS" datasets were 100% and 100% (average $RAC_{[simple\ EoS]}$ = 100%), and the *RAC*s for the "Interaction true effects" datasets were -1% and -4% (average $RAC_{[simple\ true\ effects]}$ = 0%), leading to *EoS-removal effectiveness* = 100%. Therefore, we should conclude again that ComBat entirely removes the bias related to simple EoS (100%) and has negligible data degradation (0%). These datasets had complex effects (e.g., the multiplication of pairs of brain regions differed between groups). However, these effects are not detectable by the lasso algorithm *without* interactions; thus, they did not influence these calculations.

The results were very different when we analyzed the same "Interaction" datasets using a lasso algorithm *with* first-order interactions (Table 3). First, all BAC were substantially higher (e.g., 80% instead of 63% for "Interaction EoS #1" without MAREoS). This increase is because this machine-learning algorithm could detect the first-order interactions present in these datasets. However, we must highlight here that, as we saw in the previous paragraph, these datasets also included simple effects, which we had to subtract before specifically studying the complex effects. For instance, for "Interaction EoS #1" without MAREoS, BAC was 80%, but we subtracted 13% (i.e., the BAC of the simple algorithm, 63%, minus 50%) for what the corrected BAC was 80% - 13% = 67%. Once we corrected all BACs, we proceeded as before. The *RAC*s for the "Interaction EoS" datasets were -5% and -13% (average $RAC_{[complex\ EoS]}$ = 0%), and the *RAC*s for the "Interaction true effects" datasets were 3% and 5% (average $RAC_{[complex\ true\ effects]}$ = 4%), leading to *EoS-removal effectiveness* = 0%. Therefore, we should conclude that ComBat does not remove the bias related to complex EoS (0%) and may show minor data degradation (4%).

2.6. Other machine-learning algorithms

We repeated the above calculations with standard random forest (Liaw and Wiener, 2002), support vector machine (Meyer et al., 2021), and gaussian processes (Karatzoglou et al., 2004) algorithms to provide insights on the use of MAREoS with these machine-learning algorithms. We used the default options of the "", "", and "" R packages, which involve radial basis function kernels for support vector machine and gaussian processes. We show again the R code to conduct such calculations at https://www.imardgroup.com/mareos-benchmark/.

With the "simple" datasets, BACs, *RAC*s, and simple EoS-removal effectiveness were similar when using lasso, random forest, support vector machine, or gaussian processes algorithms (Table 4). There were differences between algorithms, but they were small and likely due to chance.

The analysis of the "interaction" datasets showed that the random forest (and, to a lesser extent, the support vector machine) algorithm detects the complex effects in these datasets (Table 5). In contrast, the Gaussian processes algorithm only detects some. We thus repeated the calculations of complex EoS-removal effectiveness for the random forest and support vector machine algorithms.

Table 6 shows that random forests and support vector machine algorithms yielded $RAC_{[complex\ EoS]}$ substantially different from the 0% calculated using the lasso algorithm with first-order interactions. As introduced earlier, we assumed that lasso with first-order interactions could detect only one type of complex EoS: those based on first-order interactions. Thus, a $RAC_{[complex\ EoS]}$ of 0% meant that ComBat does not remove these complex EoS. However, $RAC_{[complex\ EoS]}$ were 34-52% for random forest and support vector machine algorithms. Therefore, we should conclude that random forest and support vector machine algorithms detect a mixture of complex EoS, some of which are removable by ComBat and others are not.

2.7. Benchmark website

For machine-learning researchers, the website (https://www.imardgroup.com/mareos-benchmark/) includes information about the types of effects detectable by different machine-learning algorithms and the effectiveness of MAREoS in removing different types of EoS. See Figure 3 for a diagram of the steps to choose an appropriate MAREoS for a specific study. Note that the numbers in the website may be slightly different from those in the manuscript because the former are based on a parallel collection of datasets for which the variable "response" is not public. We created the latter datasets to keep objective ranks of the effectiveness of the different MAREoS for different types of EoS.

Table 3
Example of measuring the "complex effects of the site (EoS) including interactions"-removal effectiveness for ComBat using the "interaction" datasets.

Algorithm: lasso *with* first-order interactions					
Dataset	Balanced accuracy		Corrected balanced accuracy [a]		Relative accuracy change (RAC)
	Without MAREoS	With ComBat	Without MAREoS	With ComBat	
"Interaction EoS #1"	80%	68%	67%	68%	-5%
"Interaction EoS #2"	85%	71%	69%	71%	-13%
Average $RAC_{[complex\ EoS]}$: [b]					0%
"Interaction true effects #1"	81%	81%	67%	67%	3%
"Interaction true effects #2"	76%	75%	66%	65%	5%
Average $RAC_{[complex\ true\ effects]}$: [b]					4%
"Complex EoS including interactions"-removal effectiveness: [b]					0%

MAREoS: Method Aiming to Remove the EoS.
[a] Subtracting the simple effects, estimated as BAC with lasso *without* interactions minus 50%.
[b] Average *RAC*s and *EoS-removal effectiveness* are limited to 0-100%. These numbers may differ slightly from those reported at https://www.imardgroup.com/mareos-benchmark/ because the latter are based on a parallel collection of datasets for which the variable "response" is not public.

Table 4
Alternative measurement with standard random forests (RF), support vector machine (SVM), and Gaussian processes (GP) algorithms of the simple effects of the site (EoS)-removal effectiveness for ComBat using the "simple" datasets.

Algorithms: random forests (RF), support vector machine (SVM), and Gaussian processes (GP)										
Dataset	Balanced accuracy							Relative accuracy change (RAC)		
	Without MAREoS			With ComBat						
	RF	SVM	GP	RF	SVM		GP	RF	SVM	GP
"Simple EoS #1"	77%	78%	78%	54%	54%		52%	84%	85%	92%
"Simple EoS #2"	75%	75%	75%	55%	52%		51%	81%	92%	94%
Average $RAC_{[simple\ EoS]}$: [a]								82%	88%	93%
"Simple true effects #1"	74%	72%	71%	74%	72%	71%		0%	1%	2%
"Simple true effects #2"	84%	83%	81%	83%	83%	81%		1%	1%	-1%
Average $RAC_{[simple\ true\ effects]}$: [a]								0%	1%	0%
"Simple-based EoS"-removal effectiveness: [a]								82%	88%	93%

MAREoS: Method Aiming to Remove the EoS.
[a] Average *RAC*s and *EoS-removal effectiveness* are limited to 0-100%. These numbers may differ slightly from those reported at https://www.imardgroup.com/mareos-benchmark/ because the latter are based on a parallel collection of datasets for which the variable "response" is not public.

Table 5
Detection of complex effects including interactions by standard random forests (RF), support vector machine (SVM), and Gaussian processes (GP) algorithms in the "interaction true effects" datasets.

Algorithms: lasso with first-order interactions (LFOI), random forests (RF), support vector machine (SVM), and Gaussian processes (GP)									
Dataset	Balanced accuracy								
	Observed					Extra accuracy compared to lasso *without* interactions			
	Lasso *without* interactions	LFOI	RF	SVM	GP	LFOI	RF	SVM	GP
"Interaction true effects #1"	64%	81%	84%	73%	65%	+17%	+20%	+9%	1%
"Interaction true effects #2"	60%	76%	77%	67%	62%	+16%	+17%	+7%	2%
Average extra accuracy: [a]						+16%	+18%	+8%	2%

MAREoS: Method Aiming to Remove the EoS.
[a] These numbers may differ slightly from those reported at https://www.imardgroup.com/mareos-benchmark/ because the latter are based on a parallel collection of datasets for which the variable "response" is not public.

Developers wanting to add a MAREoS to the website should download these datasets and conduct calculations analogous to the ones described in the example, except for not fitting/applying the simple machine-learning algorithm because the response to the treatment is non-public. Instead, they must save the pre-processed MRI data of the training and test subsets of each fold, along with an identification of these sets (e.g., "train_fold1", "test_fold1", "train_fold2", etcetera). For instance, in the first fold of the cross-validation, users should: a) find the EoS and the linear effects of the covariates using individuals labeled to be in folds 2 to 10; b) remove these effects from these individuals and

Table 6
Alternative measurement with random forests (RF) and standard support vector machine (SVM) algorithms of the "complex effects of the site (EoS) including interactions"-removal effectiveness for ComBat using the "interaction" datasets.

Algorithms: random forests (RF) and support vector machine (SVM)										
Dataset	Balanced accuracy				Corrected balanced accuracy [a]				Relative accuracy change (RAC)	
	Without MAREoS		With ComBat		Without MAREoS		With ComBat			
	RF	SVM	RF	SVM	RF	SVM	RF	SVM	RF	SVM
"Interaction EoS #1"	82%	73%	54%	55%	69%	60%	54%	55%	81%	47%
"Interaction EoS #2"	86%	87%	65%	66%	70%	71%	65%	66%	22%	21%
Average RAC$_{[complex\ EoS]}$: [b]									52%	34%
"Interaction true effects #1"	84%	73%	84%	74%	70%	59%	70%	60%	-1%	-4%
"Interaction true effects #2"	77%	67%	77%	67%	67%	57%	66%	57%	3%	0%
Average RAC$_{[complex\ true\ effects]}$: [b]									1%	0%
"Complex EoS including interactions"-removal effectiveness: [b]									51%	34%

MAREoS: Method Aiming to Remove the EoS.
[a] Subtracting the simple effects, estimated as BAC with the lasso algorithm *without* interactions minus 50%.
[b] Average *RACs* and *EoS-removal effectiveness* are limited to 0-100%. These numbers may differ slightly from those reported at https://www.imardgroup.com/mareos-benchmark/ because the latter are based on a parallel collection of datasets for which the variable "response" is not public.

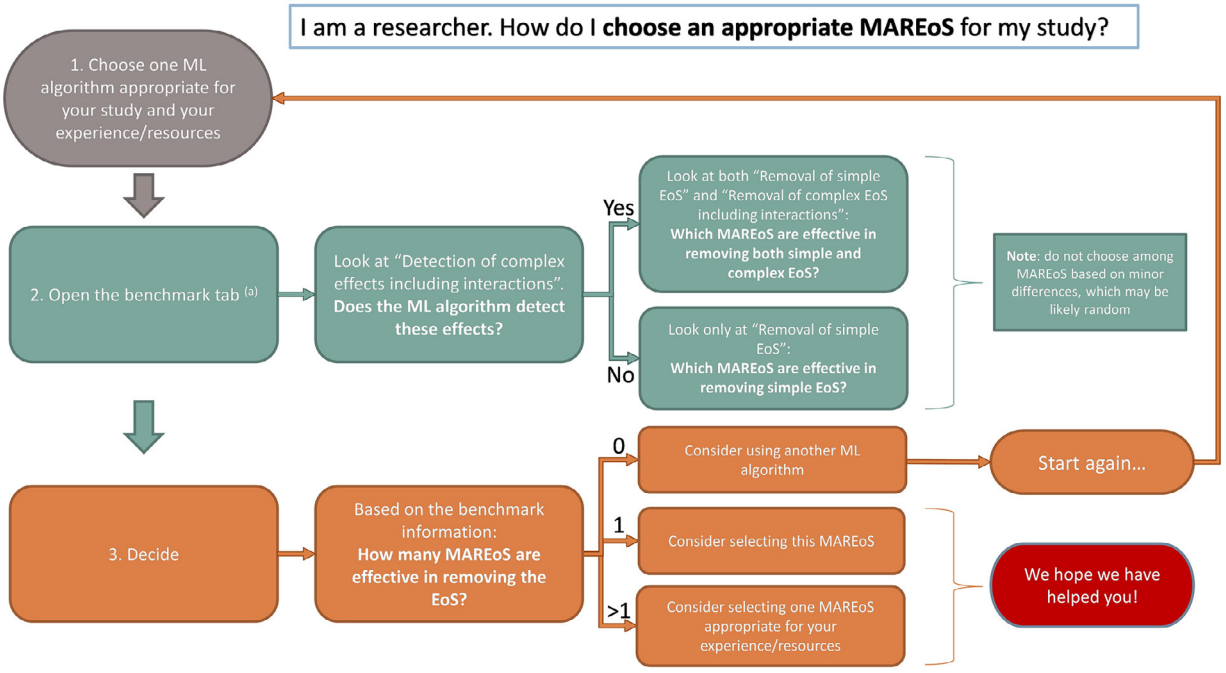

Figure 3. Steps to choose an appropriate MAREoS for a specific study – for machine-learning researchers.

save the resulting data with the set identification "train_fold1"; and c) remove these effects from individuals labeled to be in fold 1 and save the resulting data with the set identification "test_fold1". See Figure 4 for a diagram of the steps to add a MAREoS to the extendable benchmark website.

Developers wanting to add a new type of EoS to the website may contact us directly. We also welcome researchers and developers wishing to add the effectiveness of an already investigated MAREoS using an alternative machine-learning algorithm.

3. Discussion

This work presents a strategy to measure the EoS-removal effectiveness of a MAREoS in multisite machine-learning studies. We provide datasets with only simple true effects, datasets with only simple EoS, datasets with complex true effects, datasets with complex EoS, and formulas to measure the EoS-removal effectiveness from the BAC obtained when fitting prediction models in these datasets. We also provide a benchmark website to rank the EoS-removal effectiveness of the different MAREoS and a relationship of the types of EoS that may bias accuracy for different machine-learning algorithms. For instance, we report that ComBat removes all simple EoS as needed for predictions based on simple lasso algorithms while it leaves residual complex EoS that may bias the predictions based on standard support vector machine algorithms. In other words, the extendable benchmark website provides the types of EoS that researchers should remove for a given machine-learning algorithm and the effectiveness of each MAREoS for removing each type of EoS.

The most important limitation of the present work is that it encompasses only one type of complex EoS: those due to first-order interactions

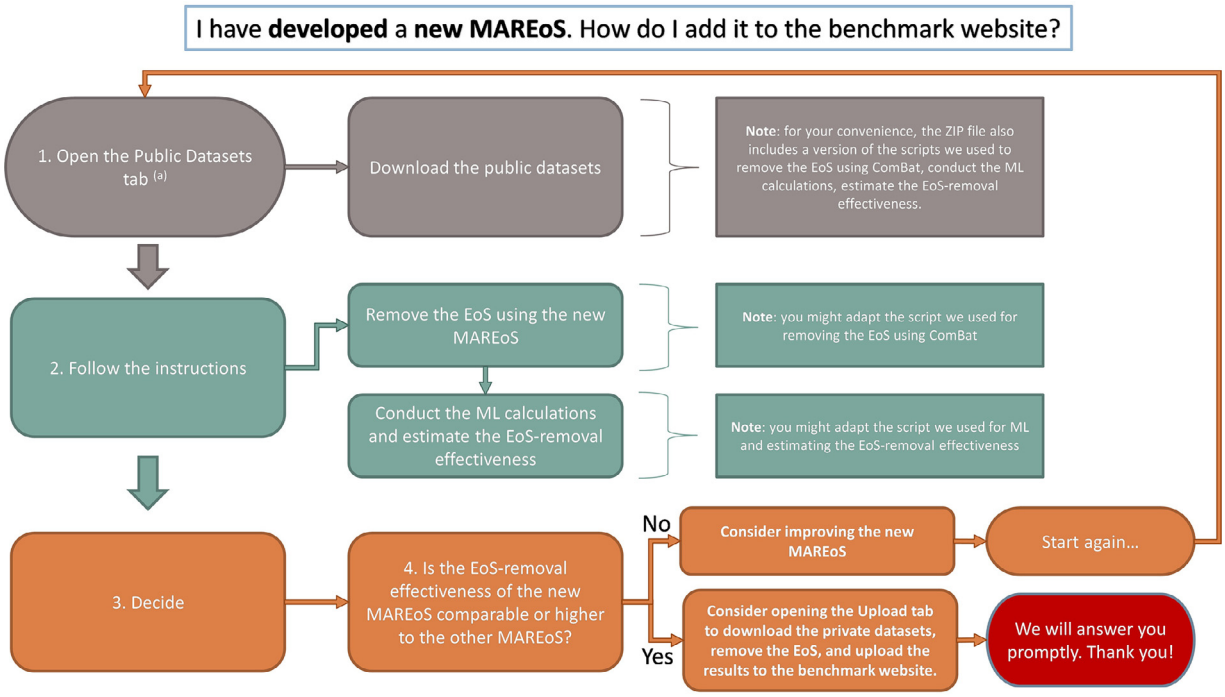

Figure 4. Steps to add a MAREoS to the extendable benchmark website – for MAREoS developers.

between brain regions. Other complex EoS could potentially derive from higher-order interactions or other regional relationships. However, we believe that the investigation of new types of complex EoS, along with the (challenging) development of methods to measure them (e.g., creating specific datasets), should be the work of future studies. We created an extendable benchmark website for this reason. Another potential limitation of this work is that we only created binary outcomes (response vs. no response). We chose this distribution for the simplicity of its definition of accuracy (percentage of correct predictions). The definitions of accuracy in other distributions may be less straightforward. For instance, for continuous outcomes, there may be several metrics (absolute or squared difference between observed and predicted, correlation between observed and predicted, etcetera). However, MAREoS remove differences between sites independently of the outcomes; thus, the benchmarking should be similar using binary or other outcomes.

We want to finish by highlighting other exciting approaches to handling EoS. We believe that, when possible, researchers should use them along with our approach to providing richer complementary insights. For instance, a small set of individuals may volunteer to be scanned in the different devices used in a multisite MRI study (Kurokawa et al., 2021; Noble et al., 2017; Tanaka et al., 2021; Tong et al., 2020). The data from these individuals, known as “traveling subjects”, have two valuable characteristics. First, they are real data and thus very likely have hidden features that our datasets may not have. Indeed, some data indicate that the traveling-subject outperforms ComBat (Maikusa et al., 2021). Second, these data allow an excellent study of the differences between MRI devices. In this scenario, differences between sites unrelated to MRI devices should be negligible. Together, these characteristics enable the development of promising deep learning-based MAREoS (Tian et al., 2022). However, the traveling-subject approach also has its drawbacks. For instance, it can only be done prospectively (i.e., it is not helpful for mega-studies based on previously acquired data, such as those in the ENIGMA consortium (Dima et al., 2022)). Also, it may be costly (requires traveling) for which the set of subjects is usually tiny, though new projects such as the BMB-HBM aim to overcome this hurdle (Koike et al., 2021).

4. Disclosures

Dr. Vieta has received grants and served as a consultant, advisor, or CME speaker for the following entities (work unrelated to the topic of this manuscript): AB-Biotics, Abbott, Allergan, Angelini, Dainippon Sumitomo Pharma, Galenica, Janssen, Lundbeck, Novartis, Otsuka, Sage, Sanofi-Aventis, and Takeda. Dr. Llufriu has received compensation for consulting services and speaker honoraria from Biogen Idec, Novartis, TEVA, Genzyme, Sanofi, and Merck

Data and code availability statement

The data and scripts used in this study are available at https://www.imardgroup.com/mareos-benchmark/

Data availability

All data are available at https://www.imardgroup.com/mareos-benchmark/

Acknowledgement

This work was supported by the Baszucki Brain Research Fund and Milken Institute (“The ENIGMA Bipolar Medications Initiative: A Global Study of Bipolar Disorder Medication Associations in the Brain” grant) and the Spanish Ministry of Science, Innovation and Universities / Economy and Competitiveness / Instituto de Salud Carlos III (CPII19/00009, PI19/00394, FI20/00047), co-financed by ERDF Funds from the European Commission ("A Way of Making Europe").

REFERENCES

Albajes-Eizagirre, A., Solanes, A., Vieta, E., Radua, J., 2019. Voxel-based meta-analysis via permutation of subject images (PSI): Theory and implementation for SDM. Neuroimage 186, 174–184.

Dima, D., Modabbernia, A., Papachristou, E., Doucet, G.E., Agartz, I., Aghajani, M., Akudjedu, T.N., Albajes-Eizagirre, A., Alnaes, D., Alpert, K.I., Andersson, M., Andreasen, N.C., Andreassen, O.A., Asherson, P., Banaschewski, T., Bargallo, N.,

Baumeister, S., Baur-Streubel, R., Bertolino, A., Bonvino, A., Boomsma, D.I., Borgwardt, S., Bourque, J., Brandeis, D., Breier, A., Brodaty, H., Brouwer, R.M., Buitelaar, J.K., Busatto, G.F., Buckner, R.L., Calhoun, V., Canales-Rodriguez, E.J., Cannon, D.M., Caseras, X., Castellanos, F.X., Cervenka, S., Chaim-Avancini, T.M., Ching, C.R.K., Chubar, V., Clark, V.P., Conrod, P., Conzelmann, A., Crespo-Facorro, B., Crivello, F., Crone, E.A., Dannlowski, U., Dale, A.M., Davey, C., de Geus, E.J.C., de Haan, L., de Zubicaray, G.I., den-Braber, A., Dickie, E.W., Di Giorgio, A., Doan, N.T., Dorum, E.S., Ehrlich, S., Erk, S., Espeseth, T., Fatouros-Bergman, H., Fisher, S.E., Fouche, J.P., Franke, B., Frodl, T., Fuentes-Claramonte, P., Glahn, D.C., Gotlib, I.H., Grabe, H.J., Grimm, O., Groenewold, N.A., Grotegerd, D., Gruber, O., Gruner, P., Gur, R.E., Gur, R.C., Hahn, T., Harrison, B.J., Hartman, C.A., Hatton, S.N., Heinz, A., Heslenfeld, D.J., Hibar, D.P., Hickie, I.B., Ho, B.C., Hoekstra, P.J., Hohmann, S., Holmes, A.J., Hoogman, M., Hosten, N., Howells, F.M., Hulshoff-Pol, H.E., Huyser, C., Jahanshad, N., James, A., Jernigan, T.L., Jiang, J., Jonsson, E.G., Joska, J.A., Kahn, R., Kalnin, A., Kanai, R., Klein, M., Klyushnik, T.P., Koenders, L., Koops, S., Kramer, B., Kuntsi, J., Lagopoulos, J., Lazaro, L., Lebedeva, I., Lee, W.H., Lesch, K.P., Lochner, C., Machielsen, M.W.J., Maingault, S., Martin, N.G., Martinez-Zalacain, I., Mataix-Cols, D., Mazoyer, B., McDonald, C., McDonald, B.C., McIntosh, A.M., McMahon, K.L., McPhilemy, G., Meinert, S., Menchon, J.M., Medland, S.E., Meyer-Lindenberg, A., Naaijen, J., Najt, P., Nakao, T., Nordvik, J.E., Nyberg, L., Oosterlaan, J., de la Foz, V.O., Paloyelis, Y., Pauli, P., Pergola, G., Pomarol–Clotet, E., Portella, M.J., Potkin, S.G., Radua, J., Reif, A., Rinker, D.A., Roffman, J.L., Rosa, P.G.P., Sacchet, M.D., Sachdev, P.S., Salvador, R., Sanchez-Juan, P., Sarro, S., Satterthwaite, T.D., Saykin, A.J., Serpa, M.H., Schmaal, L., Schnell, K., Schumann, G., Sim, K., Smoller, J.W., Sommer, I., Soriano-Mas, C., Stein, D.J., Strike, L.T., Swagerman, S.C., Tamnes, C.K., Temmingh, H.S., Thomopoulos, S.I., Tomyshev, A.S., Tordesillas-Gutierrez, D., Trollor, J.N., Turner, J.A., Uhlmann, A., van den Heuvel, O.A., van den Meer, D., van der Wee, N.J.A., van Haren, N.E.M., Van't Ent, D., van Erp, T.G.M., Veer, I.M., Veltman, D.J., Voineskos, A., Volzke, H., Walter, H., Walton, E., Wang, L., Wang, Y., Wassink, T.H., Weber, B., Wen, W., West, J.D., Westlye, L.T., Whalley, H., Wierenga, L.M., Williams, S.C.R., Wittfeld, K., Wolf, D.H., Worker, A., Wright, M.J., Yang, K., Yoncheva, Y., Zanetti, M.V., Ziegler, G.C., Thompson, P.M., Frangou, S., Karolinska Schizophrenia, P., 2022. Subcortical volumes across the lifespan: Data from 18,605 healthy individuals aged 3-90 years. Hum. Brain Mapp. 43, 452–469.

Favre, P., Pauling, M., Stout, J., Hozer, F., Sarrazin, S., Abe, C., Alda, M., Alloza, C., Alonso-Lana, S., Andreassen, O.A., Baune, B.T., Benedetti, F., Busatto, G.F., Canales-Rodriguez, E.J., Caseras, X., Chaim-Avancini, T.M., Ching, C.R.K., Dannlowski, U., Deppe, M., Eyler, L.T., Fatjo-Vilas, M., Foley, S.F., Grotegerd, D., Hajek, T., Haukvik, U.K., Howells, F.M., Jahanshad, N., Kugel, H., Lagerberg, T.V., Lawrie, S.M., Linke, J.O., McIntosh, A., Melloni, E.M.T., Mitchell, P.B., Polosan, M., Pomarol–Clotet, E., Repple, J., Roberts, G., Roos, A., Rosa, P.G.P., Salvador, R., Sarro, S., Schofield, P.R., Serpa, M.H., Sim, K., Stein, D.J., Sussmann, J.E., Temmingh, H.S., Thompson, P.M., Verdolini, N., Vieta, E., Wessa, M., Whalley, H.C., Zanetti, M.V., Leboyer, M., Mangin, J.F., Henry, C., Duchesnay, E., Houenou, J., Group, E.B.D.W., 2019. Widespread white matter microstructural abnormalities in bipolar disorder: evidence from mega- and meta-analyses across 3033 individuals. Neuropsychopharmacology 44, 2285–2293.

Fortin, J.P., Cullen, N., Sheline, Y.I., Taylor, W.D., Aselcioglu, I., Cook, P.A., Adams, P., Cooper, C., Fava, M., McGrath, P.J., McInnis, M., Phillips, M.L., Trivedi, M.H., Weissman, M.M., Shinohara, R.T., 2018. Harmonization of cortical thickness measurements across scanners and sites. Neuroimage 167, 104–120.

Friedman, J., Hastie, T., Tibshirani, R., 2010. Regularization Paths for Generalized Linear Models via Coordinate Descent. J Stat Softw 33, 1–22.

Hosseini, M., Powell, M., Collins, J., Callahan-Flintoft, C., Jones, W., Bowman, H., Wyble, B., 2020. I tried a bunch of things: The dangers of unexpected overfitting in classification of brain data. Neurosci. Biobehav. Rev. 119, 456–467.

Johnson, W.E., Li, C., Rabinovic, A., 2007. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics 8, 118–127.

Karatzoglou, A., Smola, A., Hornik, K., Zeileis, A., 2004. kernlab - An S4 Package for Kernel Methods in R. J Stat Softw 11, 1–20.

Koike, S., Tanaka, S.C., Okada, T., Aso, T., Yamashita, A., Yamashita, O., Asano, M., Maikusa, N., Morita, K., Okada, N., Fukunaga, M., Uematsu, A., Togo, H., Miyazaki, A., Murata, K., Urushibata, Y., Autio, J., Ose, T., Yoshimoto, J., Araki, T., Glasser, M.F., Van Essen, D.C., Maruyama, M., Sadato, N., Kawato, M., Kasai, K., Okamoto, Y., Hanakawa, T., Hayashi, T., Brain, M.B.H.B.M.R.I.G., 2021. Brain/MINDS beyond human brain MRI project: A protocol for multi-level harmonization across brain disorders throughout the lifespan. Neuroimage Clin 30, 102600.

Kurokawa, R., Kamiya, K., Koike, S., Nakaya, M., Uematsu, A., Tanaka, S.C., Kamagata, K., Okada, N., Morita, K., Kasai, K., Abe, O., 2021. Cross-scanner reproducibility and harmonization of a diffusion MRI structural brain network: A traveling subject study of multi-b acquisition. Neuroimage 245, 118675.

Liaw, A., Wiener, M., 2002. Classification and Regression by randomForest. R News 2, 18–22.

Maikusa, N., Zhu, Y., Uematsu, A., Yamashita, A., Saotome, K., Okada, N., Kasai, K., Okanoya, K., Yamashita, O., Tanaka, S.C., Koike, S., 2021. Comparison of traveling–subject and ComBat harmonization methods for assessing structural brain characteristics. Hum. Brain Mapp. 42, 5278–5287.

Meyer, D., Dimitriadou, D., Hornik, K., Weingessel, A., Leisch, F., 2021. e1071: Misc Functions of the Department of Statistics. Probability Theory Group (Formerly: E1071). TU, Wien.

Noble, S., Scheinost, D., Finn, E.S., Shen, X., Papademetris, X., McEwen, S.C., Bearden, C.E., Addington, J., Goodyear, B., Cadenhead, K.S., Mirzakhanian, H., Cornblatt, B.A., Olvet, D.M., Mathalon, D.H., McGlashan, T.H., Perkins, D.O., Belger, A., Seidman, L.J., Thermenos, H., Tsuang, M.T., van Erp, T.G.M., Walker, E.F., Hamann, S., Woods, S.W., Cannon, T.D., Constable, R.T., 2017. Multisite reliability of MR-based functional connectivity. Neuroimage 146, 959–970.

Radua, J., Vieta, E., Shinohara, R., Kochunov, P., Quide, Y., Green, M.J., Weickert, C.S., Weickert, T., Bruggemann, J., Kircher, T., Nenadic, I., Cairns, M.J., Seal, M., Schall, U., Henskens, F., Fullerton, J.M., Mowry, B., Pantelis, C., Lenroot, R., Cropley, V., Loughland, C., Scott, R., Wolf, D., Satterthwaite, T.D., Tan, Y., Sim, K., Piras, F., Spalletta, G., Banaj, N., Pomarol-Clotet, E., Solanes, A., Albajes-Eizagirre, A., Canales-Rodriguez, E.J., Sarro, S., Di Giorgio, A., Bertolino, A., Stablein, M., Oertel, V., Knochel, C., Borgwardt, S., du Plessis, S., Yun, J.Y., Kwon, J.S., Dannlowski, U., Hahn, T., Grotegerd, D., Alloza, C., Arango, C., Janssen, J., Diaz-Caneja, C., Jiang, W., Calhoun, V., Ehrlich, S., Yang, K., Cascella, N.G., Takayanagi, Y., Sawa, A., Tomyshev, A., Lebedeva, I., Kaleda, V., Kirschner, M., Hoschl, C., Tomecek, D., Skoch, A., van Amelsvoort, T., Bakker, G., James, A., Preda, A., Weideman, A., Stein, D.J., Howells, F., Uhlmann, A., Temmingh, H., Lopez-Jaramillo, C., Diaz-Zuluaga, A., Fortea, L., Martinez-Heras, E., Solana, E., Llufriu, S., Jahanshad, N., Thompson, P., Turner, J., van-Erp, T.collaborators, E.C., 2020. Increased power by harmonizing structural MRI site differences with the ComBat batch adjustment method in ENIGMA. Neuroimage 218, 116956.

Solanes, A., Palau, P., Fortea, L., Salvador, R., Gonzalez-Navarro, L., Llach, C.D., Valenti, M., Vieta, E., Radua, J., 2021. Biased accuracy in multisite machine-learning studies due to incomplete removal of the effects of the site. Psychiatry Res Neuroimaging 314, 111313.

Tanaka, S.C., Yamashita, A., Yahata, N., Itahashi, T., Lisi, G., Yamada, T., Ichikawa, N., Takamura, M., Yoshihara, Y., Kunimatsu, A., Okada, N., Hashimoto, R., Okada, G., Sakai, Y., Morimoto, J., Narumoto, J., Shimada, Y., Mano, H., Yoshida, W., Seymour, B., Shimizu, T., Hosomi, K., Saitoh, Y., Kasai, K., Kato, N., Takahashi, H., Okamoto, Y., Yamashita, O., Kawato, M., Imamizu, H., 2021. A multi-site, multi-disorder resting-state magnetic resonance image database. Sci Data 8, 227.

Thompson, P.M., Stein, J.L., Medland, S.E., Hibar, D.P., Vasquez, A.A., Renteria, M.E., Toro, R., Jahanshad, N., Schumann, G., Franke, B., Wright, M.J., Martin, N.G., Agartz, I., Alda, M., Alhusaini, S., Almasy, L., Almeida, J., Alpert, K., Andreasen, N.C., Andreassen, O.A., Apostolova, L.G., Appel, K., Armstrong, N.J., Aribisala, B., Bastin, M.E., Bauer, M., Bearden, C.E., Bergmann, O., Binder, E.B., Blangero, J., Bockholt, H.J., Boen, E., Bois, C., Boomsma, D.I., Booth, T., Bowman, I.J., Bralten, J., Brouwer, R.M., Brunner, H.G., Brohawn, D.G., Buckner, R.L., Buitelaar, J., Bulayeva, K., Bustillo, J.R., Calhoun, V.D., Cannon, D.M., Cantor, R.M., Carless, M.A., Caseras, X., Cavalleri, G.L., Chakravarty, M.M., Chang, K.D., Ching, C.R., Christoforou, A., Cichon, S., Clark, V.P., Conrod, P., Coppola, G., Crespo-Facorro, B., Curran, J.E., Czisch, M., Deary, I.J., de-Geus, E.J., den Braber, A., Delvecchio, G., Depondt, C., de Haan, L., de Zubicaray, G.I., Dima, D., Dimitrova, R., Djurovic, S., Dong, H., Donohoe, G., Duggirala, R., Dyer, T.D., Ehrlich, S., Ekman, C.J., Elvsashagen, T., Emsell, L., Erk, S., Espeseth, T., Fagerness, J., Fears, S., Fedko, I., Fernandez, G., Fisher, S.E., Foroud, T., Fox, P.T., Francks, C., Frangou, S., Frey, E.M., Frodl, T., Frouin, V., Garavan, H., Giddaluru, S., Glahn, D.C., Godlewska, B., Goldstein, R.Z., Gollub, R.L., Grabe, H.J., Grimm, O., Gruber, O., Guadalupe, T., Gur, R.E., Gur, R.C., Goring, H.H., Hagenaars, S., Hajek, T., Hall, G.B., Hall, J., Hardy, J., Hartman, C.A., Hass, J., Hatton, S.N., Haukvik, U.K., Hegenscheid, K., Heinz, A., Hickie, I.B., Ho, B.C., Hoehn, D., Hoekstra, P.J., Hollinshead, M., Holmes, A.J., Homuth, G., Hoogman, M., Hong, L.E., Hosten, N., Hottenga, J.J., Hulshoff-Pol, H.E., Hwang, K.S., Jack Jr., C.R., Jenkinson, M., Johnston, C., Jonsson, E.G., Kahn, R.S., Kasperaviciute, D., Kelly, S., Kim, S., Kochunov, P., Koenders, L., Kramer, B., Kwok, J.B., Lagopoulos, J., Laje, G., Landen, M., Landman, B.A., Lauriello, J., Lawrie, S.M., Lee, P.H., Le-Hellard, S., Lemaitre, H., Leonardo, C.D., Li, C.S., Liberg, B., Liewald, D.C., Liu, X., Lopez, L.M., Loth, E., Lourdusamy, A., Luciano, M., Macciardi, F., Machielsen, M.W., Macqueen, G.M., Malt, U.F., Mandl, R., Manoach, D.S., Martinot, J.L., Matarin, M., Mather, K.A., Mattheisen, M., Mattingsdal, M., Meyer-Lindenberg, A., McDonald, C., McIntosh, A.M., McMahon, F.J., McMahon, K.L., Meisenzahl, E., Melle, I., Milaneschi, Y., Mohnke, S., Montgomery, G.W., Morris, D.W., Moses, E.K., Mueller, B.A., Munoz Maniega, S., Muhleisen, T.W., Muller-Myhsok, B., Mwangi, B., Nauck, M., Nho, K., Nichols, T.E., Nilsson, L.G., Nugent, A.C., Nyberg, L., Olvera, R.L., Oosterlaan, J., Ophoff, R.A., Pandolfo, M., Papalampropoulou-Tsiridou, M., Papmeyer, M., Paus, T., Pausova, Z., Pearlson, G.D., Penninx, B.W., Peterson, C.P., Pfennig, A., Phillips, M., Pike, G.B., Poline, J.B., Potkin, S.G., Putz, B., Ramasamy, A., Rasmussen, J., Rietschel, M., Rijpkema, M., Risacher, S.L., Roffman, J.L., Roiz-Santianez, R., Romanczuk-Seiferth, N., Rose, E.J., Royle, N.A., Rujescu, D., Ryten, M., Sachdev, P.S., Salami, A., Satterthwaite, T.D., Savitz, J., Saykin, A.J., Scanlon, C., Schmaal, L., Schnack, H.G., Schork, A.J., Schulz, S.C., Schur, R., Seidman, L., Shen, L., Shoemaker, J.M., Simmons, A., Sisodiya, S.M., Smith, C., Smoller, J.W., Soares, J.C., Sponheim, S.R., Sprooten, E., Starr, J.M., Steen, V.M., Strakowski, S., Strike, L., Sussmann, J., Samann, P.G., Teumer, A., Toga, A.W., Tordesillas-Gutierrez, D., Trabzuni, D., Trost, S., Turner, J., Van den Heuvel, M., van der Wee, N.J., van Eijk, K., van Erp, T.G., van Haren, N.E., van-'t-Ent, D., van Tol, M.J., Valdes Hernandez, M.C., Veltman, D.J., Versace, A., Volzke, H., Walker, R., Walter, H., Wang, L., Wardlaw, J.M., Weale, M.E., Weiner, M.W., Wen, W., Westlye, L.T., Whalley, H.C., Whelan, C.D., White, T., Winkler, A.M., Wittfeld, K., Woldehawariat, G., Wolf, C., Zilles, D., Zwiers, M.P., Thalamuthu, A., Schofield, P.R., Freimer, N.B., Lawrence, N.S., Drevets, W.Alzheimer's Disease Neuroimaging Initiative, E.C.I.C.S.Y.S.G., 2014. The ENIGMA Consortium: large-scale collaborative analyses of neuroimaging and genetic data. Brain Imaging Behav 8, 153–182.

Tian, D., Zeng, Z., Sun, X., Tong, Q., Li, H., He, H., Gao, J.H., He, Y., Xia, M., 2022. A deep learning-based multisite neuroimage harmonization framework established with a traveling-subject dataset. Neuroimage 257, 119297.

Tong, Q., He, H., Gong, T., Li, C., Liang, P., Qian, T., Sun, Y., Ding, Q., Li, K., Zhong, J., 2020. Multicenter dataset of multi-shell diffusion MRI in healthy traveling adults with identical settings. Sci Data 7, 157.