NeuroImage 193 (2019) 35–45

Contents lists available at ScienceDirect

NeuroImage

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

Ten simple rules for predictive modeling of individual differences in neuroimaging

Dustin Scheinost [a,b,c,d,*], Stephanie Noble [d], Corey Horien [d], Abigail S. Greene [d], Evelyn MR. Lake [a], Mehraveh Salehi [e], Siyuan Gao [f], Xilin Shen [a], David O'Connor [f], Daniel S. Barron [g], Sarah W. Yip [c,g], Monica D. Rosenberg [h], R. Todd Constable [a,d,i]

[a] Department of Radiology and Biomedical Imaging, Yale School of Medicine, USA
[b] Department of Statistics and Data Science, Yale University, USA
[c] Child Study Center, Yale School of Medicine, USA
[d] Interdepartmental Neuroscience Program, Yale School of Medicine, USA
[e] Department of Electrical Engineering, Yale University, USA
[f] Department of Biomedical Engineering, Yale University, USA
[g] Department of Psychiatry, Yale School of Medicine, USA
[h] Department of Psychology, Yale University, USA
[i] Department of Neurosurgery, Yale School of Medicine, USA

ARTICLE INFO

Keywords:
Machine learning
Connectome
Classification
Cross-validation
Neural networks

ABSTRACT

Establishing brain-behavior associations that map brain organization to phenotypic measures and generalize to novel individuals remains a challenge in neuroimaging. Predictive modeling approaches that define and validate models with independent datasets offer a solution to this problem. While these methods can detect novel and generalizable brain-behavior associations, they can be daunting, which has limited their use by the wider connectivity community. Here, we offer practical advice and examples based on functional magnetic resonance imaging (fMRI) functional connectivity data for implementing these approaches. We hope these ten rules will increase the use of predictive models with neuroimaging data.

1. Introduction

A primary goal of neuroimaging research is to associate brain organization with individual phenotypes. Although thousands of research papers have modeled brain-behavior associations, these models tend to be *explanatory*. Unfortunately, because the goal of an explanatory analysis is to identify neuroimaging measures related to phenotypic measures, such analyses often do not generalize to novel individuals and have inadequate clinical utility (Rosenberg et al., 2018). To address this limitation, researchers are beginning to build *predictive* models that predict individual differences in phenotypes from neuroimaging data. Because models are defined and validated with independent data, they promise to improve our ability to uncover generalizable brain-behavior associations. Yet, like any method, predictive modeling has its own set of limitations and considerations that may be unfamiliar to the wider neuroimaging community.

We present "ten simple rules" for applying predictive modeling to brain connectivity data. These rules explain common issues aimed at both novice and experienced users of predictive models with the hope of encouraging more researchers to use these approaches. These rules are general and apply to most neuroimaging studies employing predictive modeling, independent of the exact algorithm used. Similarly, while the examples we provide are based on functional magnetic resonance imaging (fMRI) connectivity data, the same concepts apply to other types of data, such as task activation or structural connectivity data. The paper is organized as follows. First, we present a brief overview of a typical predictive modeling study. Table 1 lists definitions for key terms used throughout this manuscript. Next, we present ten rules for using predictive models, listed in Table 2 with key references, divided into three sections: **Validating predictive models with independent data: why and how?**, **Measuring model performance**, and **Accounting for confounds and interpreting results.** Finally, we offer some limitations and concluding remarks.

* Corresponding author. Yale School of Medicine, Radiology and Biomedical Imaging, Magnetic Resonance Research Center, New Haven, CT 06520, United States.
E-mail address: dustin.scheinost@yale.edu (D. Scheinost).

https://doi.org/10.1016/j.neuroimage.2019.02.057
Received 14 November 2018; Received in revised form 28 January 2019; Accepted 21 February 2019
Available online 1 March 2019

Table 1
Definitions of common terms.

Term	Definition
confound	Variable that affects the study variables and systemically differs across individuals
cross-validation	Methods of internal model validation in which a single dataset is divided into testing and training data several times
explanatory modeling	The generation and use of statistical models to test hypotheses about associations between observed data.
exploratory research	Research conducted at an early stage of inquiry to generate new hypotheses and establish a framework for future analyses
external validation	Testing predictive model performance on an independently collected dataset
false negative	Cases incorrectly classified by the model (e.g., patients incorrectly identified as non-patients)
false positive	Non-cases incorrectly classified by the model (e.g., non-patients incorrectly identified as patients)
generalizability	How well results from a sample population reflect the population at large
hyperparameters	Free parameters for an algorithm that need to be determined *a priori* or learned from the data and validated
interpretability	The ability of a researcher to understand a model and use it to better understand brain-behavior associations
model	An equation that maps a set of independent variables (*i.e.* neuroimaging data) to a set of dependent variables (*i.e.* phenotypic data)
multiple comparisons	Evaluating many hypotheses via statistical inferences simultaneously which may lead to observing a significant result simply by chance
nested cross-validation	A validation approach where, in each fold of a cross-validation, a second cross-validation is used to estimate a free parameter.
nuisance variable	Variables that are associated with study variables causing increased data variability, but have no pertinent neurobiological meaning to the study question.
overfitting	The tendency for statistical models to mistakenly fit sample-specific noise as if it were signal, leading to inflated effect size estimates
p-hacking	Selectively choosing data or analysis pipelines to obtain significant results
phenotype	An observable characteristics of an individual (e.g. behavior)
predictive modeling	The generation and use of statistical models to estimate new or future information. The term is synonymous with machine learning.
sensitivity	The proportion of true positives (see below for definition) relative to the number of cases (e.g. the number of patients correctly classified as patients divided the number of patients); also referred to as recall
specificity	The proportion of true negatives (see below for definition) relative to the number of non-cases (e.g. the number of non-patients correctly classified as non-patients divided by the number of non-patients).
testing data	Data used for model application and evaluation. The model built on the training data is applied to testing data for prediction.
training data	Data used to generate a statistical model that will be applied to a testing data set
true negative	Non-cases correctly classified by the model (e.g., non-patients correctly classified as non-patients)
true positive	Cases correctly classified by the model (e.g., patients correctly classified as patients)
underfitting	The failure of a model to capture the relationship between the response and predictor, generally due to inadequate model complexity, resulting in poor model performance in both training and testing data
validation	An unbiased evaluation of a model performance on data independent from the data used to generate the model

2. Conceptual overview: a predictive modeling study with fMRI data

In the context of neuroimaging, the goal of predictive modeling is typically to estimate a state or trait (phenotypic) characteristic of an individual from their neuroimaging data (e.g., a connectivity matrix). To do this, most studies follow the same template for analysis (see Fig. 1 for a workflow). First, neuroimaging data and associated phenotypic data (which may be binary, such as group membership, or continuous, such as IQ or symptom) from either a single or multiple datasets are separated into independent training data and testing data. Next, the training data are submitted to a predictive modeling algorithm. Using only the training data, the chosen algorithm selects the most relevant features from the data, and summarizes these features to produce a mathematical function, or model, that maps high dimensional neuroimaging data onto low dimensional phenotypic data. The model is then applied to previously unseen testing data to predict a phenotypic measure from each individual's neuroimaging data. Depending on the model validation strategy, the process may be repeated multiple times for different combinations of training and testing data. Finally, the model's performance is evaluated by comparing the predicted phenotypic measure against the actual phenotypic measure. In other words, the training data is used to define the model, while the testing data is used to evaluate its performance, or predictive power. Many different algorithms can be used to build predictive models from brain data, including support vector machines/regression (SVM/R), partial least squares regression (PLSR), neural networks, penalized regression (e.g., ridge, least absolute shrinkage and selection operator [LASSO], and elastic net), and random forests. Given the generality of the preceding template to the choice of algorithm—for simplicity—the ten rules below will be based on this template for predictive modeling.

Table 2
List of rules and key references.

Section	Number	Rule	Key references
Validating predictive models with independent data: why and how?	#1	Use out-of-sample prediction to generate more accurate and generalizable models	(Whelan and Garavan, 2014; Woo et al., 2017; Yarkoni and Westfall, 2017)
	#2	Keep training and testing data independent	(Gabrieli et al., 2015; Shmueli, 2010).
	#3	Use internal validation (i.e., cross-validation) as a practical solution for validating predictive models	(Breiman and Spector, 1992; Kohavi, 1995; Varoquaux et al., 2017)
	#4	Share data, code, and models to facilitate external validation and open science	(Milham, 2012; Nichols et al., 2017; Woo et al., 2017)
Measuring model performance	#5	Consider the question of interest when choosing a performance metric and properly assess statistical significance	(Alexander et al., 2015; Baldi et al., 2000)
	#6	Be mindful of sample characteristics	(Schnack and Kahn, 2016) (Barron et al., 2018).
	#7	Apply nested cross-validation or multiple comparisons correction when testing multiple models and parameters	(Varoquaux et al., 2017; Westfall et al., 1993)
Accounting for confounds and interpreting results	#8	Check to see if you are predicting what you think you are predicting	(Orban et al., 2018; Rao et al., 2017; Siegel et al., 2017)
	#9	Do not expect one model to fit all traits, states, or populations	(Rosenberg et al., 2018) (Greene et al., 2018)
	#10	Remember: interpretability matters	(Haufe et al., 2014; LeCun et al., 2015; Shen et al., 2017)

3. Validating predictive models with independent data: why and how?

A fundamental feature of predictive modeling—and one to which we will repeatedly return—is that models are trained in one sample and

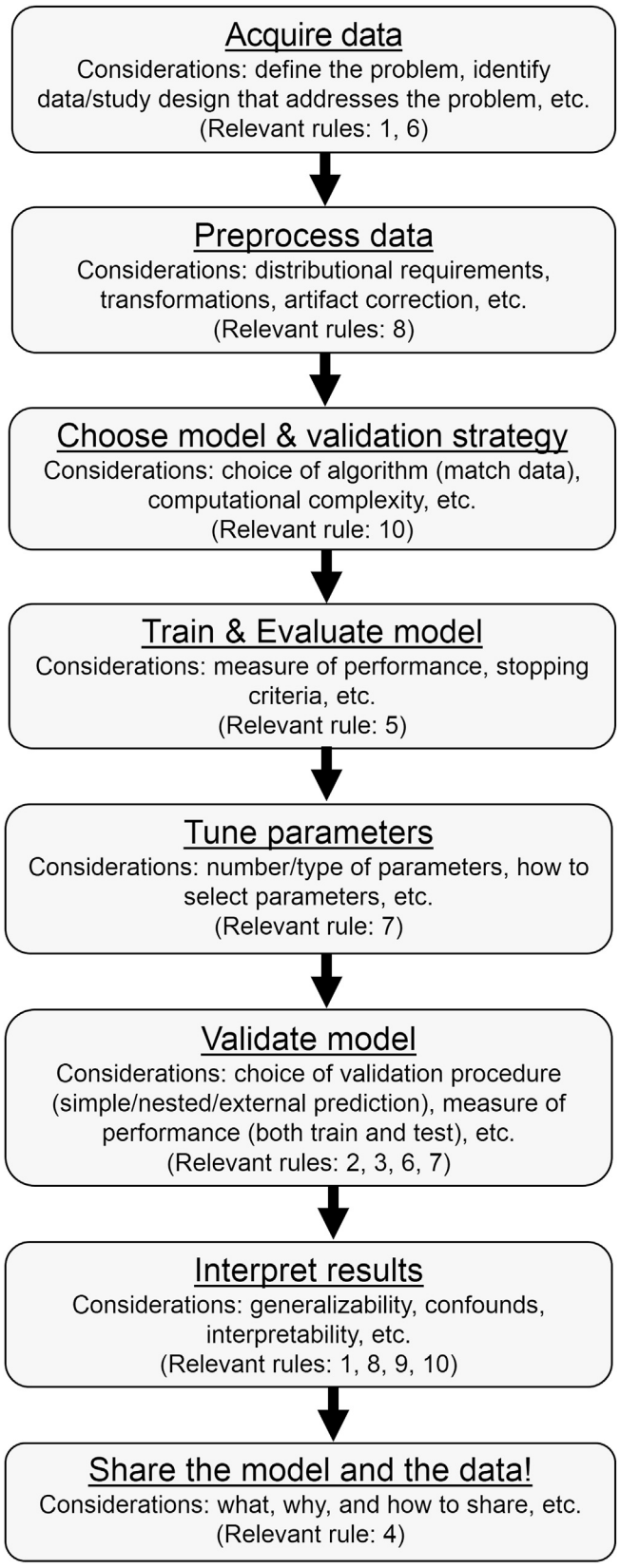

Fig. 1. General workflow for a predictive modeling study using neuroimaging data. Each box illustrates a different step in a typical study, along with relevant considerations. Pertinent rules discussed in the text are highlighted in each box as appropriate.

tested in another sample that was not used to build the model. In this section, we explore the need for, consequences of, and methods for keeping the training and testing data independent.

3.1. Rule #1: use out-of-sample prediction to generate more accurate and generalizable models

Explanatory models have populated the literature with exciting associations between brain and phenotypic measures, but these findings often do not generalize to other samples. In contrast, a common concern about predictive modeling in human neuroscience is that resulting models often seem to explain disappointingly little of the variance in the predicted measure, particularly when compared to results derived from explanatory models. Two primary causes of this phenomenon are overfitting and small sample sizes.

Overfitting, or "the tendency for statistical models to mistakenly fit sample-specific noise as if it were signal" (Yarkoni and Westfall, 2017), yields models that fail to generalize and that have high performance variance. By selecting for models that yield significant effects, effect size estimates for overfit models will be inflated. This is particularly relevant in human neuroimaging analyses, as the number of predictors (e.g., voxels, functional connections) is usually far greater than the number of observations (e.g., individuals; Whelan and Garavan, 2014). In contrast to explanatory models that use all available data to generate a model, predictive models attempt to detect overfitting by validating the model on novel individuals with strict separation of the training and testing data (see **Rules #2, #3, #4,** and **#7**). Nevertheless, overfitting can only be overcome by reducing the degrees of freedom for a model relative to the number of data points. In contrast, underfitting occurs when a model is not complex enough to capture the relationship between the neuroimaging and phenotypic data, generally resulting in poor model performance in both training and test data.

Second, as enthusiasm about "big data" (Poldrack and Gorgolewski, 2014) and data-driven analysis approaches has grown, sample sizes have increased while effect sizes have decreased. The use of large datasets has increased the power to detect associations between weak, widely distributed neural circuits and phenotypes, which should be expected to have small effect size estimates (Cremers et al., 2017). Large datasets also decrease the likelihood of overfitting. As larger samples are more representative of the general population, models from these samples are more likely to pick up on generalizable—rather than idiosyncratic—features, leading to decreased inflation of, and variance in, effect size estimates (Cremers et al., 2017; Whelan and Garavan, 2014; Yarkoni and Westfall, 2017). **Rule #6** further describes issues related to sample size and other characteristics for predictive models.

Despite critical differences between explanatory and predictive models, analysis results are often evaluated using the same metrics, making it easy to forget that they cannot be directly compared. For example, explained variance, often measured with coefficient of determination, is a common way to summarize results for both explanatory analyses and predictive models. In explanatory analyses, the coefficient of determination reflects the fit of a regression line to the dependent variable, while in predictive analyses, the coefficient of determination reflects how each observed value differs from its predicted value. As noted above, one would expect less variance to be explained by predictive models than by explanatory models, and this must be considered when using previous results to benchmark the performance of a predictive model. **Rule #5** provides further discussion on evaluating model performance.

In sum, models that fit sample idiosyncrasies or that are built in small, homogeneous datasets often demonstrate excellent explained variance and yield large effects, but in many cases fail to generalize (Woo et al., 2017). Prediction, by minimizing overfitting and discovering distributed effects from larger samples, leads to more generalizable models (see **Rules #8** and **#9** for reasons models might not generalize), and the robustness of predictive modeling to the specific features of a dataset

allows researchers to reveal fundamental brain-behavior associations (see **Rule #10** for an overview of model interpretability).

3.2. Rule #2: keep training and testing data independent

As introduced in **Rule #1**, predictive models are unique in that they are built to predict phenotypic data from previously unseen individuals—that is, data from new individuals, new time points, or both (Gabrieli et al., 2015; Shmueli, 2010; Yarkoni and Westfall, 2017). Because testing a model on new data is necessary for evaluating its generalizability, a critical rule of prediction is that previously unseen data must remain previously unseen. In other words, training and testing data—the observations used to define and evaluate a model—must be independent.

Several analysis choices during training and testing data separation can inadvertently compromise their independence. Consider, for example, the "curse of dimensionality" in human neuroimaging: features (e.g., voxels, functional connections) usually outnumber samples (e.g., participants, trials), increasing the risk of overfitting (see **Rule #1**) and complicating model interpretation (Mwangi et al., 2014; Varoquaux et al., 2017). The common and related practices of dimensionality reduction and feature selection seek to address this problem by reducing the number of potential features and identifying those most closely related to an outcome of interest (though in the extreme, these practices can also compromise model fit), but also introduce an opportunity for "peeking"—that is, failing to maintain independence of training and test data. As a concrete example, imagine a research group using functional connectivity data to predict clinical symptoms. Using data from all 100 participants, they perform independent component analysis (ICA) to identify functional networks and normalize the symptom severity scores. Using *K*-fold cross-validation (see **Rule #3**), they then train a model using connectivity measures from 80 individuals, and apply it to predict symptoms in the left-out 20. In this case, the training and testing data are not independent, because data from the left-out 20 influenced the ICA results and the symptom score normalization. In the context of this example, though potentially a computational burden to re-estimate all components, the feature selection and normalization steps should instead be performed *inside* the cross-validation loop (keeping the left-out 20 separate in each step). To identify similar methodological pitfalls, one can ask: had test data never been collected at all, would the model differ *in any way*? If the answer is yes, the test data have likely "contaminated" the model.

In the next two rules, we discuss strategies for avoiding such "peeking" at test data when building predictive models. These approaches fall into two overarching categories: internal validation (i.e., cross-validation; see **Rule #3**) and external validation (see **Rule #4**). Later, in **Rule #7**, we describe the importance of using nested cross-validation when tuning hyperparamters (*i.e.* free parameters for an algorithm) or trying multiple predictive modeling methods.

3.3. Rule #3: use internal validation (i.e., cross-validation) as a practical solution for validating predictive models

Internal validation, also known as cross-validation, refers to validation strategies that divide a single dataset into independent samples: training data used to build models and testing data held aside for testing a model's generalizability. Although testing a model in a separate, external dataset offers the strongest evidence of model generalization (see **Rule #4**), having an external dataset is often not practical. In most cases, cross-validation is a more feasible alternative. Nevertheless, care has to be taken when choosing an internal validation strategy.

Given the need for independent training and testing data (see **Rule #2**), several approaches exist to divide the original dataset into independent samples. Common methods include *K*-fold, leave-one-out, and split-half cross-validation. In *K*-fold cross-validation, the original dataset is randomly divided into K equally sized, non-overlapping subsets. Next, $K-1$ subsets are assigned as training data, with the remaining subset reserved as testing data. The assignment of subsets as training and testing data is then repeated K times, with each of the K subsets used exactly once as the testing data. Prediction performance from each combination of training and testing data can then be averaged to produce a single estimate of prediction performance. Both leave-one-out ($K =$ number of individuals) and split-half ($K = 2$) cross-validation can be viewed as special cases of *K*-fold cross-validation. A further distinction between the special case of leave-one-out cross-validation and the more general *K*-fold cross-validation is that leave-one-out cross-validation is exhaustive, while *K*-fold cross-validation is generally non-exhaustive. Exhaustive cross-validation strategies use every combination of training and testing data, whereas non-exhaustive strategies use only a limited combination of training and testing data. Though exhaustive strategies are feasible for leave-one-out cross-validation, as K becomes smaller, they can become computationally infeasible and, accordingly, are not generally used.

While *K*-fold cross-validation is the default cross-validation strategy in most cases, the choice of K affects prediction performance and must be performed with care. Increasing K will generally decrease bias—the difference between the predicted value and the true value—as more observations will be used for training. However, increasing K will increase variance—the sensitivity of the model to changes caused by different training data—as the predictive model has less data for training in each sample selection. In other words, one wants to have sufficiently large amount of training data to minimize bias (e.g. a large K), while retaining enough testing data to minimize variance (e.g. a low K). Due to this bias-variance tradeoff, the choice of K is dependent on the amount of available data. If the dataset is small, leave-one-out cross-validation (the largest possible choice of K) is generally used as more data are available for model training. However, leave-one-out cross-validation and small sample sizes can still lead to overfitting (Varoquaux et al., 2017). With larger sample sizes ($n > 200$), 5- or 10-fold cross-validation represents a good compromise between model bias and variance (Breiman and Spector, 1992; Kohavi, 1995; Pereira et al., 2009). An emerging standard seems to be to perform many iterations of 5-fold cross-validation (Varoquaux et al., 2017). Fig. 2 illustrates this bias-variance trade-off for several methods of validation. In summary, for most cases, the choice of K will scale inversely with sample size such that lower K values should be used with larger sample sizes.

3.4. Rule #4: share data, code, and models to facilitate external validation and open science

While cross-validation offers a practical way of validating a predictive model (see **Rule #3**), the best practice to maximize model generalizability is to use an independently collected, or external, dataset as testing data. This type of validation is called external validation. Using this approach, researchers can train a model—possibly preregistering it in a publication or other outlet—before collecting or downloading testing data, and/or share the model with other groups for independent validation. In such a situation it is impossible for the test data to affect the predictive model. More importantly, validating models in different datasets helps to ensure that the model is not fitting features that are dataset specific by making the testing data more heterogeneous and more representative of the general population (Woo et al., 2017).

While the need to collect two or more independent datasets has historically limited the use of external validation, the increasing availability of high-throughput neuroimaging samples and the emerging norm of data- and model-sharing in psychology and neuroscience are making external validation a routine option (Bzdok and Yeo, 2017; Milham, 2012; Nichols et al., 2017; Poldrack et al., 2013; Rosenberg et al., 2018). Best practices in open science suggest that, when sharing data, the code and, ideally, the trained and validated model itself should be shared as well.

Although the topic of sharing neuroimaging data and software has received much attention in recent years (Mennes et al., 2013; Nichols

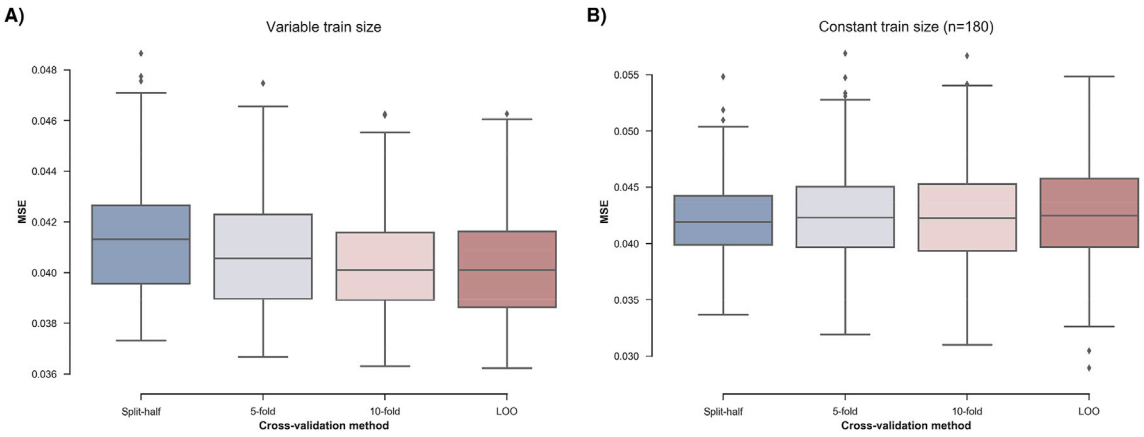

Fig. 2. Comparison of standardized MSE for different cross-validation methods for either A) variable training data size or B) constant training data size. A) Using 200 iterations of random sampling of 500 individuals from the Human Connectome (HCP) dataset, connectome-based predictive modeling (CPM) was applied to predict a measure of fluid intelligence (PMAT) with 4 different cross-validation strategies: split-half, 5-fold, 10-fold, and leave-one-out (LOO) cross-validation. For each strategy, the size of the training data was variable (i.e. the total sample was held constant) with split-half cross-validation using the least individuals for training (N = 250) and leave-one-out using the most individuals for training (N = 499). All cross-validation strategies give similar prediction performance with leave-out-one cross-validation performing the best due to the greater amount of training data. B) In contrast, when using 200 iterations of random sampling of individuals from the HCP dataset but keeping the number of individuals in training data constant (N = 180) (i.e. the total sample for each strategy was variable), leave-out-one cross-validation exhibited the largest variance in performance. Additionally, split-half cross-validation exhibited the smallest variance in performance. These data demonstrate the bias-variance tradeoff of different cross-validation strategies. See Supplemental Methods for further methodological details.

et al., 2017; Woo et al., 2017), the sharing of predictive models is not widely practiced. When sharing a model, two general approaches exist. The first approach is to share all aspects of a predictive model. On top of sharing the exact data and software used in the experiment, specific aspects of the predictive modeling pipeline should be shared, including, but not limited to, input features (*i.e.* connectivity and phenotypic data), predictive modeling methods, tested hyperparameter(s), and validation methods. This ensures that others can easily regenerate the model and test the modeling pipeline on their own data and against other predictive modeling methods.

However, in the case where the training methods are highly computationally intensive or sharing the data/software is not possible, the second approach is to share just the trained model—that is, the features used for prediction and how to combine them. When sharing trained models (as with sharing any other data), it is necessary to use standardized file and data formats. For example, scikit-learn and TensorFlow are popular machine learning toolkits for the programming language Python, which allows sharing of a trained model in the form of a NumPy file, a standard file type in Python. Other platforms provide similarly standardized formats such as ".mat" files in MATLAB, or ".json" files in JavaScript. Additionally, toolboxes exist to read these standards into other platforms and convert between different standards.

One drawback of sharing only a model is that the model is tied to the different processing choices made when it was generated. For example, the shared model may only work for the specific parcellation (or atlas) that was used for creating the connectivity matrices. If a different parcellation was used, then a new model would have to be generated. As such, carefully documenting and sharing all aspects of the modeling pipeline, when possible, is more useful than just sharing the model, as is sharing the raw data.

Finally, it is important to remember that if a model fails to generalize to a novel dataset, it may not be because the model is invalid. Different factors (e.g. multi-site data and interesting biological effects) may exist in different datasets (see **Rules #8** and **#9** for further discussion).

4. Measuring model performance

After a model is trained and applied to test data, its performance must be quantified and tested for significance. In this section, we present the most common methods to quantify model performance while accounting for different questions of interest, multiple comparisons, and sample characteristics.

4.1. Rule #5: consider the question of interest when choosing a performance metric and properly assess statistical significance

Model performance generally involves measuring the differences between observed (actual) and predicted (model generated) values and may be measured in several ways. As such, the metric of choice should be defined prior to the analysis to avoid multiple comparisons and will depend on the question of interest.

Measures of model performance can be unstandardized or standardized. Unstandardized measures provide a direct comparison between the predicted and observed outcomes and are presented in the units of the phenotypic measure. These measures are best for instances where the units and the spread of the phenotypic measure are well-known. For example, when predicting the due date of an expecting mother, unstandardized measures are readily interpretable. Time between predicted and actual date of delivery has intrinsic meaning and, given a known normative length of pregnancy, establishing acceptable levels of performance is straightforward. However, as the goal of predictive modeling with neuroimaging data is to discover brain-behavior associations, it is often of interest to use a model designed for one phenotypic measure to predict another measure to further test for specificity of the original brain-behavior association (Rosenberg et al., 2016a; Woo et al., 2017). If the units and scales between the different measures are not comparable, standardized measures—which remove units—may help to improve interpretability and comparisons of performance. Generally, there is a one-to-one mapping between unstandardized and standardized measures, which makes reporting both often trivial but recommended.

4.2. Rule #5A: evaluating continuous predictions

Continuous predictions are typically assessed by the squared or absolute difference between the predicted and observed values. Common unstandardized measures include mean squared error (MSE; $MSE(observed,\ predicted) = \frac{1}{n}\sum_{i=1}^{n}(observed_i - predicted_i)^2$), root mean squared error (RMSE; $RMSE(observed,\ predicted) = \sqrt{MSE(observed,}$

predicted)), and mean absolute error (MAE; $MAE(observed, predicted) = \frac{1}{n}\sum_{i=1}^{n}\left|observed_i - predicted_i\right|$. These measures can be normalized by the dispersion of the observed and/or predicted values to provide standardized measures. For example, MSE can be normalized by $MSE(observed, \mu)$ where μ is the mean of the observed variable, $MSE(observed, 0)$, or $MSE(observed, 0)$*$MSE(predicted, 0)$. MAE is often normalized by $\max(predicted) - \min(predicted)$. Whereas MSE represents the average squared difference between observed and predicted values, normalized MSE (NMSE; $NMSE(observed, predicted) = \frac{MSE(observed,predicted)}{MSE(observed,\mu)}$) represents the proportional improvement of the model over simply predicting the mean. Finally, normalizing the phenotypic measures (*i.e.* $observed_i^{'} = \frac{observed_i - \mu}{\sigma^2}$, where μ and σ^2 are the mean and variance of the phenotypic measures) will also produce standardized measures. However, if cross-validation is used, this normalization needs to be repeated in each fold with μ and σ^2 being estimated from the training data for that fold.

Many of these measures have analogous definitions based on regression models between the observed and predicted values. As mentioned in **Rule #1**, these definitions are not the same and are not readily comparable. For example, the R^2 from linear regression is $R^2 = 1 - \frac{MSE(observed,\hat{y})}{MSE(observed,\bar{\mu})}$, where $\hat{y}$ is the fitted value of *predicted* from linear regression. This definition of R^2 is the equivalent to the squared correlation between the predicted and observed values and reflects the error between the predicted values and its fit to the regression line, not the error between the predicted and observed values (Alexander et al., 2015; Dubois et al., 2018a). In contrast, prediction R^2 (also call cross-validation R^2 or q^2) is defined as $prediction\ R^2 = 1 - \frac{MSE(predicted,observed)}{MSE(observed,\bar{\mu})} = 1 - \text{NMSE}$. Note that, unlike R^2 for linear regression, prediction R^2 can be negative. In this case, the predictive model performs worse than simply guessing the mean of the phenotypic measure. Fig. 3 highlights the differences between the two versions of R^2. For assessing numerical accuracy, measures should be calculated directly from the predicted and observed values (e.g. prediction R^2), rather than from regression (e.g. correlation).

Nevertheless, as relative rankings, in addition to high numerical accuracy, are meaningful in establishing brain-behavior associations, correlation between predicted and observed values remains a valuable metric for evaluating prediction performance (as do other linear regression approaches). For the case of perfect correspondence between predicted and observed values, linear regression will fit a line with an intercept of 0 and a slope of 1. An intercept other than 0 indicates that the model adds a constant to every predicted value. A slope other than 1 indicates that the model will over/under predict values at the tails of the phenotypic measure. When reporting correlation between predicted and observed measures, it is recommended to report another metric that directly compares the predicted and observed values, such as prediction R^2 (Alexander et al., 2015).

4.3. Rule #5B: evaluating categorical predictions

Evaluating categorical performance involves different considerations. Categorical predictions are evaluated based on the amount of correct and incorrect predictions, typically given as a proportion. The most common measure is overall accuracy, or the proportion of correct classifications across observations (Baldi et al., 2000; Yarkoni and Westfall, 2017). However, because overall accuracy may not translate to accuracy for individual classes (e.g., cases and controls), accuracy for each class is often reported separately. Class-specific accuracy can be measured with sensitivity and specificity, two of several measures that broadly reflect rates of true positives, false positives, true negatives, or false negatives (Baldi et al., 2000). Sensitivity reflects the true positive rate

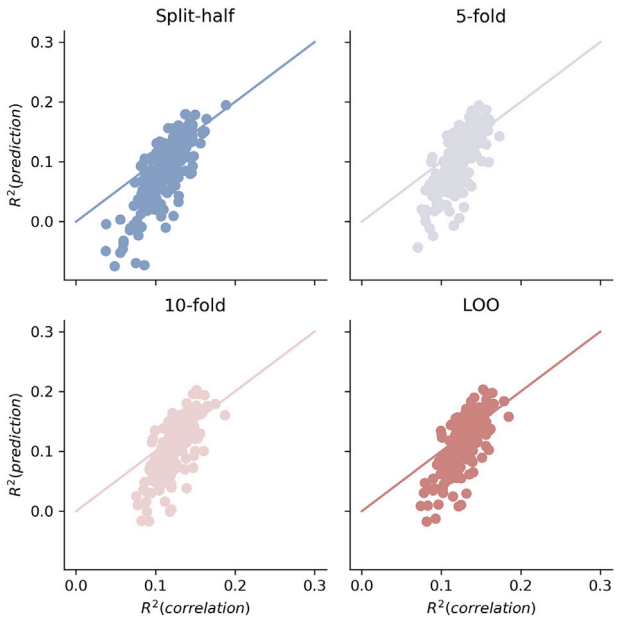

Fig. 3. Comparison of prediction R^2 calculated directly from comparing observed and predicted values and explanatory R^2 calculated from linear regression. Using 200 iterations of 400 individual for training and 400 individuals for testing randomly selected from the HCP dataset, CPM was used to predict PMAT using split-half, 5-fold, 10-fold, and leave-one-out (LOO) cross-validation. Each point represents the same CPM model evaluated with prediction R^2 (on the y-axis) and explanatory R^2 (on the x-axis). Prediction R^2 was calculated as 1 minus normalized mean squared error between the observed and predicted values (see **Rule #5**), while explanatory R^2 was calculated as the square of the Pearson correlation between the observed and predicted values. For all cross-validation strategies, R^2 from linear regression over-estimates performance when compared to R^2 calculated directly from comparing observed and predicted values. This bias is the greatest at lower prediction performance and reduces for better predicting models. The line in each plot represents the y = x line. See Supplemental Methods for further methodological details.

($true\ positives/(true\ positives + false\ negatives)$; also called recall), or percent of correctly identified cases. Specificity reflects the true negative rate ($true\ negative/(true\ negative + false\ positives)$), or percent of correctly identified controls. Although in principle, the 'best' model would result in high detection of true effects and rejection of false effects, one measure may in practice be prioritized over the others. A model with a high sensitivity but low specificity may be acceptable for preliminary screening prior to additional testing. Conversely, a model with a low sensitivity but high specificity may be more appropriate for assigning individuals to a high-risk intervention.

The relationship between sensitivity and specificity of a given classification model is characterized using a receiver operating characteristic (ROC) curve (Bradley, 1997), in which the true positive rate (sensitivity) is plotted on the y axis and the false positive rate (1-specificity) is plotted on the x axis across a range of decision thresholds (i.e., classification boundaries). In this context, overall model performance may be quantified as the area under the curve (AUC), or the portion of the area of the unit space falling below the ROC curve (Fawcett, 2006). The resultant scalar value will range between 0 and 1, with 0.5 indicating performance at the level of random chance and higher values indicating better model performance (Fawcett, 2006) (for details on calculating AUC, see (Fawcett, 2006)). Based on the ROC curve, it is also possible to determine an optimal 'cut point' or the point at which the greatest number of individuals are correctly characterized (Unal, 2017). However, note that multiple methods of determining the cut point exist (e.g., Youden index (J) method, concordance probability method) (Unal, 2017) and that the

criteria for determining these should be determined *a priori* using indpendent or simulated data to avoid over-fitting.

Important considerations in evaluating overall accuracy, sensitivity, and specificity arise when group sizes are unequal. First, chance model performance for individual classes may be more or less than 50% if groups have unequal sample sizes. For example, a model that performs at chance in a sample of 70 cases and 30 controls, randomly guessing that 70% of individuals are cases based on base rates, will be 70% accurate for cases (i.e., 70% sensitive) but 50% accurate overall. Furthermore, a model that predicts that everyone in this sample is a patient will be 100% correct for patients, but 0% correct for controls and 70% correct overall. Thus, especially when group sizes are unequal, it is beneficial to report complementary measures of accuracy that depend on case prevalence, such as positive predictive value ($true\ positive/(true\ positives + false\ positives)$; also called precision), negative predictive value ($true\ negative/(true\ negative + false\ negative)$), or the F-score ($2 \times \frac{precision \times recall}{precision + recall}$, also called F_1 score). For additional considerations related to unequal sample sizes, see **Rule #6**.

4.4. Rule #5C: assessing significance

For both continuous and categorical performance measure, statistical significance is best evaluated using permutation testing when using cross-validation. Results from each fold of the cross-validation are not independent and the degrees of freedom for parametric statistics will be over-estimated. Permutation testing involves repeatedly shuffling labels on the data points (*i.e.* randomly assigning one individual's phenotypic measures to another individual's imaging data), re-running the predictive modeling algorithm on the randomly shuffled data, calculating prediction performance of the model based on the shuffled data, and creating a null distribution of the prediction performance measure. From this null distribution, a one-sided p-value is calculated as the proportion of permutations where prediction performance is either 1) greater than or equal to the prediction performance of the original, un-shuffled data if a lower metric indicates better performance or 2) less than or equal to the prediction performance of the original, un-shuffled data if a higher metric indicates better performance. When using external validation with an independent dataset, p-values based on parametric statistics, such as chi-square or correlation, remain valid.

4.5. Rule #6: Be mindful of sample characteristics

Predictive models are only as good as the data that is used to train and validate them. As such, it is important to be mindful of how sample characteristics can influence the final model. Common sample characteristics include sample size, the distribution of the phenotypic measures, and unequal group sizes (e.g. when patients are rare relative to controls).

Increasing the sample size for training and testing will help to create more generalizable models and less variable prediction results (Cui and Gong, 2018; Schnack and Kahn, 2016; Varoquaux et al., 2017). Nevertheless, different objectives and modeling approaches invariably require different sample sizes for satisfactory performance. While models that generalize to external datasets can be generated with relatively few individuals (e.g. N = 25 for (Rosenberg et al., 2016a; Rosenberg et al., 2016b)), most questions and methods will require datasets of over 100 individuals. Fig. 4 shows how performance changes as a function of sample size for an example situation.

Continuous phenotypic measures can have very different distributions, which can affect results in unanticipated ways. As such, understanding how a phenotype varies across the sample is an important step. Many predictive algorithms assume that variables follow a specific distribution, e.g. a normal distribution. Therefore, before training a model, the distribution of each predicted phenotype should be evaluated and, if necessary, transformed to follow the expected distribution. For example, the Fisher transform is often used in functional connectivity analyses to convert correlation coefficients to an approximately normal distribution. In addition to accounting for the distribution, a sufficient range of measurements for a given phenotypic measure is needed to model individual differences. In specific cases—such as a phenotypic measure wherein healthy individuals perform at a maximum level—values might again need to be transformed to provide a meaningful range to permit individual variation and, therefore, prediction. An ideal model would be constructed from an even distribution of scores across the entire range of a phenotype.

Datasets with unbalanced groups are common and present problems when performing cross-validation and assessing prediction performance. For example, consider a study of 100 patients with 70 randomized to a placebo and 30 to a drug group. The larger placebo group could bias the classifier towards predicting placebo over drug. This classification bias would not be readily detected in the testing data because subsampling this data would also result in unequal numbers of each group. One

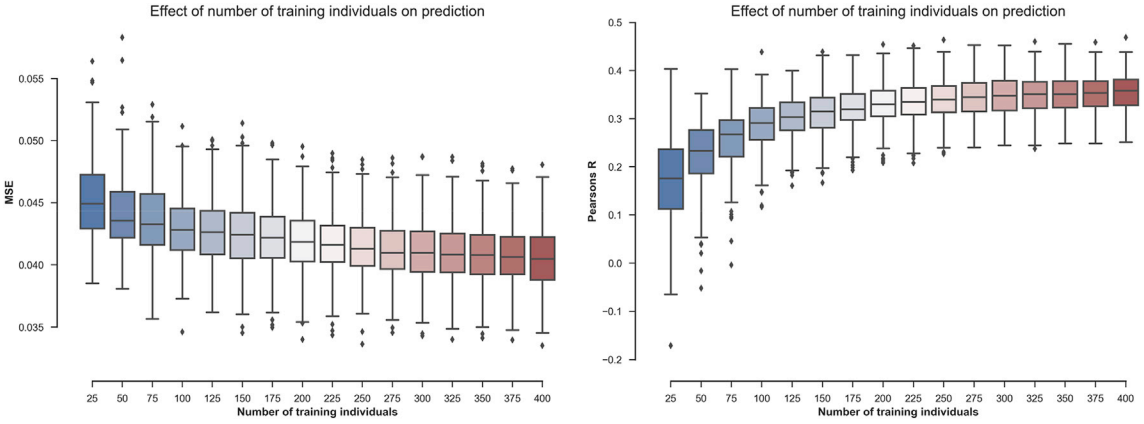

Fig. 4. Comparison of prediction performance as a function of the number of individuals in the training data. Using 200 iterations of 400 individuals for training and 400 individuals for testing randomly selected from the HCP dataset, CPM was used to predict PMAT using a variable number of individuals in the training data, starting with 25 individual up to 400 individual in steps of 25 individuals. Each CPM model was then evaluated on the same 400 test subjects, for each iteration. Increasing the number of individuals in the training data increased the performance of the CPM model with performance beginning to plateau with >200 individuals for training. The panel on the left shows model performance evaluated with standardized MSE. The panel on the right shows model performance evaluated with Pearson's correlation. See Supplemental Methods for further methodological details.

strategy for overcoming this classification bias is to 'down-weight' or penalize classifications for the larger group (in this case, the placebo group) (Thai-Nghe et al., 2010). Another strategy to overcome this is to balance the training and testing data to have equal group memberships (Barron et al., 2018). This can be done by randomly sampling equal numbers of drug and placebo groups to form the training data (under-sampling the majority class and/or oversampling the minority class using duplicate or synthetic observations). In addition, selecting equal numbers of drug and placebo groups in the testing data can avoid this problem. Because the dataset as a whole is unbalanced, this would necessarily leave some individuals from the larger placebo group unused in the classification analysis; here, iterating and randomly selecting balanced groups can make better use of the entire dataset. This example also illustrates a common pitfall known as the Accuracy Paradox. Suppose a predictive model performs at 66% accuracy by correctly identifying 20 out of 30 individuals in the testing data. In this unbalanced scenario, this accuracy would increase to 70% simply by identifying every individual in the testing data as placebo. As discussed in **Rule #5**, reporting other measures that better handle unbalanced groups (e.g., the F Score) in addition to total accuracy is often useful. Even if classes are balanced, comparing different models (i.e. by using different classification algorithms) with reference to accuracy alone can present a comparable problem because improvements to the model are not necessarily reflected by increased accuracy.

4.6. Rules #7: apply nested cross-validation or multiple comparisons correction when testing multiple models and parameters

Even if models are validated using the previously discussed approaches (see **Rules #3** and **#4**), further measures must be taken when testing multiple models to ensure that a model does not perform well simply by chance. In a given study a large number of models and hyperparameters (e.g., regularization parameters that adjust model sparsity/complexity by controlling feature selection and weighting; see **Rule #2**), may be considered and correction for these multiple tests must be taken into account. For example, in an exploratory study, a researcher may want to test two different predictive modeling strategies for predicting variable "X" in a dataset. Each strategy may require two hyperparameters, and they may decide to test 20 values for each. They may then wish to test two different preprocessing approaches. Finally, they might explore whether the imaging data can also predict variables "Y" and "Z". In sum, this researcher will have conducted 480 tests (2 strategies x 2 hyperparameters per strategy x 20 hyperparameter values x 2 preprocessing approaches x 3 outcome variables). Although many of these tests are not independent, it becomes increasingly likely that one model may appear to make accurate predictions by chance.

Thoughtful dimensionality reduction (see **Rule #2**) can mitigate this problem by decreasing the number of tests performed. For example, rather than testing the capacity of a given model to predict many individual variables, one can derive a single summary variable (e.g., via factor analysis), which may have the added benefit of improving the measurement quality of the predicted variable (Dubois et al., 2018b).

If an experiment includes multiple hyperparameters or models, it is prudent to perform nested cross-validation to ultimately validate only a single model (Varoquaux et al., 2017). In nested cross-validation, the original data is randomly divided into three subsets: training, validation, and testing data. First, all models are trained with training data and cross-validated with validation data, as described in **Rule #3**. Data for training and validation are often repeatedly reassigned, and the best-performing model on average is selected for final testing. These two steps are referred to as the inner loop. Finally, the model selected from the inner loop is validated in the outer loop with the testing data. It is important to keep data used in the inner and outer loops independent (see **Rule #2**). Individuals may be reassigned to subsets repeatedly to assess the robustness of model selection and performance. However, reassignment may introduce dependence between tests.

Standard corrections for multiple comparisons may be used instead of, or in addition to, nested cross-validation. Common methods include Bonferroni correction and False Discovery Rate (FDR) correction. Bonferroni correction limits the rate of at least one false positive in a family of tests (Westfall et al., 1993) and is often considered overly conservative. FDR correction, on the other hand, balances the expected proportion of false positives given a total number of positives.

More generally, best practices appropriate to the stage of the discovery (i.e., confirmatory or exploratory) alleviate many multiple testing concerns (Wagenmakers et al., 2012). A confirmatory design involving *a priori* restriction of the number of models tested reduces the probability of obtaining a false positive test by chance. That said, exploratory analyses are acceptable when many models are potentially of interest. Nevertheless, exploratory approaches must be accompanied by clear documentation of the exploratory nature of the analyses and cautious interpretations (e.g., hypothesis-generating rather than hypothesis-confirming language) to avoid "*p*-hacking" (Yarkoni and Westfall, 2017).

5. Accounting for confounds and interpreting results

In the final section, we will discuss important factors affecting model performance including confounds and nuisance variables, biologically-meaningful variables, and model interpretability.

5.1. Rule #8: check to see if you are predicting what you think you are predicting

Navigating confounds and nuisance variables in predictive modeling—although sometimes overlooked or downplayed—is crucial to ensuring that a predictive model captures meaningful individual differences in neural circuitry (Rao et al., 2017; Siegel et al., 2017). A confound is an extraneous variable that affects the neuroimaging and/or phenotypic data and systemically differs across individuals in the sample or overall population (Rao et al., 2017). As discussed in **Rule #6**, even if confounds do not exist in the sample as a whole, they may exist in particular folds of cross-validation. In contrast, nuisance variables are extraneous variable that have an association with the neuroimaging and/or phenotypic data, but have no pertinent neurobiological meaning to the study question. Unlike confounds, nuisance variables tend to increase the variability within the data, instead of systematically differing between subsets of the data (Sanderman et al., 2006). To check for these effects, researchers should correlate known or suspected confounds and nuisance variables with the neuroimaging data, the phenotypic measure, and the predicted values in the training and testing data.

The first step to minimizing the effects of confounds and nuisance variables is at the data collection and preprocessing stage. Random sampling and best practices in data collection can reduce the likelihood of collecting data with confounds (Sanderman et al., 2006). Additionally, state-of-the-art preprocessing strategies can remove many confound and nuisance variables that cannot be accounted for in data collection, such as physiological and head motion artifacts (Murphy et al., 2013; Power et al., 2012; Siegel et al., 2017; Tagliazucchi and Laufs, 2014). However, these effects very frequently persist and need to be accounted for when generating predictive models.

When generating predictive models, suspected confounds and nuisance variables can be accounted for in a number of ways. One strategy is to exclude individuals with outlying values (e.g. large head motion, or individuals on medication). However, this approach can bias the sample reducing the generalizability of the model, and runs the risk of misestimating the effect of interest (Rao et al., 2017). Alternatively, partial correlation methods that regress confounds and nuisance variables out of the input data can be used to select features (Hsu et al., 2018). Also, the final model may include additional terms to capture any prediction variance attributed to confounding or nuisance variables effects. As stated in **Rule #2**, any regression of, or covarying for,

confounding or nuisance variables needs to be performed independently for the testing and training data if cross-validation is being used. Otherwise, controlling for confounding or nuisance variables across the entire dataset will compromise the independence of the testing and training data and is a common error when trying to remove the effects of confounding or nuisance variables. As noted in (Linn et al., 2016), the regression approach may not be robust when considering multiple variables jointly. In these cases, methods that differentially weight individual data points based on propensity scores—the probability of a data point being assigned to a particular group given a set of confounding or nuisance variables—may offer more robust control of external factors. Approaches like this are beginning to be developed for use in neuroimaging (Linn et al., 2016). In general, trying different ways to account for confounds and nuisance variables is best explored using a nested cross-validation approach or left as an exploratory analysis to avoid problems with multiple comparisons (see **Rule #7**).

Finally, as open-science has made large, multi-site datasets more broadly available (see **Rule #4**), site effects in predictive models represent a common confound (Dansereau et al., 2017; Orban et al., 2018). Harmonizing data acquisition for both neuroimaging and phenotypic data in a multi-site study can minimize, but may not fully eliminate, site effects (Abraham et al., 2017; Noble et al., 2017). Other site differences can include systematic missing data or different proportions of patients versus controls. For example, some sites may not collect or release certain information for logistical reasons (e.g. lack of facilities/resources or human subject concerns). In all of these cases, several approaches exist to address site effects, such as including site as a fixed or random effect in a model (Dansereau et al., 2017) and using performance metrics that take into account the unbalanced nature of the sample (see **Rule #6**). Perhaps the simplest and most powerful procedure is to use cross-validation approaches where all the data from one or more sites are left out for later use as testing data.

Overall, care to control for confounds and nuisance variables improves the generalizability of predictive models (Orban et al., 2018). Yet, as discussed next in **Rule #9**, nuisance variables may be of neurobiological interest in some contexts and can be a topic for further investigation.

5.2. Rule #9: do not expect one model to fit all traits, states, or populations

As mentioned in **Rule #8**, what might be considered a nuisance variable in one study may in fact be a phenotype of interest in another study. For example, sex and age differences have been reported in functional connectivity (Andrews-Hanna et al., 2007; Biswal et al., 2010; Dosenbach et al., 2010; Geerligs et al., 2014; Greene et al., 2018; Kilpatrick et al., 2006; Satterthwaite et al., 2015; Scheinost et al., 2015; Tomasi and Volkow, 2012). The underlying predictive features may be different for different populations as defined by these factors, thus yielding different models. That is, sex or age differences may require different models to predict the same phenotypic measure, reflecting group differences in underlying neural circuitry. Such disparities between models do not invalidate any model. Rather, as the goal of predictive modeling is to establish brain-behavior associations, they highlight opportunities for better resolving these associations by carefully considering variables that could influence the model. For instance, if model differences across development are observed, further work could involve grouping participants based on age and evaluating when models succeed and fail, in order to illuminate developmental changes influencing the link between the brain and phenotypic measures (Rosenberg et al., 2018).

Possible model differences due to cultural/ethnic factors also must be taken into account. In the context of pain-imaging research, for example, culture, ethnicity, and various psychological factors contribute to individual variability in the experience of pain (Borsook and Kalso, 2013; Bushnell et al., 2013; Davis et al., 2017), and such factors can present a challenge when attempting to predict a measure of interest. Hence, it is important that researchers consider these variables.

In addition, the influence of brain state is emerging as a critical consideration when generating predictive models. In this sense, *state* can refer to brain differences in short-time-scale fluctuations (dynamics of fMRI), associated with executing different tasks (such as a cognitive task versus rest), or in physiological state (such as sleep or due to a drug). It has been shown previously that by having individuals complete tasks in the scanner, participant-specific connectivity features are more readily observable (Finn et al., 2017; Vanderwal et al., 2017) and lead to better model performance when predicting fluid intelligence (Greene et al., 2018) and measures of attention (Rosenberg et al., 2016a). In effect, tasks can serve as a means to enhance differences in connectivity between individuals above and beyond those detected at rest (Finn et al., 2017; Vanderwal et al., 2017): hence, using tasks to evoke specific brain states might allow higher predictive capacity of the phenotypic measure of interest (Greene et al., 2018; Rosenberg et al., 2016a). This might be especially relevant when generating models for patients with a psychiatric disease (Finn et al., 2017): a task could be used to probe specific circuits underlying a given phenotypic measure that might go undetected at rest, facilitating the generation of effective predictive models—similar to how a cardiac stress test might uncover abnormal heart rhythms that are otherwise unnoticed when a patient is not exercising (Finn et al., 2017). Thus, by carefully considering brain states, as well as other differences of interest, researchers have the potential to better understand the association between the brain and phenotypic measures.

5.3. Rule #10: remember: interpretability matters

One final challenge in predictive modeling is "model interpretability". One of the ultimate goals in generating predictive models from neuroimaging data is to identify the underlying associations between the brain and phenotypic measures. Although predictive models with larger and more complex features may capture more subtle brain-behavior associations and yield better performance, they can be more difficult to map onto the brain (*i.e.* harder to interpret). In contrast, models that are more easily described—even if their performance is worse—may ultimately be more useful for understanding brain-behavior associations. Therefore, model complexity and interpretability needs to be taken into consideration.

Models that are easier to interpret have two main advantages over more complex models. First, these models may offer more neurobiological insights. Models that provide a one-to-one mapping back to the imaging data allow for easier visualization and investigation of the underlying brain features that contribute to the model (Shen et al., 2017). Second, models that are easier to understand are also easier to modify, improve, or fix. For example, models derived by techniques such as deep learning are very difficult to interpret and relate back to tangible features in the brain (LeCun et al., 2015). Likewise, these techniques require significant expertise to improve in a principled manner (LeCun et al., 2015).

As in most cases, the final trade-off between model complexity and interpretability depends on the scientific application. A clinical treatment, such as transcranial magnetic stimulation (Fox et al., 2012), requires interpretable models to define the target; whereas other applications, such as brain–computer interfaces (Ruiz et al., 2014; Weiskopf et al., 2004), may pursue the best prediction performance regardless of model interpretability and, in this case, treating the model as a black box is fine.

When using cross-validation, it is often difficult to interpret the underlying neurobiology associated with feature weights, as the selected features for a model may vary across different folds. Typically, one of two approaches is used. The first approach is to combine the features across every fold of cross-validation. For example, if linear methods are used, these features can be averaged across every fold. Similarly, features that only appear in every fold or a large majority of folds (e.g. 90% of the folds) can be retained for interpretation as in (Rosenberg et al., 2016a).

The second approach is to re-run the modeling algorithm on all the data after hyperparameters have been chosen through cross-validation. In this approach, the goal is not to assess model performance. Rather, the features are used only to make interpretations about the underlying neurobiological meaning of the model. When using all the data to create a final set of features, overfitting is likely, as this analysis is inherently explanatory, and hyperparameters may have already been trained. Hence, model performance should only be assessed using external, independent dataset as described in **Rule #2**.

Many studies that focus on interpretability recommend linear classifiers or regression approaches (Grosenick et al., 2013). These studies rely on interpreting the weights of the models; yet, linearity alone does not guarantee interpretability (Grosenick et al., 2013). For example, linear methods could yield unstable coefficients if the input variables are highly correlated (Hastie et al., 2001) (*i.e.* the variance is arbitrarily split between the two covariates that are correlated). In addition, most linear classifiers return a dense set of feature weights (Grosenick et al., 2013). In other words, every input variable is assigned a non-zero weight needed for prediction. Setting variables with low weights to zero using penalized approaches (i.e. LASSO) can help to create a sparser set of features and aid interpretability. Finally, a large linear weight on a feature does not equate to greater importance of that feature for prediction (Haufe et al., 2014).

To assess the interpretability of features from a complex model, there are methods that employ post-hoc analyses to vary aspects of the model and evaluate how these variations affect model performance (Baehrens et al., 2010; Hansen et al., 2011; Ribeiro et al., 2016). A simple approach to this would be to iterate through all features, removing a single feature at each iteration, in order to quantify the changes in prediction performance associated with each feature. Features that degrade performance the most upon removal would then be assigned greater importance in the model. Note that because these approaches are used to assess the importance of features in existing models, they are considered post hoc, and they are not subject to the multiple comparisons correction discussed in **Rule #6**.

6. Limitations

While the goal of this work is to promote the use and understanding of predictive models in neuroimaging, we would also suggest that traditional explanatory models are needed. For a discussion of the pros of explanatory models, we point the interested reader to a series of comments and replies (Friston, 2012, 2013; Lindquist et al., 2013; Reiss, 2015) on this topic.

Predictive models based on neuroimaging data will only ever account for a fraction of the variance. Neuroimaging studies are limited by how much information the signal can capture about the measure of interest. At the same time, these studies are also limited by the chosen phenotypic measure used. While the success of a model is evaluated by how well it predicts a phenotypic measure (and these phenotypic measures have to be treated as gold standards), it is well known that such measures are not always the ground truth but themselves suffer from confounds and noise. When studying brain-behavior associations, one must keep in mind how extraordinary it is that neuroimaging data can be distilled to approximate phenotypic measures that reflect a simplification of multiple complex features. Thus, even modest results are reasonable and remarkable. For a discussion on the reliability of phenotypic measures in the context of predictive modeling, we point the interested reader to: (Dubois et al., 2018a, 2018b; Gignac and Bates, 2017).

Finally, most of these rules are covered in greater detail in classic machine learning/predictive modeling textbooks. For further reading, we point to the interested reader to following textbooks (Hastie et al., 2001; James et al., 2014; Wasserman, 2006, 2010):

7. Conclusion

Predictive modeling can be an intimidating. Yet, such models provide a powerful means to establish the link between the brain and phenotypic measures, revealing novel associations that other methods cannot provide. As such, we hope these ten rules will assist neuroimaging researchers to interpret the growing literature and to adopt these approaches into their analysis toolboxes. Ultimately, such work will help advance our understanding of the brain.

Acknowledgements

Data were provided in part by the Human Connectome Project, WU-Minn Consortium (Principal Investigators: David Van Essen and Kamil Ugurbil; 1U54MH091657) funded by the 16 NIH Institutes and Centers that support the NIH Blueprint for Neuroscience Research; and by the McDonnell Center for Systems Neuroscience at Washington University.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.neuroimage.2019.02.057.

References

Abraham, A., Milham, M.P., Di Martino, A., Craddock, R.C., Samaras, D., Thirion, B., Varoquaux, G., 2017. Deriving reproducible biomarkers from multi-site resting-state data: an Autism-based example. Neuroimage 147, 736–745.

Alexander, D.L., Tropsha, A., Winkler, D.A., 2015. Beware of R(2): simple, unambiguous assessment of the prediction accuracy of QSAR and QSPR models. J. Chem. Inf. Model. 55, 1316–1322.

Andrews-Hanna, J.R., Snyder, A.Z., Vincent, J.L., Lustig, C., Head, D., Raichle, M.E., Buckner, R.L., 2007. Disruption of large-scale brain systems in advanced aging. Neuron 56, 924–935.

Baehrens, D., Schroeter, T., Harmeling, S., Kawanabe, M., Hansen, K., M, K.-R., 2010. How to explain individual classification decisions. #252, ller J. Mach. Learn. Res. 11, 1803–1831.

Baldi, P., Brunak, S., Chauvin, Y., Andersen, C.A., Nielsen, H., 2000. Assessing the accuracy of prediction algorithms for classification: an overview. Bioinformatics 16, 412–424.

Barron, D.S., Salehi, M., Browning, M., Harmer, C.J., Constable, R.T., Duff, E., 2018. Exploring the prediction of emotional valence and pharmacologic effect across fMRI studies of antidepressants. Neuroimage: Clinical 20, 407–414.

Biswal, B.B., Mennes, M., Zuo, X.N., Gohel, S., Kelly, C., Smith, S.M., Beckmann, C.F., Adelstein, J.S., Buckner, R.L., Colcombe, S., Dogonowski, A.M., Ernst, M., Fair, D., Hampson, M., Hoptman, M.J., Hyde, J.S., Kiviniemi, V.J., Kötter, R., Li, S.J., Lin, C.P., Lowe, M.J., Mackay, C., Madden, D.J., Madsen, K.H., Margulies, D.S., Mayberg, H.S., McMahon, K., Monk, C.S., Mostofsky, S.H., Nagel, B.J., Pekar, J.J., Peltier, S.J., Petersen, S.E., Riedl, V., Rombouts, S.A., Rypma, B., Schlaggar, B.L., Schmidt, S., Seidler, R.D., Siegle, G.J., Sorg, C., Teng, G.J., Veijola, J., Villringer, A., Walter, M., Wang, L., Weng, X.C., Whitfield-Gabrieli, S., Williamson, P., Windischberger, C., Zang, Y.F., Zhang, H.Y., Castellanos, F.X., Milham, M.P., 2010. Toward discovery science of human brain function. Proc. Natl. Acad. Sci. U. S. A. 107, 4734–4739.

Borsook, D., Kalso, E., 2013. Transforming pain medicine: adapting to science and society. Eur. J. Pain 17, 1109–1125.

Bradley, A.P., 1997. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recogn. 30, 1145–1159.

Breiman, L., Spector, P., 1992. Submodel selection and evaluation in regression. The X-random case. Int. Stat. Rev./Rev. Int. Stat. 60, 291–319.

Bushnell, M.C., Ceko, M., Low, L.A., 2013. Cognitive and emotional control of pain and its disruption in chronic pain. Nat. Rev. Neurosci. 14, 502–511.

Bzdok, D., Yeo, B.T.T., 2017. Inference in the age of big data: future perspectives on neuroscience. Neuroimage 155, 549–564.

Cremers, H.R., Wager, T.D., Yarkoni, T., 2017. The relation between statistical power and inference in fMRI. PLoS One 12, e0184923.

Cui, Z., Gong, G., 2018. The effect of machine learning regression algorithms and sample size on individualized behavioral prediction with functional connectivity features. Neuroimage 178, 622–637.

Dansereau, C., Benhajali, Y., Risterucci, C., Pich, E.M., Orban, P., Arnold, D., Bellec, P., 2017. Statistical power and prediction accuracy in multisite resting-state fMRI connectivity. Neuroimage 149, 220–232.

Davis, K.D., Flor, H., Greely, H.T., Iannetti, G.D., Mackey, S., Ploner, M., Pustilnik, A., Tracey, I., Treede, R.D., Wager, T.D., 2017. Brain imaging tests for chronic pain: medical, legal and ethical issues and recommendations. Nat. Rev. Neurol. 13, 624–638.

Dosenbach, N.U., Nardos, B., Cohen, A.L., Fair, D.A., Power, J.D., Church, J.A., Nelson, S.M., Wig, G.S., Vogel, A.C., Lessov-Schlaggar, C.N., Barnes, K.A., Dubis, J.W.,

Feczko, E., Coalson, R.S., Pruett, J.R., Barch, D.M., Petersen, S.E., Schlaggar, B.L., 2010. Prediction of individual brain maturity using fMRI. Science 329, 1358–1361.
Dubois, J., Galdi, P., Han, Y., Paul, L.K., Adolphs, R., 2018a. Resting-state functional brain connectivity best predicts the personality dimension of openness to experience. Personal Neurosci. 1.
Dubois, J., Galdi, P., Paul, L.K., Adolphs, R., 2018b. A distributed brain network predicts general intelligence from resting-state human neuroimaging data. Philos. Trans. R. Soc. Lond. B Biol. Sci. 373.
Fawcett, T., 2006. An introduction to ROC analysis. Pattern Recogn. Lett. 27, 861–874.
Finn, E.S., Scheinost, D., Finn, D.M., Shen, X., Papademetris, X., Constable, R.T., 2017, Oct 15. Can brain state be manipulated to emphasize individual differences in functional connectivity? Neuroimage 160, 140–151.
Fox, M.D., Buckner, R.L., White, M.P., Greicius, M.D., Pascual-Leone, A., 2012. Efficacy of transcranial magnetic stimulation targets for depression is related to intrinsic functional connectivity with the subgenual cingulate. Biol. Psychiatry 72, 595–603.
Friston, K., 2012. Ten ironic rules for non-statistical reviewers. Neuroimage 61, 1300–1310.
Friston, K., 2013. Sample size and the fallacies of classical inference. Neuroimage 81, 503–504.
Gabrieli, J.D., Ghosh, S.S., Whitfield-Gabrieli, S., 2015. Prediction as a humanitarian and pragmatic contribution from human cognitive neuroscience. Neuron 85, 11–26.
Geerligs, L., Renken, R.J., Saliasi, E., Maurits, N.M., Lorist, M.M., 2014, Jul. A brain-wide study of age-related changes in functional connectivity. Cerebr. Cortex 25 (7), 1987–1999.
Gignac, G.E., Bates, T.C., 2017. Brain volume and intelligence: the moderating role of intelligence measurement quality. Intelligence 64, 18–29.
Greene, A.S., Gao, S., Scheinost, D., Constable, R.T., 2018. Task-induced brain state manipulation improves prediction of individual traits. Nat. Commun. 9, 2807.
Grosenick, L., Klingenberg, B., Katovich, K., Knutson, B., Taylor, J.E., 2013. Interpretable whole-brain prediction analysis with GraphNet. Neuroimage 72, 304–321.
Hansen, K., Baehrens, D., Schroeter, T., Rupp, M., Müller, K.R., 2011. Visual interpretation of kernel-based prediction models. Mol. Inform. 30, 817–826.
Hastie, T., Tibshirani, R., Friedman, J., 2001. The Elements of Statistical Learning. Springer New York Inc., New York, NY, USA.
Haufe, S., Meinecke, F., Görgen, K., Dähne, S., Haynes, J.D., Blankertz, B., Bießmann, F., 2014. On the interpretation of weight vectors of linear models in multivariate neuroimaging. Neuroimage 87, 96–110.
Hsu, W.T., Rosenberg, M.D., Scheinost, D., Constable, R.T., Chun, M.M., 2018. Resting-state functional connectivity predicts neuroticism and extraversion in novel individuals. Soc. Cognit. Affect Neurosci. 13, 224–232.
James, G., Witten, D., Hastie, T., Tibshirani, R., 2014. An Introduction to Statistical Learning: with Applications in R. Springer Publishing Company, Incorporated.
Kilpatrick, L.A., Zald, D.H., Pardo, J.V., Cahill, L.F., 2006. Sex-related differences in amygdala functional connectivity during resting conditions. Neuroimage 30, 452–461.
Kohavi, R., 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: Proceedings of the 14th International Joint Conference on Artificial Intelligence, vol. 2. Morgan Kaufmann Publishers Inc., Montreal, Quebec, Canada, pp. 1137–1143.
LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521, 436–444.
Lindquist, M.A., Caffo, B., Crainiceanu, C., 2013. Ironing out the statistical wrinkles in "ten ironic rules". Neuroimage 81, 499–502.
Linn, K.A., Gaonkar, B., Doshi, J., Davatzikos, C., Shinohara, R.T., 2016. Addressing confounding in predictive models with an application to neuroimaging. Int. J. Biostat. 12, 31–44.
Mennes, M., Biswal, B.B., Castellanos, F.X., Milham, M.P., 2013. Making data sharing work: the FCP/INDI experience. Neuroimage 82, 683–691.
Milham, M.P., 2012. Open neuroscience solutions for the connectome-wide association era. Neuron 73, 214–218.
Murphy, K., Birn, R.M., Bandettini, P.A., 2013. Resting-state fMRI confounds and cleanup. Neuroimage 80, 349–359.
Mwangi, B., Tian, T.S., Soares, J.C., 2014. A review of feature reduction techniques in neuroimaging. Neuroinformatics 12, 229–244.
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., 2017. Best practices in data analysis and sharing in neuroimaging using MRI. Nat. Neurosci. 20, 299–303.
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., 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.
Orban, P., Dansereau, C., Desbois, L., Mongeau-Pérusse, V., Giguère, C., Nguyen, H., Mendrek, A., Stip, E., Bellec, P., 2018. Multisite generalizability of schizophrenia diagnosis classification based on functional brain connectivity. Schizophr. Res. 192, 167–171.
Pereira, F., Mitchell, T., Botvinick, M., 2009. Machine learning classifiers and fMRI: a tutorial overview. Neuroimage 45, S199–S209.
Poldrack, R.A., Barch, D.M., Mitchell, J.P., Wager, T.D., Wagner, A.D., Devlin, J.T., Cumba, C., Koyejo, O., Milham, M.P., 2013. Toward open sharing of task-based fMRI data: the OpenfMRI project. Front. Neuroinf. 7, 12.
Poldrack, R.A., Gorgolewski, K.J., 2014. Making big data open: data sharing in neuroimaging. Nat. Neurosci. 17, 1510–1517.
Power, J.D., Barnes, K.A., Snyder, A.Z., Schlaggar, B.L., Petersen, S.E., 2012. Spurious but systematic correlations in functional connectivity MRI networks arise from subject motion. Neuroimage 59, 2142–2154.
Rao, A., Monteiro, J.M., Mourao-Miranda, J., Initiative, A.s.D., 2017. Predictive modelling using neuroimaging data in the presence of confounds. Neuroimage 150, 23–49.
Reiss, P.T., 2015. Cross-validation and hypothesis testing in neuroimaging: an irenic comment on the exchange between Friston and Lindquist et al. Neuroimage 116, 248–254.
Ribeiro, M.T., Singh, S., Guestrin, C., 2016. Why should I trust you?. In: Explaining the Predictions of Any Classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, San Francisco, California, USA, pp. 1135–1144.
Rosenberg, M.D., Casey, B.J., Holmes, A.J., 2018. Prediction complements explanation in understanding the developing brain. Nat. Commun. 9, 589.
Rosenberg, M.D., Finn, E.S., Scheinost, D., Papademetris, X., Shen, X., Constable, R.T., Chun, M.M., 2016a. A neuromarker of sustained attention from whole-brain functional connectivity. Nat. Neurosci. 19, 165–171.
Rosenberg, M.D., Zhang, S., Hsu, W.T., Scheinost, D., Finn, E.S., Shen, X., Constable, R.T., Li, C.S., Chun, M.M., 2016b. Methylphenidate modulates functional network connectivity to enhance attention. J. Neurosci. 36, 9547–9557.
Ruiz, S., Buyukturkoglu, K., Rana, M., Birbaumer, N., Sitaram, R., 2014. Real-time fMRI brain computer interfaces: self-regulation of single brain regions to networks. Biol. Psychol. 95, 4–20.
Sanderman, R., Coyne, J.C., Ranchor, A.V., 2006. Age: nuisance variable to be eliminated with statistical control or important concern? Patient Educ. Counsel. 61, 315–316.
Satterthwaite, T.D., Wolf, D.H., Roalf, D.R., Ruparel, K., Erus, G., Vandekar, S., Gennatas, E.D., Elliott, M.A., Smith, A., Hakonarson, H., Verma, R., Davatzikos, C., Gur, R.E., Gur, R.C., 2015. Linked sex differences in cognition and functional connectivity in youth. Cerebr. Cortex 25, 2383–2394.
Scheinost, D., Finn, E.S., Tokoglu, F., Shen, X., Papademetris, X., Hampson, M., Constable, R.T., 2015. Sex differences in normal age trajectories of functional brain networks. Hum. Brain Mapp. 36, 1524–1535.
Schnack, H.G., Kahn, R.S., 2016. Detecting neuroimaging biomarkers for psychiatric disorders: sample size matters. Front. Psychiatry 7, 50.
Shen, X., Finn, E.S., Scheinost, D., Rosenberg, M.D., Chun, M.M., Papademetris, X., Constable, R.T., 2017. Using connectome-based predictive modeling to predict individual behavior from brain connectivity. Nat. Protoc. 12, 506–518.
Shmueli, G., 2010. To explain or to predict? Stat. Sci. 25, 289–310.
Siegel, J.S., Mitra, A., Laumann, T.O., Seitzman, B.A., Raichle, M., Corbetta, M., Snyder, A.Z., 2017. Data quality influences observed links between functional connectivity and behavior. Cerebr. Cortex 27, 4492–4502.
Tagliazucchi, E., Laufs, H., 2014. Decoding wakefulness levels from typical fMRI resting-state data reveals reliable drifts between wakefulness and sleep. Neuron 82, 695–708.
Thai-Nghe, N., Gantner, Z., Schmidt-Thieme, L., 2010. Cost-sensitive learning methods for imbalanced data. In: The 2010 International Joint Conference on Neural Networks (IJCNN), pp. 1–8.
Tomasi, D., Volkow, N.D., 2012. Aging and functional brain networks. Mol. Psychiatr. 17 (471), 549-458.
Unal, I., 2017. Defining an optimal cut-point value in ROC analysis: an alternative approach. Comput. Math. Method. Med. 2017. Article ID 3762651, 14 pages, 2017.
Vanderwal, T., Eilbott, J., Finn, E.S., Craddock, R.C., Turnbull, A., Castellanos, F.X., 2017. Individual differences in functional connectivity during naturalistic viewing conditions. Neuroimage 157, 521–530.
Varoquaux, G., Raamana, P.R., Engemann, D.A., Hoyos-Idrobo, A., Schwartz, Y., Thirion, B., 2017. Assessing and tuning brain decoders: cross-validation, caveats, and guidelines. Neuroimage 145, 166–179.
Wagenmakers, E.J., Wetzels, R., Borsboom, D., van der Maas, H.L., Kievit, R.A., 2012. An agenda for purely confirmatory research. Perspect. Psychol. Sci. 7, 632–638.
Wasserman, L., 2006. All of Nonparametric Statistics (Springer Texts in Statistics). Springer-Verlag.
Wasserman, L., 2010. All of Statistics: A Concise Course in Statistical Inference. Springer Publishing Company, Incorporated.
Weiskopf, N., Mathiak, K., Bock, S.W., Scharnowski, F., Veit, R., Grodd, W., Goebel, R., Birbaumer, N., 2004. Principles of a brain-computer interface (BCI) based on real-time functional magnetic resonance imaging (fMRI). Biomed. Eng. IEEE Trans. 51, 966–970.
Westfall, P.H., Young, S.S., Wright, S.P., 1993. On adjusting P-values for multiplicity. Biometrics 49, 941–945.
Whelan, R., Garavan, H., 2014. When optimism hurts: inflated predictions in psychiatric neuroimaging. Biol. Psychiatry 75, 746–748.
Woo, C.W., Chang, L.J., Lindquist, M.A., Wager, T.D., 2017. Building better biomarkers: brain models in translational neuroimaging. Nat. Neurosci. 20, 365–377.
Yarkoni, T., Westfall, J., 2017. Choosing prediction over explanation in psychology: lessons from machine learning. Perspect. Psychol. Sci. 12, 1100–1122.