Review

Biological Psychiatry: CNNI

Machine Learning With Neuroimaging: Evaluating Its Applications in Psychiatry

Ashley N. Nielsen, Deanna M. Barch, Steven E. Petersen, Bradley L. Schlaggar, and Deanna J. Greene

ABSTRACT

Psychiatric disorders are complex, involving heterogeneous symptomatology and neurobiology that rarely involves the disruption of single, isolated brain structures. In an attempt to better describe and understand the complexities of psychiatric disorders, investigators have increasingly applied multivariate pattern classification approaches to neuroimaging data and in particular supervised machine learning methods. However, supervised machine learning approaches also come with unique challenges and trade-offs, requiring additional study design and interpretation considerations. The goal of this review is to provide a set of best practices for evaluating machine learning applications to psychiatric disorders. We discuss how to evaluate two common efforts: 1) making predictions that have the potential to aid in diagnosis, prognosis, and treatment and 2) interrogating the complex neurophysiological mechanisms underlying psychopathology. We focus here on machine learning as applied to functional connectivity with magnetic resonance imaging, as an example to ground discussion. We argue that for machine learning classification to have translational utility for individual-level predictions, investigators must ensure that the classification is clinically informative, independent of confounding variables, and appropriately assessed for both performance and generalizability. We contend that shedding light on the complex mechanisms underlying psychiatric disorders will require consideration of the unique utility, interpretability, and reliability of the neuroimaging features (e.g., regions, networks, connections) identified from machine learning approaches. Finally, we discuss how the rise of large, multisite, publicly available datasets may contribute to the utility of machine learning approaches in psychiatry.

Keywords: Computational psychiatry, Feature selection, Functional connectivity, Machine learning, Neurophysiological mechanisms, Prediction

https://doi.org/10.1016/j.bpsc.2019.11.007

Psychiatric disorders are complex in their clinical phenomenology, with patient profiles often involving heterogeneous symptoms that change or fluctuate in severity over time. The neurophysiology underlying psychopathology parallels this complexity, as brain disorders rarely involve only single brain structures (1,2). To better describe and understand the complexities of psychiatric disorders, machine learning and other pattern classification approaches have become increasingly used to harness the rich information observed with human neuroimaging (3–7). Machine learning capitalizes on multivariate data, detecting complex patterns in the brain that may identify abnormalities present in psychiatric disorders. Broadly, these types of approaches can be categorized into supervised or unsupervised learning strategies, with supervised learning using known attributes of individuals to identify relevant brain patterns, and unsupervised learning using coherent brain patterns to generate novel attributes or subgroups of patients.

While machine learning holds promise as a tool for studying psychiatric disorders, these approaches also come with unique challenges and trade-offs, requiring additional considerations (8–12). In this review, we discuss the importance of evaluating the application of machine learning to psychiatric disorders, particularly focusing on supervised learning approaches. Specifically, we discuss issues that can arise when using supervised machine learning to 1) make predictions about individuals and 2) uncover the mechanisms underlying psychopathology. In part 1, we discuss best practices for making individual-level predictions about patients with machine learning, providing guidelines for the use of clinically informative training labels, appropriate assessment of classification performance and generalizability, and avoidance or benchmarking of confounding variables. In part 2, we discuss best practices for making inferences about the mechanisms underlying psychiatric disorders using machine learning, providing guidelines for evaluating the unique utility, interpretability, and reliability of a set of features. This review is intended to highlight important considerations for interpreting machine learning results in psychiatry, as an experimenter, reviewer, or critical reader of the literature.

To ground these points, we discuss specific examples applying supervised machine learning to functional connectivity magnetic resonance imaging (MRI) data in particular. Functional connectivity MRI measures the temporal correlation between the intrinsic [or sometimes task-evoked (13–15)]

ISSN: 2451-9022

functional MRI activity of pairs of regions (16), yielding a rich characterization of the brain's functional network architecture (17–19). Regions comprising a "functional network" are linked by strong, positive correlations at rest (e.g., default mode, visual) (17,18), termed functional connectivity (16). While our examples focus on functional connectivity, note that most points of consideration discussed here also apply to machine learning studies using other neuroimaging measures to understand psychiatric disorders.

Supervised machine learning identifies relationships between multivariate features (e.g., functional connections) and subject labels (e.g., patient vs. healthy control subject) using a learning algorithm (e.g., support vector machines). When applied to psychiatry, training labels are often different diagnoses (e.g., Tourette syndrome vs. healthy, tic-free) but can also be different states within a patient (e.g., depressed vs. remitted) or different task conditions (e.g., viewing happy vs. fearful faces). A number of supervised learning algorithms (e.g., k-nearest neighbor, support vector machines, decision trees) combine information across features in different ways. Algorithm selection depends on many factors including the research question, type of data, and nature of the training data (20). In general, machine learning procedures involve training (i.e., feature selection, feature weight optimization, and cross-validation) and testing (i.e., model performance, model generalizability). The patterns of features that best classify individuals in the training set according to their labels are weighted and combined in a resulting classifier that can be applied to a distinct set of individuals in the testing phase. While "classifier" often implies a binary model (e.g., patient or healthy control subject), here we use "classifier" to describe any trained multivariate model (i.e., binary, categorical, or continuous). For a review of the general procedures involved in supervised machine learning strategies and their applications to neuroimaging data, see Figure 1 and other reviews (3,21–23).

PART 1: EVALUATING PREDICTION WITH MACHINE LEARNING IN PSYCHIATRIC DISORDERS

Machine learning is well poised to address a major goal in psychiatry: making predictions about individual patients. For example, will a given child go on to develop a psychiatric disorder? Will treatment A or B work better for this individual patient? In most cases, this level of clinical utility has yet to be reached. Barriers to clinically useful, neuroimaging-based classifiers include classifier predictions that do not go beyond known information (e.g., current diagnosis), ambiguous metrics of classifier performance, poor model generalizability to future datasets, and predictions that are correlated

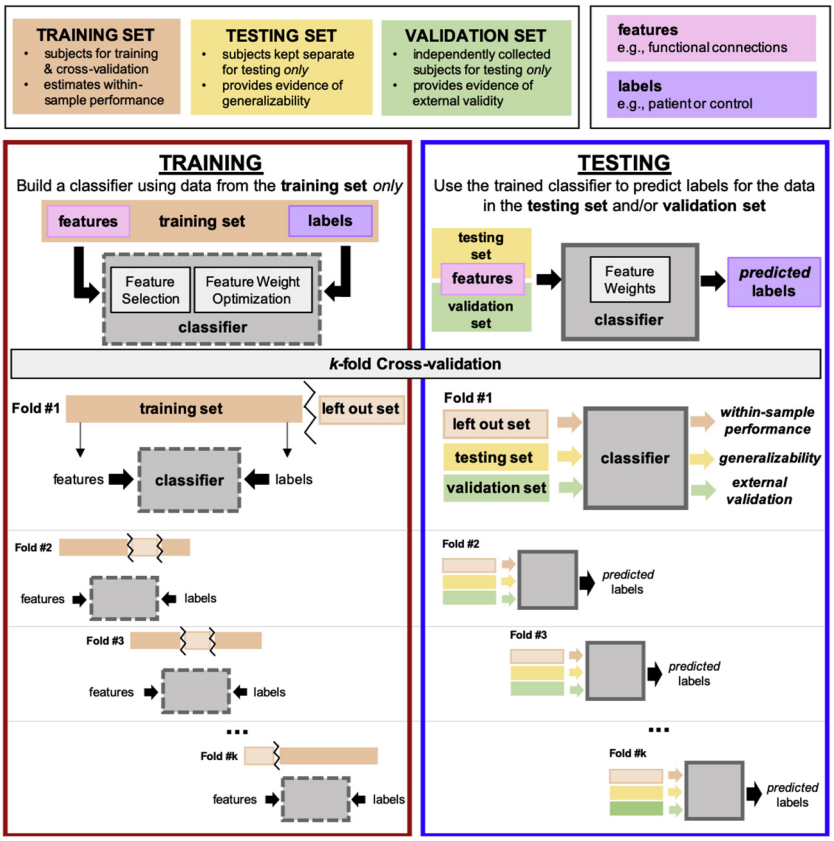

Figure 1. Overview of machine learning components and procedures. Different datasets serve distinct purposes in machine learning approaches. The training set is the group of individuals used to build a classifier. The testing set is a group of individuals collected at the same time as the training set but kept completely separate from training. Reporting performance in a testing set can provide evidence of classifier generalizability. The validation set is a group of individuals collected separately from the training set (e.g., different site, different scanner, separate study). Reporting performance in a validation set provides additional evidence of classifier validity. Each of these datasets comprises features and labels. Features are the multivariate data that, in aggregate, are used to build and make predictions with a classifier. Labels are binary, categorical, or continuous characteristics of individuals that are used to train a classifier and are subsequently predicted. Machine learning procedures involve training (red box) and testing (blue box). Training identifies relationships between multivariate features (e.g., functional connections) and subject labels (e.g., patient vs. healthy control subject) using a learning algorithm (e.g., support vector machines). The patterns of features that best classify individuals in the training set are then weighted and combined in a resulting classifier. Training can also involve feature selection, data- or hypothesis-driven selection of a reduced set of features. Training procedures should only be performed in the training set (and separately across folds of cross-validation). Testing involves applying the trained classifier to new individuals never used in training. Commonly, classifiers are assessed using k-fold cross-validation. For each fold, a portion of individuals is left out of the training set (left-out set), and a classifier is built using the remaining individuals in the training set. The trained classifier is then used to classify the left-out set of individuals and, if available, the independent testing and validation sets. Cross-validation can assess whether the performance and feature weights of a classifier depend upon which individuals are in the training set.

with confounding variables. Below, we discuss these challenges as well as recommendations to facilitate research aimed at developing clinically relevant multivariate classifiers.

Clinically Informative Training Labels

Early work applying supervised machine learning approaches to study psychiatric disorders has provided a solid proof of concept, predicting known patient characteristics (e.g., diagnosis, symptom severity) from concurrent neuroimaging data. For example, we and others have demonstrated that functional connectivity can be used to successfully classify individuals by diagnosis (i.e., patient vs. healthy control subject) in a number of psychiatric disorders, including attention-deficit/hyperactivity disorder (24), autism (25), depression (26), schizophrenia (27,28), and Tourette syndrome (29,30). Additionally, predicting age or other normative characteristics of individuals may be clinically informative. In the case of age, many psychiatric disorders have developmental origins (31,32) and are often interpreted in terms of developmental progress [e.g., brain immaturity (33,34)]. Such interpretations have led to research questions aimed at understanding deviations from normative developmental trajectories as risk factors for psychopathology. Thus, predicting the developmental status of patients may illuminate features about the disorder, especially when it involves atypical brain development. Several studies of typical development have shown that patterns of functional connectivity can successfully predict an individual's age (35–41). Capitalizing on these results, we found that a classifier trained to predict age in typical development can elucidate atypical development in children and adults with Tourette syndrome (30). Furthermore, brain age predicted from functional connectivity under different contexts (rest, implicit positive/negative emotion task) has been linked to an individual's risky behavior outside of the scanner (40).

Of course, classification of known characteristics, such as a person's diagnostic status, alone does not provide clinical utility. Therefore, there is a growing amount of literature on using machine learning methods with prospective imaging studies, in which neuroimaging data are collected before the emergence of distinguishing behaviors or symptoms (e.g., before treatment outcome or clinical diagnosis) to determine whether patterns of neuroimaging features can predict subsequent diagnosis, prognosis, or treatment efficacy. Several studies that have prospectively collected functional connectivity data have been able to predict future psychiatric outcomes (25,42,43). For example, functional connectivity was successfully used to classify which 6-month-old infants at high risk for developing autism were subsequently diagnosed with autism at 24 months (25) and to predict which individuals seeking treatment for substance abuse subsequently completed an intensive rehabilitation program (42).

Another important issue is that diagnosis alone may not fully capture the heterogeneity in psychiatric disorders, potentially leading to clinically less informative classifiers. Thus, categorical or continuous training labels that encompass subgroups or dimensionality of psychiatric symptoms [e.g., Research Domain Criteria (44)] may yield classifiers that better represent underlying symptomatology (45). Furthermore, unsupervised learning strategies may provide additional clinical utility by identifying novel subgroups of patients with categorically different patterns of neuroimaging features. These subgroups may exhibit different treatment outcomes or symptom trajectories. For example, subtypes of depression generated using unsupervised machine learning and functional connectivity were subsequently able to predict responsiveness to transcranial magnetic stimulation therapy (46).

Performance and Generalizability

Classifier success is typically assessed by testing how well a classifier can predict the labels of a set of individuals never used for training, either across folds of cross-validation or in an independent testing set (Figure 1). For binary or other categorical classifiers, total accuracy (percentage of patients and healthy control subjects correctly labeled) is often reported but may not sufficiently convey a classifier's performance. Classifier bias (e.g., classifying all individuals as patients) and imbalanced training sets (e.g., 75% patients, 25% healthy control subjects) can obfuscate whether a binary classifier (e.g., patient vs. healthy control subjects) that accurately classifies more than 50% of individuals in the test set is actually performing better than chance. Nonparametric tests like permutation testing (i.e., randomizing the labels in the training set) can establish an appropriate null for evaluating whether a classifier performs better than chance.

For regression models, metrics of success include the numerical accuracy of predictions (e.g., mean squared error) and the relational accuracy of predictions (e.g., R^2). If interested in the accurate prediction of a specific individual (e.g., identifying an individual with vulnerability to psychopathology), metrics that quantify the numerical accuracy of predictions should be used. Alternatively, if interested in the prediction of the variance in the sample (e.g., determining whether variance in treatment response is represented in neuroimaging data), metrics that quantify relational accuracy are sufficient. As in assessing binary or categorical classification accuracy, regression models can be assessed with permutation testing to establish an appropriate null for the amount of error expected by chance.

Beyond evaluating performance, it is critical to determine how well a trained classifier can generalize to data from novel subjects. Assessment of performance through cross-validation is a reasonable first step, yet alone it is not sufficient to demonstrate generalizability. Importantly, to avoid inflation of performance metrics, all procedures used to train the classifier, such as feature selection, model selection, and parameter optimization, should be conducted only in the training set and separately across folds of cross-validation. For most studies it is feasible to provide evidence of generalizability by setting aside a group of subjects for testing at the start, building a classifier with data in the training set, and then reporting performance of the classifier in the testing set. Ultimately, external validation with an independently collected validation set is best.

Poor generalizability, when a trained (and often published) classifier does not accurately classify new subjects, might arise for a variety of reasons. First, cross-validation (as just discussed) can be prone to poor generalizability. For example, leave-one-out cross-validation, in which a single individual is left out of the training set in each fold, has been shown to

produce less reliable estimates of classification accuracy than k-fold cross-validation, in which a percentage of individuals (e.g., 10%) are left out of the training set in each fold (47). Second, the quality of the data used in training can affect the resulting classifier; unreliable neuroimaging data might contribute to poor generalizability. For example, the reliability of functional connectivity is impacted by artifact correction, subject arousal, and the amount of functional MRI data collected (48–50), and thus these factors may also affect the reliability of a classifier. Third, a classifier will theoretically generalize best when trained with precise and ecologically valid labels. Because diagnosis alone may not fully capture the heterogeneity in psychiatric disorders, training labels that encompass the dimensionality of psychiatric symptoms as well as other ecological factors like comorbid diagnoses or medication use may also improve the ultimate performance of a classifier in a real-world setting (2,51). Finally, the number of subjects used for training can affect the resulting classifier (52,53), although there is currently no prescriptive sample size for generalizable performance.

In some unique cases, poor generalizability can be informative of the nature of a disorder. In previous work, we trained a classifier to distinguish children with Tourette syndrome from healthy control subjects with functional connectivity (30). This classifier generalized to an independent test set of children but not to an independent test set of adults. Similarly, a classifier trained to distinguish adults with Tourette syndrome from healthy control subjects could not accurately classify diagnosis in children. Poor generalizability across age groups suggested that different patterns of functional connectivity underlie childhood and adulthood Tourette syndrome. Thus, cross-sample classifiers may illuminate the nature of atypical functional connectivity in psychiatric disorders.

Confounding Variables

Another important concern when evaluating a classifier is whether the resulting predictions are confounded by other uninteresting variables. For example, one problematic and commonly observed confounding variable in functional connectivity data is head motion in the scanner (54–56). Movement (even submillimeter) in the scanner has been shown to be significantly correlated with several demographic variables (e.g., body mass index, tobacco use, education), behavioral and cognitive abilities (e.g., fluid intelligence, emotion recognition, vocabulary, spatial orientation), and subthreshold clinical symptoms (e.g., impulsivity, antisocial, somatosensory problems) (57,58). Machine learning algorithms are very sensitive to any differentiating characteristics, and hence, a diagnostic classifier may detect motion-related differences in functional connectivity rather than, or in addition to, disorder-related differences. Fortunately, several strategies can help mitigate these effects. First, strategies to reduce the amount of head movement during data collection include real-time motion monitoring (59), behavioral interventions (60), and stabilizing padding (e.g., CaseForge head cases [CaseForge, Inc., Berkley, CA]). Second, processing strategies have been developed and benchmarked to reduce motion-related artifacts in functional connectivity and are particularly useful for mitigating between-group differences in motion (54,61). Finally, matching the amount of head motion between groups in the training set (even after motion denoising) reduces the likelihood that head motion can be used by a classifier. One strategy that we have used to assess the impact of head motion is to intentionally train a classifier to predict individual differences in head motion. We demonstrated that the performance of a head motion classifier was dramatically affected by adequate motion denoising (before denoising: R^2 = .50, after denoising: R^2 = .04) (35). As head motion also affects volumetric, tractography, and task-evoked brain estimates (62–65), these effects should be adequately addressed when using machine learning with other neuroimaging data, along with other potentially confounding variables (e.g., scanner sequence, amount of data).

PART 2: EVALUATING INTERPRETATIONS OF NEURAL MECHANISMS UNCOVERED WITH MACHINE LEARNING

In addition to prediction, investigators hope that machine learning can provide insight into the complex neural mechanisms underlying psychiatric disorders, revealing which regions, connections, networks, or other neuroimaging measures are disrupted. Determining the specific neural circuitry involved, how these features are affected, and how disruption relates to symptom severity or vulnerability has the potential to inform targets for treatment. Generally, two approaches are used to interrogate which features can classify psychiatric disorders: feature selection and feature weight interrogation. When interpreting results from these approaches, it is important to consider the unique utility, interpretability, and reliability of the identified set of neuroimaging features (regions, networks, connections, etc.). Below, we provide suggestions for making inferences about neural mechanisms when using machine learning techniques.

Unique Utility of a Set of Features

Many machine learning approaches involve feature selection. The resulting reduced set of features is often reported, visualized, and interpreted as the archetypal set of features underlying classification, and hence the disorder being studied. However, before inferring that specific features characterize a disorder, it is important to compare the performance of these features with the performance of an appropriate null, as the utility of these features may not be unique. For example, when investigating typical development with machine learning and functional connectivity (35), we used a common, data-driven strategy to select functional connections with the strongest univariate relationships with age (i.e., feature ranking). Age prediction using this feature selection strategy was fairly successful (R^2 = .45 for the top 1000 features). However, this performance was not unique to the selected connections, as a classifier trained to predict age using randomly selected connections was just as successful (average R^2 = .42 $\pm$ .05). In this study, the top ranked features, which are typically interpreted as the most important, performed no better than randomly selected features. This result highlights that classifier performance using a reduced feature set must be evaluated against an appropriate null to claim the unique utility of those features.

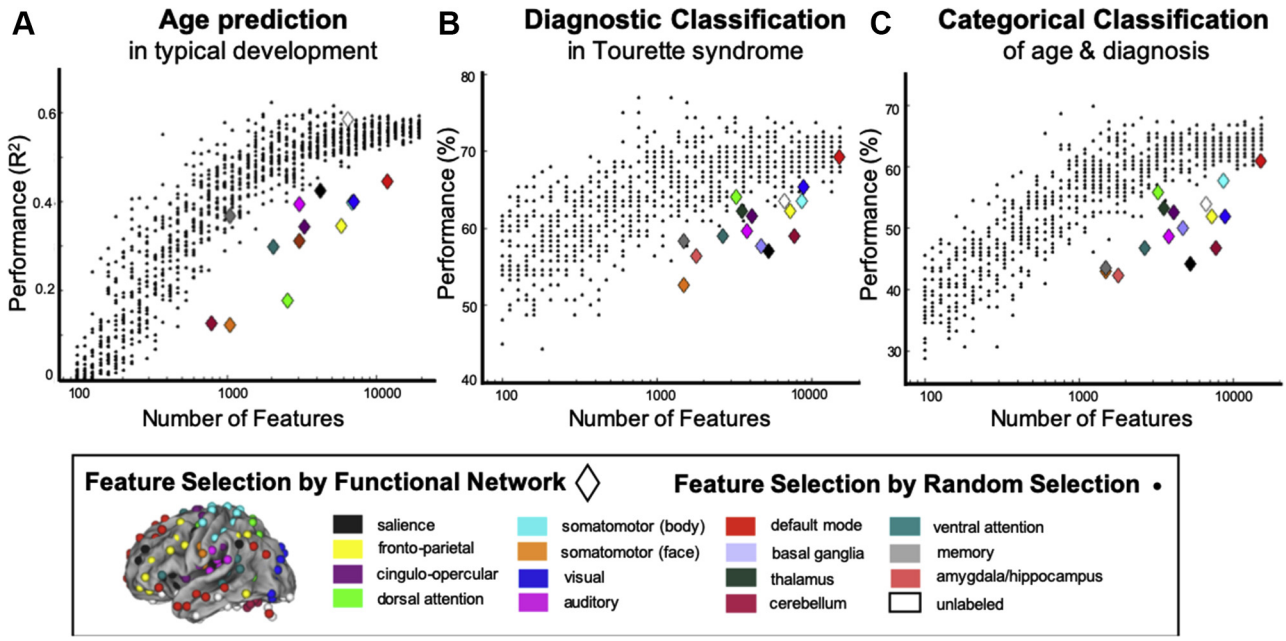

Figure 2. Comparing classifiers using features from different functional networks is confounded by feature number. Performance of classifiers improved with increasing numbers of features; performance of classifiers built from a single functional network was poorer than classifiers built from randomly selected functional connections. Support vector machine learning approaches were used to identify patterns of functional connectivity that **(A)** related to age in a training set of typically developing individuals (n = 129, 7–31 years), **(B)** distinguished individuals with Tourette syndrome from healthy control subjects (n = 156, 7–35 years), and **(C)** categorized individuals by both age group (child, adult) and diagnosis (Tourette syndrome, healthy control subject) (n = 156, 7–35 years). For each case, classifiers were built using either the functional connections associated with a single functional network (e.g., within the default mode and between the default mode and other networks) or randomly selected sets of functional connections that ranged from 100 to 20,000 features.

Another approach, hypothesis-driven feature selection, involves comparing the performance between classifiers that use different a priori feature sets (e.g., functional connections from a particular network, like the default mode network). Such comparisons require careful consideration of potential confounds between feature sets in addition to evaluation against a null as discussed above. We compared the performance of classifiers trained using functional connections selected from different functional networks optimized for age prediction in typical development (regression model) (Figure 2A), diagnostic classification of Tourette syndrome (binary model) (Figure 2B), and categorization of children and adults with and without Tourette syndrome (categorical model) (Figure 2C). While performance did vary by network in each case, performance was highly correlated with network size, i.e., the number of features used for training (e.g., the default mode network is the largest and performed the best). Similarly, when trained with randomly selected functional connections, performance increased with feature number, regardless of network identity. Therefore, we could not determine whether certain functional networks carried more relevant information due to their identity or due to their size. Feature number is one example, but other potential confounds must be carefully considered when comparing the performance of classifiers using a priori feature sets.

Feature Weight Interpretability and Reliability

Another approach used to investigate which features are most affected in a disorder is feature weight interrogation, in which features that are strongly weighted by a classifier are examined. However, the interpretability of feature weights is not always straightforward. First, feature weight interpretability differs across learning algorithms depending on how features are combined (11). Linear regression, support vector machines, and artificial neural networks (i.e., deep learning) all involve the linear weighting of features, but they differ in the number of nonlinear steps (support vector machines: class loss penalties; deep learning: hidden layers, activation function). There is a trade-off [see Figure 1 in Bzdok and Ioannidis (11)], as models with added nonlinearities can better fit complex training data, but the feature weights derived from these models cannot be easily mapped onto digestible descriptions of the underlying mechanism (e.g., increased/decreased functional connectivity in patients) (66). Furthermore, this additional complexity may not be necessary to describe the training data—deep neural networks (more complex) and kernel regression (less complex) achieve comparable accuracies for functional connectivity prediction of behavior and demographics (67). Second, the feature weights from a trained classifier reflect a multidimensional pattern, and thus, consideration of individual feature weights is inappropriate. Classifier performance relies upon the combination of all selected features, from the most strongly weighted to the most weakly weighted. While it may be enlightening to determine if strongly weighted features are organized by a biological principle, such as belonging to a particular functional network, these features should not be interpreted as the only functional connections responsible for classification/prediction.

Interrogating feature weights can also be problematic if those feature weights are unreliable. Unreliable feature weights can occur if a model is not generalizable, as discussed in part 1, or if it overfits the training data. Another possible contributor to this poor reliability is the collinearity of neuroimaging measures. In fact, we demonstrated that functional connections that were most strongly related to the training labels (in our case, age) were also more intercorrelated than expected by chance (35). Intercorrelated features provide redundant information, potentially explaining why randomly selected features predicted age as well as top ranked features. Two features that alone carry equally relevant information may be weighted differently based on the environment of other features, and hence, the distribution of feature weights will be unreliable. Functional connectivity data may be particularly susceptible to such redundancies, in part by definition, as functional networks are composed of regions with similar patterns of connectivity (17,49).

Dimensionality reduction may mitigate this redundancy but may not improve interpretability. Data-driven dimensionality reduction techniques (e.g., principal component analysis) yield orthogonal components, but these components reflect a weighted combination of individual features. Thus, when these components are subsequently weighted by a classifier, the underlying pattern is much less transparent. Alternatively, knowledge-driven dimensionality reduction, reducing features according to an organizing principle (e.g., averaging connections within/between separate functional networks), may reduce redundancy and maintain interpretability. Unfortunately, for functional connectivity (and other neuroimaging measures), which organizing principle best captures the variance related to psychiatric disorders (e.g., areas, functional networks, connector hubs) has not yet been determined.

TRADITIONAL UNIVARIATE VS. MACHINE LEARNING APPROACHES

For certain research questions, multivariate machine learning approaches can provide significant advances over traditional univariate approaches. By combining information across many features, machine learning approaches can often detect differences in neuroimaging data that might not be detected with traditional univariate approaches. Testing for differences among thousands of functional connections using standard statistical approaches can be too conservative with multiple comparisons correction. Additionally, machine learning approaches prevail for work that aims to make predictions for a single individual rather than describing the central tendency of the group. These types of approaches align well with many goals in psychiatry targeting early diagnosis and individualized treatment.

Not all questions are best suited for machine learning (11). While machine learning approaches are well suited to classify and make predictions, they can only indirectly test hypotheses about neurobiological mechanisms. In theory, machine learning methods provide an unbiased approach to identifying disrupted brain mechanisms in psychiatric disorders. However, machine learning algorithms value the utility rather than the relevance of the features used for classification (68), i.e., a feature may be relevant to a psychiatric disorder (e.g., differ between patients and healthy control subjects) but carry redundant information that reduces the utility of any single feature for multivariate classification. Thus, the feature weights of a classifier are not designed to and may not necessarily reveal a complete picture of the brain features affected in psychiatric disorders. Traditional statistical univariate (or multivariate) approaches (e.g., *t* test, analysis of variance, linear regression) prevail in interpretability and may be more appropriate for research questions in which understanding the underlying mechanisms is the primary outcome. Nevertheless, careful use of machine learning methods can provide insight into the nature of atypical brain features, sparking hypotheses for future study with traditional statistical approaches (69,70).

The rise of very large, publicly available datasets, such as the NIH Human Connectome Project (>1000 adults) and the Adolescent Brain Cognitive Development study (>11,800 children), may shift the relative utility of traditional univariate and machine learning approaches. These very large samples will be useful for determining reasonable standards, such as the number of subjects required to produce reliable classifiers. Additionally, large multisite samples like the Adolescent Brain Cognitive Development study can provide a unique resource with which data collected at different sites or across different waves can be used for external validation to demonstrate generalizability. These classifiers trained with very large samples can then be applied to smaller patient samples to identify atypical brain patterns. Finally, it is possible that with the large amount of training data provided by these unprecedented neuroimaging samples, machine learning may be able to uncover a more reliable picture of the complex relationships among features. While these large datasets are accompanied by many advantages, there are also challenges such as confounding variables that must be overcome. Since functional connectivity from the Adolescent Brain Cognitive Development study systematically varies according to acquisition site and scanner manufacturer (71), proper harmonization (72,73) should precede the application of machine learning techniques.

Conclusions

In this targeted review, we have discussed how machine learning can be a useful tool for identifying patterns in multivariate data that have the potential to aid in diagnosis, prognosis, and treatment and to uncover complex mechanisms underlying psychopathology. These goals can only be achieved if best practices are followed. A classifier with the most promise for clinical utility will be one that successfully generalizes to new, independent data and does not rely upon confounding features. While our discussion has focused on examples from functional connectivity MRI, the points raised here apply to other neuroimaging measures and even non-neuroimaging data that share key characteristics, such as large numbers of features (e.g., genes, microbiome, blood biomarkers) or attempts to combine data of many different types. Applying best practices that enhance the likelihood of generalization and replicability, reduce the potential influence of confounds, and increase the interpretability of the data will help machine learning approaches move the field forward in informative and useful ways.

ACKNOWLEDGMENTS AND DISCLOSURES

This project was supported by a grant from the NIH (Grant No. K01MH104592 [to DJG]), by the Tourette Association of America (to DJG), and by the McDonnell Center for Systems Neuroscience (to SEP).

The authors report no biomedical financial interests or potential conflicts of interest.

ARTICLE INFORMATION

From the Institute for Innovations in Developmental Sciences (ANN) and Department of Medical Social Sciences (ANN), Northwestern University, Chicago, Illinois; Department of Psychological and Brain Sciences (DMB), Washington University in St. Louis, and Department of Psychiatry (DMB, DJG), Department of Neurology (SEP), and Department of Neuroscience (SEP), and Mallinckrodt Institute of Radiology (SEP, DJG), Washington University School of Medicine, St. Louis, Missouri; Kennedy Krieger Institute (BLS), Baltimore, Maryland; and Department of Neurology (BLS) and Department of Pediatrics (BLS), Johns Hopkins University School of Medicine, Baltimore, Maryland.

Address correspondence to Ashley Nielsen, Ph.D., 633 N. St. Clair, Chicago, IL 60611; E-mail: ashley.nielsen@northwestern.edu.

Received Aug 5, 2019; revised Oct 29, 2019; accepted Nov 17, 2019.

REFERENCES

1. Bray S, Chang C, Hoeft F (2009): Applications of multivariate pattern classification analyses in developmental neuroimaging of healthy and clinical populations. Front Hum Neurosci 3:32.
2. Lessov-Schlaggar CN, Rubin JB, Schlaggar BL (2016): The fallacy of univariate solutions to complex systems problems. Front Neurosci 10:267.
3. Du Y, Fu Z, Calhoun VD (2018): Classification and prediction of brain disorders using functional connectivity: promising but challenging. Front Neurosci 12:525.
4. Vieira S, Pinaya WHL, Mechelli A (2017): Using deep learning to investigate the neuroimaging correlates of psychiatric and neurological disorders: Methods and applications. Neurosci Biobehav Rev 74:58–75.
5. Margulies DS, Böttger J, Long X, Lv Y, Kelly C, Schäfer A, *et al.* (2010): Resting developments: a review of fMRI post-processing methodologies for spontaneous brain activity. MAGMA 23:289–307.
6. Yoo K, Rosenberg MD, Hsu W-T, Zhang S, Li C-SR, Scheinost D, *et al.* (2018): Connectome-based predictive modeling of attention: Comparing different functional connectivity features and prediction methods across datasets. Neuroimage 167:11–22.
7. Yoo K, Rosenberg MD, Noble S, Scheinost D, Constable RT, Chun MM (2019): Multivariate approaches improve the reliability and validity of functional connectivity and prediction of individual behaviors. Neuroimage 197:212–223.
8. Arbabshirani MR, Plis S, Sui J, Calhoun VD (2017): Single subject prediction of brain disorders in neuroimaging: Promises and pitfalls. Neuroimage 145:137–165.
9. Janssen RJ, Mourão-Miranda J, Schnack HG (2018): Making individual prognoses in psychiatry using neuroimaging and machine learning. Biol Psychiatry Cogn Neurosci Neuroimaging 3:798–808.
10. Dwyer DB, Falkai P, Koutsouleris N (2018): Machine learning approaches for clinical psychology and psychiatry. Annu Rev Clin Psychol 14:91–118.
11. Bzdok D, Ioannidis JPA (2019): Exploration, inference, and prediction in neuroscience and biomedicine. Trends Neurosci 42:251–262.
12. Varoquaux G, Raamana PR, Engemann DA, Hoyos-Idrobo A, Schwartz Y, Thirion B (2017): Assessing and tuning brain decoders: Cross-validation, caveats, and guidelines. Neuroimage 145:166–179.
13. Gratton C, Laumann TO, Nielsen AN, Greene DJ, Gordon EM, Gilmore AW, *et al.* (2018): Functional brain networks are dominated by stable group and individual factors, not cognitive or daily variation. Neuron 98:439–452.
14. Gratton C, Laumann TO, Gordon EM, Adeyemo B, Petersen SE (2016): Evidence for two independent factors that modify brain networks to meet task goals. Cell Rep 17:1276–1288.
15. Finn ES, Scheinost D, Finn DM, Shen X, Papademetris X, Todd Constable R (2017): Can brain state be manipulated to emphasize individual differences in functional connectivity? Neuroimage 160:140–151.
16. Biswal B, Yetkin FZ, Haughton VM, Hyde JS (1995): Functional connectivity in the motor cortex of resting human brain using echo-planar MRI. Magn Reson Med 34:537–541.
17. Power JD, Cohen AL, Nelson SM, Wig GS, Barnes KA, Church JA, *et al.* (2011): Functional network organization of the human brain. Neuron 72:665–678.
18. Yeo BTT, Krienen FM, Sepulcre J, Sabuncu MR, Lashkari D, Hollinshead M, *et al.* (2011): The organization of the human cerebral cortex estimated by intrinsic functional connectivity. J Neurophysiol 106:1125–1165.
19. Power JD, Schlaggar BL, Lessov-Schlaggar CN, Petersen SE (2013): Evidence for hubs in human functional brain networks. Neuron 79:798–813.
20. Davatzikos C (2019): Machine learning in neuroimaging: Progress and challenges. Neuroimage 197:652–656.
21. Woo C-W, Chang LJ, Lindquist MA, Wager TD (2017): Building better biomarkers: brain models in translational neuroimaging. Nat Neurosci 20:365–377.
22. Rosenberg MD, Casey BJ, Holmes AJ (2018): Prediction complements explanation in understanding the developing brain. Nat Commun 9:589.
23. Scheinost D, Noble S, Horien C, Greene AS, Lake EMR, Salehi M, *et al.* (2019): Ten simple rules for predictive modeling of individual differences in neuroimaging. Neuroimage 193:35–45.
24. Fair DA, Nigg JT, Iyer S, Bathula D, Mills KL, Dosenbach NUF, *et al.* (2012): Distinct neural signatures detected for ADHD subtypes after controlling for micro-movements in resting state functional connectivity MRI data. Front Syst Neurosci 6:80.
25. Emerson RW, Adams C, Nishino T, Hazlett HC, Wolff JJ, Zwaigenbaum L, *et al.* (2017): Functional neuroimaging of high-risk 6-month-old infants predicts a diagnosis of autism at 24 months of age. Sci Transl Med 9:393.
26. Craddock RC, Holtzheimer PE 3rd, Hu XP, Mayberg HS (2009): Disease state prediction from resting state functional connectivity. Magn Reson Med 62:1619–1628.
27. Zeng L-L, Wang H, Hu P, Yang B, Pu W, Shen H, *et al.* (2018): Multi-site diagnostic classification of schizophrenia using discriminant deep learning with functional connectivity MRI. EBioMedicine 30:74–85.
28. Arbabshirani MR, Kiehl KA, Pearlson GD, Calhoun VD (2013): Classification of schizophrenia patients based on resting-state functional network connectivity. Front Neurosci 7:133.
29. Greene DJ, Church JA, Dosenbach NUF, Nielsen AN, Adeyemo B, Nardos B, *et al.* (2016): Multivariate pattern classification of pediatric Tourette syndrome using functional connectivity MRI. Dev Sci 19:581–598.
30. Nielsen A, Gratton C, Church JA, Dosenbach NUF, Black KJ, Petersen SE, *et al.* (2020): Atypical functional connectivity in Tourette syndrome differs between children and adults. Biol Psychiatry 87:164–173.
31. Gross C, Hen R (2004): The developmental origins of anxiety. Nat Rev Neurosci 5:545.
32. Swanson JD, Wadhwa PM (2008): Developmental origins of child mental health disorders. J Child Psychol Psychiatry 49:10.
33. Shaw P, Eckstrand K, Sharp W, Blumenthal J, Lerch JP, Greenstein D, *et al.* (2007): Attention-deficit/hyperactivity disorder is characterized by a delay in cortical maturation. Proc Natl Acad Sci U S A 104:19649–19654.
34. Worbe Y, Malherbe C, Hartmann A, Pélégrini-Issac M, Messé A, Vidailhet M, *et al.* (2012): Functional immaturity of cortico-basal ganglia networks in Gilles de la Tourette syndrome. Brain 135:1937–1946.
35. Nielsen AN, Greene DJ, Gratton C, Dosenbach NUF, Petersen SE, Schlaggar BL (2019): Evaluating the prediction of brain maturity from functional connectivity after motion artifact denoising. Cereb Cortex 29:2455–2469.
36. Smyser CD, Dosenbach NUF, Smyser TA, Snyder AZ, Rogers CE, Inder TE, *et al.* (2016): Prediction of brain maturity in infants using machine-learning algorithms. Neuroimage 136:1–9.

37. Dosenbach NUF, Nardos B, Cohen AL, Fair DA, Power JD, Church JA, *et al.* (2010): Prediction of individual brain maturity using fMRI. Science 329:1358–1361.
38. Pruett JR Jr, Kandala S, Hoertel S, Snyder AZ, Elison JT, Nishino T, *et al.* (2015): Accurate age classification of 6 and 12 month-old infants based on resting-state functional connectivity magnetic resonance imaging data. Dev Cogn Neurosci 12:123–133.
39. Satterthwaite TD, Wolf DH, Ruparel K, Erus G, Elliott MA, Eickhoff SB, *et al.* (2013): Heterogeneous impact of motion on fundamental patterns of developmental changes in functional connectivity during youth. Neuroimage 83:45–57.
40. Rudolph MD, Miranda-Domínguez O, Cohen AO, Breiner K, Steinberg L, Bonnie RJ, *et al.* (2017): At risk of being risky: The relationship between "brain age" under emotional states and risk preference. Dev Cogn Neurosci 24:93–106.
41. Li H, Satterthwaite TD, Fan Y (2018): Brain age prediction based on resting-state functional connectivity patterns using convolutional neural networks. Proc IEEE Int Symp Biomed Imaging 2018:101–104.
42. Steele VR, Maurer JM, Arbabshirani MR, Claus ED, Fink BC, Rao V, *et al.* (2018): Machine learning of functional magnetic resonance imaging network connectivity predicts substance abuse treatment completion. Biol Psychiatry Cogn Neurosci Neuroimaging 3:141–149.
43. Gifford G, Crossley N, Fusar-Poli P, Schnack HG, Kahn RS, Koutsouleris N, *et al.* (2017): Using neuroimaging to help predict the onset of psychosis. Neuroimage 145:209–217.
44. Cuthbert BN (2014): The RDoC framework: facilitating transition from ICD/DSM to dimensional approaches that integrate neuroscience and psychopathology. World Psychiatry 13:28–35.
45. Insel TR, Cuthbert BN (2015): Brain disorders? Precisely. Science 348:499–500.
46. Drysdale AT, Grosenick L, Downar J, Dunlop K, Mansouri F, Meng Y, *et al.* (2017): Resting-state connectivity biomarkers define neurophysiological subtypes of depression. Nat Med 23:28–38.
47. Wong T-T (2015): Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation. Pattern Recognit 48:2839–2846.
48. Laumann TO, Gordon EM, Adeyemo B, Snyder AZ, Joo SJ, Chen M-Y, *et al.* (2015): Functional system and areal organization of a highly sampled individual human brain. Neuron 87:657–670.
49. Gordon EM, Laumann TO, Gilmore AW, Newbold DJ, Greene DJ, Berg JJ, *et al.* (2017): Precision functional mapping of individual human brains. Neuron 95:791–807.e7.
50. Noble S, Scheinost D, Constable RT (2019): A decade of test-retest reliability of functional connectivity: A systematic review and meta-analysis. Neuroimage 203:116157.
51. Greene DJ, Black KJ, Schlaggar BL (2016): Considerations for MRI study design and implementation in pediatric and clinical populations. Dev Cogn Neurosci 18:101–112.
52. 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.
53. Varoquaux G (2018): Cross-validation failure: Small sample sizes lead to large error bars. Neuroimage 180:68–77.
54. Power JD, Barnes KA, Snyder AZ, Schlaggar BL, Petersen SE (2012): Spurious but systematic correlations in functional connectivity MRI networks arise from subject motion. Neuroimage 59:2142–2154.
55. Van Dijk KRA, Sabuncu MR, Buckner RL (2012): The influence of head motion on intrinsic functional connectivity MRI. Neuroimage 59:431–438.
56. Satterthwaite TD, Elliott MA, Gerraty RT, Ruparel K, Loughead J, Calkins ME, *et al.* (2013): An improved framework for confound regression and filtering for control of motion artifact in the preprocessing of resting-state functional connectivity data. Neuroimage 64:240–256.
57. Siegel JS, Mitra A, Laumann TO, Seitzman BA, Raichle M, Corbetta M, Snyder AZ (2017): Data quality influences observed links between functional connectivity and behavior. Cereb Cortex 27:4492–4502.
58. Kong X-Z, Zhen Z, Li X, Lu H-H, Wang R, Liu L, *et al.* (2014): Individual differences in impulsivity predict head motion during magnetic resonance imaging. PLoS One 9:e104989.
59. Dosenbach NUF, Koller JM, Earl EA, Miranda-Dominguez O, Klein RL, Van AN, *et al.* (2017): Real-time motion analytics during brain MRI improve data quality and reduce costs. Neuroimage 161:80–93.
60. Greene DJ, Koller JM, Hampton JM, Wesevich V, Van AN, Nguyen AL, *et al.* (2018): Behavioral interventions for reducing head motion during MRI scans in children. Neuroimage 171:234–245.
61. Ciric R, Wolf DH, Power JD, Roalf DR, Baum GL, Ruparel K, *et al.* (2017): Benchmarking of participant-level confound regression strategies for the control of motion artifact in studies of functional connectivity. Neuroimage 154:174–187.
62. Reuter M, Tisdall MD, Qureshi A, Buckner RL, van der Kouwe AJW, Fischl B (2015): Head motion during MRI acquisition reduces gray matter volume and thickness estimates. Neuroimage 107:107–115.
63. Ling J, Merideth F, Caprihan A, Pena A, Teshiba T, Mayer AR (2012): Head injury or head motion? Assessment and quantification of motion artifacts in diffusion tensor imaging studies. Hum Brain Mapp 33:50–62.
64. Yendiki A, Koldewyn K, Kakunoori S, Kanwisher N, Fischl B (2014): Spurious group differences due to head motion in a diffusion MRI study. Neuroimage 88:79–90.
65. Siegel JS, Power JD, Dubis JW, Vogel AC, Church JA, Schlaggar BL, Petersen SE (2014): Statistical improvements in functional magnetic resonance imaging analyses produced by censoring high-motion data points. Hum Brain Mapp 35:1981–1996.
66. Haufe S, Dähne S, Nikulin VV (2014): Dimensionality reduction for the analysis of brain oscillations. Neuroimage 101:583–597.
67. He T, Kong R, Holmes AJ, Nguyen M, Sabuncu MR, Eickhoff SB, *et al.* (2019): Deep neural networks and kernel regression achieve comparable accuracies for functional connectivity prediction of behavior and demographics. Neuroimage 206:116276.
68. Guyon I, Elisseeff A (2003): An introduction to variable and feature selection. J Mach Learn Res 3:1157–1182.
69. Oquendo MA, Baca-Garcia E, Artés-Rodríguez A, Perez-Cruz F, Galfalvy HC, Blasco-Fontecilla H, *et al.* (2012): Machine learning and data mining: strategies for hypothesis generation. Mol Psychiatry 17:956–959.
70. Holm EA (2019): In defense of the black box. Science 364:26–27.
71. Marek S, Tervo-Clemmens B, Nielsen AN, Wheelock MD, Miller RL, Laumann TO, *et al.* (2019): Identifying reproducible individual differences in childhood functional brain networks: An ABCD Study [published online ahead of print Sept 19]. Dev Cogn Neurosci 40:100706.
72. Fortin J-P, Parker D, Tunç B, Watanabe T, Elliott MA, Ruparel K, *et al.* (2017): Harmonization of multi-site diffusion tensor imaging data. Neuroimage 161:149–170.
73. Fortin J-P, Cullen N, Sheline YI, Taylor WD, Aselcioglu I, Cook PA, *et al.* (2018): Harmonization of cortical thickness measurements across scanners and sites. Neuroimage 167:104–120.