Biological Psychiatry: Cognitive Neuroscience and Neuroimaging

Making individual prognoses in psychiatry using neuroimaging and machine learning

--Manuscript Draft--

Manuscript Number:	BPSC-D-17-00185R2
Full Title:	Making individual prognoses in psychiatry using neuroimaging and machine learning
Article Type:	Special Issue - Review
Corresponding Author:	Ronald Johannes Janssen, Ph.D. University Medical Center Utrecht Utrecht, Utrecht NETHERLANDS
Order of Authors:	Ronald Johannes Janssen, Ph.D.
	Janaina Mourao-Miranda, Ph.D.
	Schnack Hugo, Ph.D.
Abstract:	Psychiatric prognosis is a difficult problem. Making a prognosis requires looking (far) into the future, as opposed to diagnosis, which is concerned with the current state. During the follow-up period, many factors will influence the course of the disease. Combined with the usually scarcer longitudinal data and the variability in the definition of outcomes/transition, this makes prognostic predictions a challenging endeavor. Employing neuroimaging data in this endeavor introduces the additional hurdle of high-dimensionality. Machine-learning techniques are especially suited to tackle this added problem. This review starts with a brief introduction to machine learning in the context of its application to medical data. We highlight a few issues that are especially relevant for prediction of outcome and transition using neuroimaging. We then review the literature applying machine learning for this purpose. Critical examination of the studies and their results with respect to the relevant issues revealed the following: i) There is growing evidence for the prognostic capability of machine-learning-based models using neuroimaging; ii) reported accuracies may be too optimistic due to small sample sizes and the lack of independent test samples. Finally, we discuss options to improve the reliability of (prognostic) prediction models. These include new methodologies and multi-modal modeling. Most importantly, however, is our conclusion that future work will need to provide properly (cross-)validated accuracy estimates of models trained on sufficiently large datasets. Nevertheless, machine learning represents a powerful tool in the search for psychiatric biomarkers.

Making individual prognoses in psychiatry using neuroimaging and machine learning

Ronald J. Janssen[1]*, Janaina Mourão-Miranda[2,3], Hugo G. Schnack[1]

[1] Department of Psychiatry, Brain Center Rudolf Magnus, University Medical Center Utrecht, Utrecht University, Netherlands

[2] Centre for Medical Image Computing, Department of Computer Science, University College London, UK.

[3] Max Planck University College London Centre for Computational Psychiatry and Ageing Research, University College London, UK

* Corresponding author: Department of Psychiatry, A01.161
University Medical Center Utrecht, Heidelberglaan 100
3584 CX Utrecht, The Netherlands
phone: +31-88-75-53386;
Email: janssen.rj@gmail.com

Short title: Psychiatric prognoses with imaging and machine learning

Key words: imaging; schizophrenia; major depressive disorder; machine learning; prognosis; prediction

250 words in the abstract.

4361 words in the main text.

1 table.

4 figures.

1 supplementary material.

Abstract

Psychiatric prognosis is a difficult problem. Making a prognosis requires looking (far) into the future, as opposed to diagnosis, which is concerned with the current state. During the follow-up period, many factors will influence the course of the disease. Combined with the usually scarcer longitudinal data and the variability in the definition of outcomes/transition, this makes prognostic predictions a challenging endeavor. Employing neuroimaging data in this endeavor introduces the additional hurdle of high-dimensionality. Machine-learning techniques are especially suited to tackle this challenging problem.

This review starts with a brief introduction to machine learning in the context of its application to clinical neuroimaging data. We highlight a few issues that are especially relevant for prediction of outcome and transition using neuroimaging.

We then review the literature applying machine learning for this purpose. Critical examination of the studies and their results with respect to the relevant issues revealed the following: i) There is growing evidence for the prognostic capability of machine-learning-based models using neuroimaging; ii) reported accuracies may be too optimistic due to small sample sizes and the lack of independent test samples.

Finally, we discuss options to improve the reliability of (prognostic) prediction models. These include new methodologies and multi-modal modeling. Paramount, however, is our conclusion that future work will need to provide properly (cross-)validated accuracy estimates of models trained on sufficiently large datasets. Nevertheless, with the technological advances enabling

acquisition of large databases of patients and healthy subjects machine learning represents a powerful tool in the search for psychiatric biomarkers.

1 Introduction

Over the past two decades, machine learning (ML) has become effective enough to be of interest beyond the artificial-intelligence field it originated from. This increased effectiveness has been achieved by developments in the algorithms themselves, but perhaps more importantly, the increased availability of computational resources and large datasets.

The ability of algorithms such as support vector machines (SVMs) to handle datasets with (many) more variables than observations is especially attractive from a neuroimaging point of view. With the growing interest in applying ML in the field of psychiatric neuroimaging for outcome (1) and transition (2), there is need of understanding the possibilities and limitations of these techniques, and how to use them properly.

This review is structured as follows: first we will discuss the basis for the effectiveness of ML. Next, we will review the application of ML in the context of predicting psychiatric outcomes, followed by an assessment of the limitations of current approaches. Finally, we give some thought to possibilities for future directions.

2 Machine learning: what it does(n't do)

In this section, we will focus on why ML might work well and provide some background to aid with interpreting results from studies employing such methods. For a more in-depth treatment of the do's and don'ts of applying ML to (psychiatric) neuroimaging, the reader is referred to Arbabshirani et al.(3), and Wolfers et al.(4)

In order to decide on the usefulness of ML it is necessary to discuss what it is and what it does. ML is a term that covers a large variety of mathematical/statistical models developed within the field of artificial intelligence. The common denominator of these models is their ability to extract patterns from previous observations, in order to predict new ones. In line with Breiman(5) and Arlot & Celise(6), we will refer to the model that extracts patterns as a training algorithm and the pattern it has extracted as a prediction model. This distinction will be important when discussing validation later on, to avoid confusion about which of the two is being validated.

2.1 How is it different from conventional statistics?

The goal of statistics is to provide a rigorous basis for reasoning under uncertainty. Essentially, this is also the problem being solved by ML techniques, even those developed from a non-statistical perspective. Hence, one could argue that ML is a part of statistics, although nomenclature can differ. For example, independent and dependent variables are often called features and targets, respectively.

The success of ML approaches is due to a focus on generalizability, i.e. the performance of models on new data. Generalizability is affected by three sources of error: underfitting, overfitting, and irreducible error. These sources are illustrated in Figure 1 using a classification example, though everything applies equally to regression. Irreducible error solely is a property of the data, it is caused by noise in the dependent variable[*]; the best possible classifier would still misdiagnose these cases. Underfitting occurs when a training algorithm is constrained in a

[*] Another way to look at this is to say that the predictors do not contain all the information necessary to reconstruct the outcome measures.

way that prevents it from producing an optimal classifier. Overfitting occurs when the ostensible pattern found by the training algorithm is partially due to noise.

2.2 Validation

To estimate generalizability, one can split their data into training and test sets (see Fig 2). Applying a training algorithm to the training set produces a prediction model that can be applied to the test set. Performance on this test set is an estimate of generalizability. However, this estimate might be highly dependent on the specific partition used.

Cross-validation (CV) addresses this by repeating the train/test steps on different splits of the data. There are several forms of CV, each defined by the method for generating these splits; we will focus on the most common, generally best performing, form: *k*-fold CV(6). In *k*-fold CV, the dataset is evenly split into *k* test-sets. The prediction model tested on such a left-out set is trained on the remaining *k*-1 sets combined. The average performance gives an estimate of the quality of models produced with this training algorithm. A special form of *k*-fold CV is when *k*=*N*, yielding *N* test sets of size 1; this procedure is called leave-one-out (LOO) CV.

As mentioned earlier, it is necessary for the training algorithm to be sufficiently flexible to avoid underfitting, yet not so flexible that it will overfit. This flexibility is usually controlled via some parameter of the training algorithm. This parameter, or hyperparameter, can be optimized by trying different values and picking the one that maximizes generalizability. It should be noted that this parameter setting is now also based on the test set. As such, the associated generalization score is no longer applicable as the test set is no longer independent. Instead, CV should be applied hierarchically, also known as nested CV(7). In this approach, illustrated in Fig.

2, there is an outer CV loop used to estimate generalizability and the training set in each iteration of this is used as input to an inner CV loop to optimize parameters.

2.3 Types of predictions

As with any analysis, it is important to carefully consider the choice of (in)dependent variables that will be used for modeling. Focusing on the dependent variable, i.e. a patient's outcome, a fundamental choice is whether to do classification (discrete outcomes, e.g. diagnostic transition) or regression (continuous outcome, e.g. functioning).

Classification has the advantage of making unambiguous statements. This makes classifiers generally easier to train than regression models; they ask a simple yes/no question, as opposed to the more detailed predictions in regression. At the same time, this also means that discrete outcomes lose information like severity. This is often compounded by the arbitrariness of outcome definition. Such information is useful e.g. when examining subclinical populations or treatment effectiveness.

In practice, many classification algorithms actually carry out something like regression under the hood and convert the continuous output to a discrete one. However, interpretation of this intermediate output is usually not straightforward. If, for example, some clinical score is thresholded to produce diagnoses, it may be preferable to do regression on that score directly. Fig. 3 demonstrates that regression will achieve the same goal as classification, while maintaining interpretability. Regression also utilizes the data more fully, i.e. it might extract more information from the data, though at the risk of increased overfitting.

So far, we have assumed that the dependent variable is given, which is the case in supervised learning. If the dependent variable cannot (easily) be defined and measured, unsupervised learning might provide some insight. Unsupervised learning aims to find some underlying structure in the observations. Example of unsupervised approaches include clustering approaches and dimensionality reduction techniques (e.g. PCA).

2.4 Considerations

An important consideration in using CV to do parameter optimization is the choice of performance measure. There are several options for either classification or regression; we define and discuss some of the more common ones in the supplementary material. Different measures emphasize different things, for example, mean-absolute-error is less tolerant of outliers than mean-squared-error.

Properly performed CV can help when sample sizes are relatively small compared to the number of degrees of freedom. However, the procedure does require sufficient data for generating training and test sets. The training set needs to be large enough to reliably train a model and the test set needs to be large enough to get a good estimate of the accuracy (8). High variability in performance over folds might indicate that the dataset is too small.

One must also be careful when comparing generalizability reported in the literature. When parameters are optimized using CV, instead of nested CV, any reported generalizability is almost certainly inflated (6, 7, 9). Even when the reported accuracies are derived from independent test sets, a seemingly high accuracy might not be significantly different from chance (3, 8). Permutation testing is a robust approach for determining such significance (10).

When using *k*-fold CV, care should also be taken in generating the folds. At a minimum, folds should be generated such that structure in the data (e.g. the order of participants) are unlikely to affect the procedure. Ideally, the folds should be stratified so that all training and test sets have similar distributions of both features and targets (9).

Finally, researcher degrees-of-freedom (11) affect ML as well (8). Trying different preprocessing steps, training algorithms, feature sets, etc., are common steps. Conventional statistics advises a clearly defined and preregistered analysis protocol, which applies to ML just as well. However, an exploratory approach is not problematic in itself, as long as it is distinguished from a confirmatory approach. An exploratory approach might identify possible avenues of research, but should never be interpreted as more than very tentative evidence. Crucial to the interpretation of exploratory research is the documentation of these researcher degrees-of-freedom and the associated accuracies. These, along with sample size, should temper any interpretation of reported results.

The conclusion, then, is that machine learning is a very useful tool. Like any tool, however, it has limitations and can be used inappropriately.

3 Review of literature on neuroimaging-based prognostic predictions in psychiatry

On November 6th, 2017, we searched in PubMed for studies that had applied machine learning to neuroimaging data to make prognostic predictions in patients with (risk of) psychiatric

disorders. (See Supplement for the precise search criteria.) We removed studies that did not make predictions about patients' future status. Studies that did not test their prediction model in unseen cases (either by using an independent test set or, most often, by using cross-validation within the training sample – see section 2.2 of this review) were also excluded. This search resulted in 33 studies: 17 studies for depression (MDD)(12–28), 10 for psychosis(29–38), 1 for bipolar disorder (BD)(39), 3 for ADHD(40–42) and 2 for ASD(43, 44) (see Table 1). Quite some interesting results can be seen from this overview:

3.1 Studies: history, numbers, aims and designs

The oldest studies date from 2009 (less than 10 years from to date). The start of appearance of outcome/transition studies more or less coincides with the start of applying ML for diagnostic classification in psychiatry (i.e., classification of being ill or not; see review by Wolfers et al(4)).

The number of prognostic ML studies (33) is much lower than the number of diagnostic ML studies (82, for schizophrenia plus mood disorders, in a 2015 review by Wolfers et al(4)). Most studies were on MDD and psychosis (17 and 10, respectively). In the following sections, we will focus our discussion on these two disorders.

Interestingly, half of the psychosis studies is about outcome/treatment response in patients and half is about predicting transition (or functional outcome) in young subjects at high risk for developing psychosis. No imaging studies about predicting transition to MDD were found.

Different study designs have been used. For MDD, all but one(20) study had a treatment-response design, with follow-up durations ranging from 2–16 weeks. For psychosis, we see the opposite: All but one(38) study had long follow-up durations: 1–7 year. Apart from the study by Khodayari-Rostamabad et al.(30), the long-term studies were naturalistic, either predicting outcome or illness course, or transition to psychosis. Naturally, the latter type of study requires long follow-up times.

3.2 Sample size and performance

Most studies reached cross-validated prediction accuracies of 70% or higher, however, sample sizes (N) were generally modest. Figure 4 shows balanced accuracy versus N for MDD and psychosis. For MDD, N ranged from 10–124, with a mean of 48.0 and a median of 34. For psychosis, N ranged from 27–212, with a mean of 60.1 and a median of 45. However, the N=212 study(36) is an 'outlier': the next largest study had N=73. There seems to be a strong negative relationship between sample size and accuracy: For MDD, a linear regression of accuracy on N yields: Acc(%) = 86.91 - 0.121×N. The corresponding correlation coefficient is -0.52. For psychosis, the number of studies is too small to reliably perform such an analysis, also because of the different designs (transition in subjects at risk and outcome in patients). However, the negative association between N and accuracy seems to be present here as well. This effect has been observed earlier in diagnostic ML studies and can be explained in terms of sample heterogeneity(45, 46) or uncertainty in accuracy estimates(8) (see section 4.2).

3.3 Follow-up duration and image modality

The lack of long-term MDD prognostic studies and of short-term prognostic studies in psychosis hinders a robust analysis of the influence of follow-up duration on prediction accuracy. Of course, many other factors play a confounding role here, including differences in patient populations (inclusion criteria) and treatment choices. Disregarding differences in study designs, a strong negative correlation (r=-0.47) was found between follow-up duration and prediction accuracy in psychosis studies. For MDD studies, at first sight there does not seem to be a clear relation between prediction accuracy and how far in the future one aims to predict response. However, closer inspection of the MDD results reveals that all (5) short-term studies(14, 18, 22, 25, 26) (at most 4 weeks) score 80% or higher. There are, however, also two 'long-term' (8 and 12 weeks, respectively) MDD studies(19, 28) that score ≥80%. To interpret this, we need also take into account the imaging modalities used for prediction.

The MDD and psychosis communities seem to favor different neuroimaging modalities. Many MDD studies used EEG or (resting state) functional MRI, while for psychosis, almost all studies used structural MRI (no (rs)fMRI) and only two used EEG. EEG seems to be high-ranking in the list of MDD treatment-response studies. However, all but one (6 weeks) EEG studies had short follow-up duration, so it is difficult to disentangle these effects. Interestingly, the 'longer-term' (12 weeks) study(19) that scored 90% accuracy turns out to be a multi-modal MRI study, combining different imaging modalities to make a prognosis. This study was the only late-life depression study.

Some studies combined imaging and non-imaging predictors in their models (e.g., clinical scores or genetic data). The best performing model by Kim et al.(41) incorporated data from many different sources, including rsfMRI data (although the imaging data was not reported to play a significant role in predicting outcome).

The entanglement of follow-up duration and imaging modality makes it difficult to draw firm conclusions about their respective influences on prediction accuracy, but one may speculate that prognostic predictions on shorter-term and/or based on combining different imaging modalities can be made with more accuracy.

3.4 Training algorithms.

Many different algorithms were used, although the support vector machine (SVM, and SVR, its regression variant) was used in 16, i.e., more than 50% of the studies, often (11 studies, see Table 1) in combination with structural MRI data. This choice may be logical, because of SVM's robustness against $p>>N$ (many more features than cases) problems, thus avoiding overfitting (see section 2). This may also explain the relative lack of DNNs (2 studies); these algorithms can generate powerful, non-linear models, but they require large sample sizes. With the emergence of large multi-site collaborations, e.g. ENIGMA (47) and the UK Biobank (48), the use of DNNs for imaging-based prognosis may become more feasible. Another algorithm that was relatively frequently encountered were different flavors of logistic regression (LR; 5 studies). Although some studies compared different algorithms, no clear 'winner algorithm' seems to be present. Choice of algorithm probably has to do with factors such as presence of expertise, availability,

convenience, etc. For some (features, output) combinations, certain algorithms are more suitable than others.

4 What are the limitations of the current studies?

The recent ML applications predicting outcome and transition from neuroimaging data represent an important advance towards the development of neurobiomarkers for psychiatric disorders. Nevertheless, these studies present some important limitations that should be carefully considered.

4.1 Clinical use: the need for validation

With an average accuracy of 76% and 25 (out of 52) studies reaching ≥80% accuracy, the clinical use of these models may appear nearby (49). However, it is unknown whether the models can maintain these accuracies when being applied to brain images acquired using different scanners and in different populations. While all studies (per our inclusion criteria) used cross-validation (CV), only three studies (27, 33, 38) used independent test samples (or holdout test subjects from the training sample), including one multi-center study that applied leave-site-out (LSO) CV(38). The advantage of these approaches is that they better test the generalizability of the prediction models. In this regard, the field of making prognostic predictions is still in its infancy, when compared to the field of applying ML for disease classification, where the use of independent test samples is more or less required. A few other studies were multi-center (20, 27, 33, 36), and large-scale international consortia for outcome prediction in psychosis are currently collecting data (50, 51).

4.2 Sample size

Perhaps the most important limitation is the fact that most studies were based on very small samples. There has been abundant evidence that the performance of neuroimaging-based classifiers measured using cross-validation (CV) in small samples is overoptimistic when compared to the model performance on an independent data set. Prediction models trained in small samples tend to have large variance in CV accuracy, which, in combination with publication bias (see section 4.3), may lead to overoptimistic estimates of generalization accuracy(8, 52). Apart from this sampling and measurement effect, smaller samples are likely to be more homogeneous, since acquisition of large samples often requires relaxing the inclusion criteria or using multi-center data, increasing the variation among subjects, particularly patients(45, 53). As a result, small-sample prediction models may be tuned to features that are shared by a homogeneous set of patients. These features may thus be specific to these patients, rather than specific to the disease. Application of small-sample prediction models to new patients will most likely lead to disappointing accuracies.

4.3 Reported and unreported accuracies.

Quite a few studies employed more than one modeling approach. Some studies used a number of different image modalities or derivatives from them, e.g., gray matter and white morphology from structural images, fMRI scans with different paradigms, or different networks from resting state scans. This leads to different (combinations of) feature sets. Other modeling variations include the use of different algorithms and/or optimization strategies for feature selection and or hyperparameter tuning (see sections 2.2 and 2.4 re. cross-validation and user degrees of

freedom). These studies reported all results, often highlighting (in the Abstract) the best-performing model. The publication of the results of all models that had been investigated is very useful, since it provides the reader insight in, e.g., the biomarker potential of each modality. In addition, it displays the variety in performance between different data sources and/or training algorithms. In some studies, the accuracies span a range of more than 30% (Fig. 4C).

Choosing the model with the highest (cross-validation) accuracy in the training dataset may lead to a reported prediction accuracy that is too optimistic. In other words, in testing the best model's generalizability in an independent test sample, a relatively large drop in accuracy may occur. If theoretical support for the superiority could be found for the 'winning' model, and if not dozens of models have been tried, this optimistic bias may be small. It can also not be ruled out that studies reporting only one model and its accuracy have tried (many) more models – probably with lower accuracies. Withholding evidence of other approaches' (poorer) performances is a form of publication bias and it renders the report less informative, while the reader is also left unaware of the fact that the accuracy of the published model may be too optimistic.

4.4 Target uncertainty

Another potential limitation of these studies is the uncertainty of the label definition (e.g., there might be some variability on how outcome or transition are defined and therefore the classes might be very heterogeneous), which will strongly affect the performance of the ML models.

While most studies present predictions of outcome class (e.g., good versus poor), some studies predicted outcome on a continuous scale, e.g., general functioning (GAF) or depressive symptom severity (HDRS). The performance of such prediction models is provided in terms of (squared) correlation coefficients between true and predicted scores (see section 2.4). If continuous outcome measures are available, predicting these may be favored over their categorized versions, since subtle differences between outcomes may get lost in this process and category (class) boundaries are usually artificial (see section 2.3).

Finally, all clinical information available is summarized into a single label (either categorical or continuous. Therefore, a lot of relevant information is potentially left out of the modeling.

5 Advanced machine learning techniques, what do they offer?

So far, most studies applying ML to predict outcome and transition from neuroimaging data have focused on categorical classification problems using a single neuroimaging modality. The successes of these models are limited by how informative is the neuroimaging modality for the prediction task considered (i.e. predicting outcome or transition), the reliability of pre-defined labels (e.g. responders vs. non-responders) and sample size available. In clinical research settings, often different sources of information are acquired in order to identify possible biomarkers including different modalities of neuroimaging data, socio-demographic information, measures based on neurocognitive testing as well as genomic data. Outcome may very well depend on, e.g., socioeconomic status, while response to a certain type of medication is likely do depend a patient's genetic makeup. One promising future direction is the

development of multi-modal prediction models combining these different sources of information. In this case the assumption is that the labels are reliable and can be used to train the prediction model. The model then learns the relative contribution of the different sources of information for prediction. This direction has been investigated in a number of studies for diagnosis (54–61) and to a lesser extent for prognosis(33, 58, 62–64).

A big challenge in training single and multi-modal ML models to predict outcome or transition is that often the labels are not reliable or unknown (i.e. there is no available longitudinal/follow-up information about the outcome or transition). In order to overcome these limitations another potential future direction is using unsupervised ML approaches to find underlying structure in the data that can be used to split a heterogeneous patient sample into more homogeneous subgroups, by identifying patients with similar biological and/or clinical characteristics (see Schnack(53), for a review of such approaches). The underlying assumption is that subgroups of patients with common characteristics are more likely to have similar outcome/transition. One example of this approach can be found in Drysdale et al.(27), where the authors used canonical correlation analysis to find a low-dimensional space capturing the association between resting state connectivity and a combination of clinical symptoms quantified by the 17-item Hamilton Depression Rating Scale (HDRS). The authors identified two canonical variates that they termed "anhedonia-related connectivity features" and "anxiety-related connectivity features" which were used in a hierarchical cluster analysis. The clustering analysis revealed four subgroups or biotypes of patients associated with different clinical-symptom profiles, which predicted responsiveness to transcranial magnetic stimulation (rTMS) therapy.

The challenge in using unsupervised approaches to identify subgroups (i.e. clustering) is the validation of the obtained subgroups, as different groupings might be obtained depending on the data used for modelling, model assumptions and potential confounds. Validation of obtained subgroups can be done using additional or follow-up information combined with metrics that characterize the stability/reproducibility of the groupings. Furthermore, considering the current debate within psychiatry between categorical versus dimensional approaches (e.g. (65)), careful considerations should also be taken to test if the underlying structure in the data supports a categorical separation in subgroups or a continuous variation from normal to abnormal (see (66) for a commentary on this issue).

ML approaches can be also used to define a normative model that can identify abnormal patterns associated to specific outcome measures. For example, in Mourão-Miranda et al.(67) a one-class support vector machine (OC-SVM) was used to define a normative model based on patterns of brain activation to sad faces. The model was trained on a sample of healthy controls and tested on healthy controls and depressed patients. The authors found that 21% of the healthy subjects and 52% of the depressed patients were considered outliers with respect to the normative base, respectively. Interestingly, among the patients classified as outliers 70% did not respond to treatment and among those classified as non-outliers 89% responded to treatment. As future research, this approach could be extended to define multi-modal normative models.

Multi-output learning approaches can be used to learn the association between neuroimaging data and multiple targets. These approaches might be advantageous as they can explore the relationship between different targets (e.g. clinical scores) and potentially better characterize the

diseases status of patient groups. A number of approaches can be used for multi-output predictions, such as Partial Least Square (PLS), Canonical Correlation Analysis(CCA) and Reduced Rank Regression (RRR). For example, in Rahim et al(61). the authors compared a number of multi-output prediction models for predicting multiple behavioral scores from neuroimaging data and found that they generalize better for a new cohort when compared to single output models.

The application of multi-task learning to predict outcome or transition is another promising direction. The main idea in multi-task learning is to solve multiple "prediction tasks" simultaneously, exploring the commonalities and differences across tasks (e.g. Evgeniou et al.(68)). This approach can be beneficial when the individual tasks are related. Multi-task learning can be used to predict multi-outputs (treating the prediction of each target as a different task, e.g. Wan et. Al.(60)) or to predict multiple related outcomes (treating the prediction of different subgroups as a different task). Another step to improve prognostic models could be by including data at multiple time-points, capturing the dynamics of the disease. The effects of environmental factors (e.g., medication) and their interactions with patient properties (e.g., neuroimaging data) can be taken into account while training a model to predict a patient's future state.

In summary, with the technological advances enabling acquisition of large volumes of patient data, ML becomes a powerful tool to discover hidden regularities or patterns in these complex and heterogeneous data that cannot be easily found by human experts. These complex patterns can be considered individual signatures that have the potential to drive precision medicine.

Nevertheless, as with other modeling approaches, ML models need to be validated and replicated using big samples before they can make an impact in real-word clinical outcome prediction.

Acknowledgements

Janaina Mourão-Miranda was supported by the Wellcome Trust under grant number WT102845/Z/13/Z.

Financial disclosures

Ronald Janssen, Janaina Mourão-Miranda and Hugo Schnack report no biomedical financial interests or potential conflicts of interest.

References

1. Dazzan P, Arango C, Fleischacker W, Galderisi S, Glenthoj B, Leucht S, *et al.* (2015): Magnetic Resonance Imaging and the Prediction of Outcome in First-Episode Schizophrenia: A Review of Current Evidence and Directions for Future Research. *Schizophr Bull*. 41: 574–583.

2. 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.

3. Arbabshirani MR, Plis S, Sui J, Calhoun VD (2017): Single subject prediction of brain disorders in neuroimaging: Promises and pitfalls. *NeuroImage*. 145: 137–165.

4. Wolfers T, Buitelaar JK, Beckmann CF, Franke B, Marquand AF (2015): From estimating activation locality to predicting disorder: A review of pattern recognition for neuroimaging-based psychiatric diagnostics. *Neurosci Biobehav Rev*. 57: 328–349.

5. Breiman L (2001): Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). *Stat Sci*. 16: 199–231.

6. Arlot S, Celisse A (2010): A survey of cross-validation procedures for model selection. *Stat Surv*. 4: 40–79.

7. Varma S, Simon R (2006): Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*. 7: 91.

8. Varoquaux G (2017): Cross-validation failure: Small sample sizes lead to large error bars. *NeuroImage*. . doi: 10.1016/j.neuroimage.2017.06.061.

9. Kohavi R (1995): A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. Morgan Kaufmann, pp 1137–1143.

10. Stelzer J, Chen Y, Turner R (2013): Statistical inference and multiple testing correction in classification-based multi-voxel pattern analysis (MVPA): Random permutations and cluster size control. *NeuroImage*. 65: 69–82.

11. Simmons JP, Nelson LD, Simonsohn U (2011): False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychol Sci*. 22: 1359–1366.

12. Costafreda SG, Chu C, Ashburner J, Fu CHY (2009): Prognostic and Diagnostic Potential of the Structural Neuroanatomy of Depression. (K. Domschke, editor) *PLoS ONE*. 4: e6353.

13. Khodayari-Rostamabad A, Reilly JP, Hasey G, de Bruin H, MacCrimmon D (2010): Using pre-treatment EEG data to predict response to SSRI treatment for MDD. IEEE, pp 6103–6106.

14. Khodayari-Rostamabad A, Reilly JP, Hasey GM, de Bruin H, MacCrimmon D (2011): Using pre-treatment electroencephalography data to predict response to transcranial magnetic stimulation therapy for major depression. IEEE, pp 6418–6421.

15. Gong Q, Wu Q, Scarpazza C, Lui S, Jia Z, Marquand A, *et al.* (2011): Prognostic prediction of therapeutic response in depression using high-field MR imaging. *NeuroImage*. 55: 1497–1503.

16. Korgaonkar MS, Williams LM, Song YJ, Usherwood T, Grieve SM (2014): Diffusion tensor imaging predictors of treatment outcomes in major depressive disorder. *Br J Psychiatry*. 205: 321–328.

17. Miller JM, Schneck N, Siegle GJ, Chen Y, Ogden RT, Kikuchi T, *et al.* (2013): fMRI response to negative words and SSRI treatment outcome in major depressive disorder: A preliminary study. *Psychiatry Res Neuroimaging*. 214: 296–305.

18. Erguzel TT, Ozekes S, Gultekin S, Tarhan N, Hizli Sayar G, Bayram A (2015): Neural Network Based Response Prediction of rTMS in Major Depressive Disorder Using QEEG Cordance. *Psychiatry Investig*. 12: 61.

19. Patel MJ, Andreescu C, Price JC, Edelman KL, Reynolds CF, Aizenstein HJ (2015): Machine learning approaches for integrating clinical and imaging features in late-life depression classification and response prediction: Prediction models for late-life depression. *Int J Geriatr Psychiatry*. 30: 1056–1067.

20. Schmaal L, Marquand AF, Rhebergen D, van Tol M-J, Ruhé HG, van der Wee NJA, *et al.* (2015): Predicting the Naturalistic Course of Major Depressive Disorder Using Clinical and Multimodal Neuroimaging Information: A Multivariate Pattern Recognition Study. *Biol Psychiatry*. 78: 278–286.

21. Williams LM, Korgaonkar MS, Song YC, Paton R, Eagles S, Goldstein-Piekarski A, *et al.* (2015): Amygdala Reactivity to Emotional Faces in the Prediction of General and Medication-Specific Responses to Antidepressant Treatment in the Randomized iSPOT-D Trial. *Neuropsychopharmacology*. 40: 2398–2408.

22. van Waarde JA, Scholte HS, van Oudheusden LJB, Verwey B, Denys D, van Wingen GA (2015): A functional MRI marker may predict the outcome of electroconvulsive therapy in severe and treatment-resistant depression. *Mol Psychiatry*. 20: 609–614.

23. Goldstein-Piekarski AN, Korgaonkar MS, Green E, Suppes T, Schatzberg AF, Hastie T, *et al.* (2016): Human amygdala engagement moderated by early life stress exposure is a biobehavioral target for predicting recovery on antidepressants. *Proc Natl Acad Sci*. 113: 11955–11960.

24. Redlich R, Opel N, Grotegerd D, Dohm K, Zaremba D, Bürger C, *et al.* (2016): Prediction of Individual Response to Electroconvulsive Therapy via Machine Learning on Structural Magnetic Resonance Imaging Data. *JAMA Psychiatry*. 73: 557.

25. Mumtaz W, Xia L, Mohd Yasin MA, Azhar Ali SS, Malik AS (2017): A wavelet-based technique to predict treatment outcome for Major Depressive Disorder. (D. Hu, editor) *PLOS ONE*. 12: e0171409.

26. Al-Kaysi AM, Al-Ani A, Loo CK, Breakspear M, Boonstra TW (2016): Predicting brain stimulation treatment outcomes of depressed patients through the classification of EEG oscillations. IEEE, pp 5266–5269.

27. Drysdale AT, Grosenick L, Downar J, Dunlop K, Mansouri F, Meng Y, *et al.* (2016): Resting-state connectivity biomarkers define neurophysiological subtypes of depression. *Nat Med*. 23: 28–38.

28. Crane NA, Jenkins LM, Bhaumik R, Dion C, Gowins JR, Mickey BJ, *et al.* (2017): Multidimensional prediction of treatment response to antidepressants with cognitive control and functional MRI. *Brain*. 140: 472–486.

29. Koutsouleris N, Meisenzahl EM, Davatzikos C, Bottlender R, Frodl T, Scheuerecker J, *et al.* (2009): Use of Neuroanatomical Pattern Classification to Identify Subjects in At-Risk Mental States of Psychosis and Predict Disease Transition. *Arch Gen Psychiatry*. 66: 700.

30. Khodayari-Rostamabad A, Hasey GM, MacCrimmon DJ, Reilly JP, Bruin H de (2010): A pilot study to determine whether machine learning methodologies using pre-treatment electroencephalography can predict the symptomatic response to clozapine therapy. *Clin Neurophysiol*. 121: 1998–2006.

31. Koutsouleris N, Borgwardt S, Meisenzahl EM, Bottlender R, Möller H-J, Riecher-Rössler A (2012): Disease Prediction in the At-Risk Mental State for Psychosis Using Neuroanatomical Biomarkers: Results From the FePsy Study. *Schizophr Bull*. 38: 1234–1246.

32. Mourão-Miranda J, Reinders AATS, Rocha-Rego V, Lappin J, Rondina J, Morgan C, *et al.* (2012): Individualized prediction of illness course at the first psychotic episode: a support vector machine MRI study. *Psychol Med*. 42: 1037–1047.

33. Koutsouleris N, Riecher-Rössler A, Meisenzahl EM, Smieskova R, Studerus E, Kambeitz-Ilankovic L, *et al.* (2015): Detecting the Psychosis Prodrome Across High-Risk Populations Using Neuroanatomical Biomarkers. *Schizophr Bull*. 41: 471–482.

34. Kambeitz-Ilankovic L, Meisenzahl EM, Cabral C, von Saldern S, Kambeitz J, Falkai P, *et al.* (2016): Prediction of outcome in the psychosis prodrome using neuroanatomical pattern classification. *Schizophr Res*. 173: 159–165.

35. Ramyead A, Studerus E, Kometer M, Uttinger M, Gschwandtner U, Fuhr P, Riecher-Rössler A (2016): Prediction of psychosis using neural oscillations and machine learning in neuroleptic-naïve at-risk patients. *World J Biol Psychiatry*. 17: 285–295.

36. Nieuwenhuis M, Schnack HG, van Haren NE, Lappin J, Morgan C, Reinders AA, *et al.* (2017): Multi-center MRI prediction models: Predicting sex and illness course in first episode psychosis patients. *NeuroImage*. 145: 246–253.

37. de Wit S, Ziermans TB, Nieuwenhuis M, Schothorst PF, van Engeland H, Kahn RS, *et al.* (2017): Individual prediction of long-term outcome in adolescents at ultra-high risk for

psychosis: Applying machine learning techniques to brain imaging data: Individual Outcome Prediction With MRI. *Hum Brain Mapp*. 38: 704–714.

38. Koutsouleris N, Wobrock T, Guse B, Langguth B, Landgrebe M, Eichhammer P, *et al.* (2017): Predicting Response to Repetitive Transcranial Magnetic Stimulation in Patients With Schizophrenia Using Structural Magnetic Resonance Imaging: A Multisite Machine Learning Analysis. *Schizophr Bull*. . doi: 10.1093/schbul/sbx114.
39. Fleck DE, Ernest N, Adler CM, Cohen K, Eliassen JC, Norris M, *et al.* (2017): Prediction of lithium response in first-episode mania using the LITHium Intelligent Agent (LITHIA): Pilot data and proof-of-concept. *Bipolar Disord*. 19: 259–272.
40. Ahmadlou M, Rostami R, Sadeghi V (2012): Which attention-deficit/hyperactivity disorder children will be improved through neurofeedback therapy? A graph theoretical approach to neocortex neuronal network of ADHD. *Neurosci Lett*. 516: 156–160.
41. Kim J-W, Sharma V, Ryan ND (2015): Predicting Methylphenidate Response in ADHD Using Machine Learning Approaches. *Int J Neuropsychopharmacol*. 18: pyv052.
42. Ishii-Takahashi A, Takizawa R, Nishimura Y, Kawakubo Y, Hamada K, Okuhata S, *et al.* (2015): Neuroimaging-Aided Prediction of the Effect of Methylphenidate in Children with Attention-Deficit Hyperactivity Disorder: A Randomized Controlled Trial. *Neuropsychopharmacology*. 40: 2676–2685.
43. Plitt M, Barnes KA, Wallace GL, Kenworthy L, Martin A (2015): Resting-state functional connectivity predicts longitudinal change in autistic traits and adaptive functioning in autism. *Proc Natl Acad Sci*. 112: E6699–E6706.

44. Yang D, Pelphrey KA, Sukhodolsky DG, Crowley MJ, Dayan E, Dvornek NC, *et al.* (2016): Brain responses to biological motion predict treatment outcome in young children with autism. *Transl Psychiatry*. 6: e948.

45. Schnack HG, Kahn RS (2016): Detecting Neuroimaging Biomarkers for Psychiatric Disorders: Sample Size Matters. *Front Psychiatry*. 7. doi: 10.3389/fpsyt.2016.00050.

46. Dluhoš P, Schwarz D, Cahn W, van Haren N, Kahn R, Španiel F, *et al.* (2017): Multi-center machine learning in imaging psychiatry: A meta-model approach. *NeuroImage*. 155: 10–24.

47. Thompson PM, Stein JL, Medland SE, Hibar DP, Vasquez AA, Renteria ME, *et al.* (2014): The ENIGMA Consortium: large-scale collaborative analyses of neuroimaging and genetic data. *Brain Imaging Behav*. . doi: 10.1007/s11682-013-9269-5.

48. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, *et al.* (2015): UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLOS Med*. 12: e1001779.

49. First M, Botteron K, Castellanos FX, Dickstein DP, Drevets WC, Kim KL, *et al.* (n.d.): Consensus Report of the APA Work Group on Neuroimaging Markers of Psychiatric Disorders. .

50. PSYSCAN | Translating neuroimaging findings from research into clinical practice (n.d.): . Retrieved November 28, 2017, from http://www.psyscan.eu/.

51. PRONIA - FP7 Research Project (n.d.): . Retrieved November 28, 2017, from https://www.pronia.eu/.

52. Mendelson AF, Zuluaga MA, Lorenzi M, Hutton BF, Ourselin S (2017): Selection bias in the reported performances of AD classification pipelines. *NeuroImage Clin*. 14: 400–416.

53. Schnack HG (2017): Improving individual predictions: Machine learning approaches for detecting and attacking heterogeneity in schizophrenia (and other psychiatric diseases). *Schizophr Res*. . doi: 10.1016/j.schres.2017.10.023.

54. Donini M, Monteiro JM, Pontil M, Shawe-Taylor J, Mourao-Miranda J (2016): A multimodal multiple kernel learning approach to Alzheimer's disease detection. IEEE, pp 1–6.

55. Filippone M, Marquand AF, Blain CRV, Williams SCR, Mourão-Miranda J, Girolami M (2012): PROBABILISTIC PREDICTION OF NEUROLOGICAL DISORDERS WITH A STATISTICAL ASSESSMENT OF NEUROIMAGING DATA MODALITIES. *Ann Appl Stat*. 6: 1883–1905.

56. Tong T, Gray K, Gao Q, Chen L, Rueckert D (2017): Multi-modal classification of Alzheimer's disease using nonlinear graph fusion. *Pattern Recognit*. 63: 171–181.

57. Meng X, Jiang R, Lin D, Bustillo J, Jones T, Chen J, *et al.* (2017): Predicting individualized clinical measures by a generalized prediction framework and multimodal fusion of MRI data. *NeuroImage*. 145: 218–229.

58. Zhang D, Shen D (2012): Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in Alzheimer's disease. *NeuroImage*. 59: 895–907.

59. Zhang D, Wang Y, Zhou L, Yuan H, Shen D (2011): Multimodal classification of Alzheimer's disease and mild cognitive impairment. *NeuroImage*. 55: 856–867.

60. Wan J, Zhang Z, Yan J, Li T, Rao BD, Fang S, *et al.* (2012): Sparse Bayesian multi-task learning for predicting cognitive outcomes from neuroimaging measures in Alzheimer's disease. *2012 IEEE Conf Comput Vis Pattern Recognit*. Presented at the 2012 IEEE Conference on Computer Vision and Pattern Recognition, pp 940–947.

61. Rahim M, Thirion B, Bzdok D, Buvat I, Varoquaux G (2017): Joint prediction of multiple scores captures better individual traits from brain images. *NeuroImage*. 158: 145–154.

62. Young J, Modat M, Cardoso MJ, Mendelson A, Cash D, Ourselin S (2013): Accurate multimodal probabilistic prediction of conversion to Alzheimer's disease in patients with mild cognitive impairment. *NeuroImage Clin*. 2: 735–745.

63. Filipovych R, Resnick SM, Davatzikos C (2011): Multi-Kernel Classification for Integration of Clinical and Imaging Data: Application to Prediction of Cognitive Decline in Older Adults. In: Suzuki K, Wang F, Shen D, Yan P, editors. *Mach Learn Med Imaging*. (Vol. 7009), Berlin, Heidelberg: Springer Berlin Heidelberg, pp 26–34.

64. Hinrichs C, Singh V, Xu G, Johnson SC (2011): Predictive markers for AD in a multi-modality framework: An analysis of MCI progression in the ADNI population. *NeuroImage*. 55: 574–589.

65. Owen MJ (2014): New Approaches to Psychiatric Diagnostic Classification. *Neuron*. 84: 564–571.

66. Barch DM (2017): Biotypes: Promise and Pitfalls. *Biol Psychiatry*. 82: 2–3.

67. Mourão-Miranda J, Hardoon DR, Hahn T, Marquand AF, Williams SCR, Shawe-Taylor J, Brammer M (2011): Patient classification as an outlier detection problem: An application of the One-Class Support Vector Machine. *NeuroImage*. 58: 793–804.

68. Evgeniou T, Micchelli CA, Pontil M (2005): Learning Multiple Tasks with Kernel Methods. *J Mach Learn Res*. 6: 615–637.

Figure captions

Fig. 1 – Pattern recognition and error. The x- and y-axes represent two simulated brain measures where patients (red squares) differ from controls (black circles) at the group level. Train (top row) and test (bottom row) data sets each consisted of 20 patients and 20 controls. For more details on the simulation procedure, see the supplementary material. The top row depicts training data and bottom row represents test data, both sets consist of 20 patients and 20 controls Red and blue shading in the top left panel represent the probability of obtaining a score in that region for patients and controls respectively. The optimal classifier is represented with a solid line, where we classify everything to the right of this line as patients. In the bottom-left panel, this bound is applied to training data and the resulting represent irreducible error errors (indicated with crosses). Middle and right columns illustrate under- and overfitting models respectively (dashed lines). The underfitted model was restricted to have only one non-zero weight (as might happen in L1-regularization). As a result, the decision boundary was forced to be either vertical or horizontal. This results in lower performance in both training and testing. The overfitted model performs better on the training data than the optimal model, but worse on the test data. This is due to the model incorporating spurious patterns occurring by chance in the training data.

Fig. 2 – Nested cross-validation. This figure illustrates nested *k*-fold cross-validation (CV). The basic idea of *k*-fold CV is to split the data into k non-overlapping test-sets (blue sections). In each fold, a model is trained on the training set (white sections) and applied to the test set. Such a procedure can be repeated for different parameter settings to obtain performance estimates for each setting. Selecting this setting, however, invalidates the performance estimate, as the

test set is no longer independent. The solution to this is to perform this optimization inside each of the k_o outer folds. The training data of the outer fold is now split into k_i inner test-sets (purple sections) and the train/test procedure is repeated for each inner fold and parameter setting. Using the performance on the inner test-sets, one can select the optimal parameter settings and retrain the model on the outer training set. Importantly, each of the outer folds might reach different conclusions about optimal parameter settings. Even if they all agree, selecting any specific model from among the k_o models based on their performance will invalidate the performance estimate.

Fig.3 – Illustration of classification versus regression. Panel A plots simulated data: Global Assessment of Functioning (GAF) score as a function of some arbitrary feature. The solid line represents a linear regression model and the dashed lines denote GAF-thresholds for diagnosis. The horizontal line is a threshold applied to the observed data at a GAF-score of 65. The vertical line corresponds to the same threshold applied to predicted GAF-score. Red dots indicate where the predicted diagnosis disagrees with "true" diagnosis. Panel B depicts the same data (with the same color coding), but after thresholding the observed scores. The dashed, horizontal line is the same as in panel A. Note that there is no improvement to be had by shifting the threshold, meaning that regression and classification achieve the same goals.

Fig. 4 – Prediction accuracy of outcome/transition versus sample size (N). Accuracy as a function of sample size is plotted for MDD (A) and psychosis (B), based on the studies listed in **Table 1.** Markers denote different choices for dependent (e.g., treatment response, transition) and independent (i.e., image modality: e.g., MRI, EEG) variables. The line in panel A represents a

linear regression model. Panel C shows the variation in prediction accuracy per study, for studies reporting the performances of several modeling approaches (e.g., different feature sets, different training algorithms).

Table

Study	Output	Follow-up	Data	Method	Validation	N	Result
Major depressive disorder (MDD)							
Cost2009 (12)	Medication response: y/n CBT response: y/n	8 wk 16 wk	sMRI gm-VBM	lin SVM	LOO-CV	9/9 6/6	88.9% n.s.
Khod2010 (13)	SSRI response: y/n	6 wk	EEG	KPLSR	L2O-CV	8/16	86.6%
Khod2011 (14)	TMS response: y/n	2 wk	EEG	MFA	L2O-CV	9/18	80%
Gong2011 (15)	Antidepr. resp.: y/n	12 wk	sMRI gm sMRI wm	lin SVM	LOO-CV	23/23	69.6 65.2%
Korg2014 (16)	Treatment resp.: y/n	8 wk	DTI-FA DTI-FA+age		CV	74	62% 74%
Mill2013 (17)	SSRI response: HDRS continuous	8 wk	fMRI emo. wrd proc.	regression	10f-CV	17	r^2=0.48
Ergu2015 (18)	rTMS response: y/n	4 wk	EEG	ANN	6f-CV	30/25	89%
Pate2015 (19)	Treatment resp.: y/n (late-life depression)	12 wk	sMRI+rsfMRI+DTI	ADTree L1-LR lin SVM RBF SVM	nLOO-CV	9-11/10-13	89.5% 84% 80% 68%
Schm2015 (20)	Outcome: chronic/ gradual improvment/ fast remission	2 y	fMRI emo. fac. expr. fMRI exec.function. sMRI gm	GPC	LOO-CV (3-cen)	23/36/59	73% 51% n.s. 42% n.s.
Will2015 (21)	Antidepressant response: y/n Remission: y/n	8 wk	fMRI +age,symptom severity	discriminant analysis clas.	LOO-CV	48/32 37/43	77% n.s.
vanW2015 (22)	ECT response: y/n	2 wk	rsfMRI netw1 rsfMRI netw2	lin SVM	L2O-CV	25/20	84.5% 78%
Gold2016	Antidepres. func. remission: y/n	8 wk	fMRI+ELS	hierar. LR	LOO-CV	19/51	81%

(23)							
Redl1016 (24)	ECT response: y/n (acute MDD) HDRS continuous score	6.6 (2.3) wk	sMRI gm-VBM	lin SVM GPC lin SVR	LOO-CV	13/10 23	75% 70% r=0.67
AlKa2016 (26)	Improved cognition: y/n Improved mood: y/n	3 wk	EEG	SVM+LDA+ELM	CV	10	90% 70%
Mumt2017 (25)	Treatment response: y/n	4 wk	EEG WT EEG STFT EEG EMD EEG 3-combined EEG 5 lit.based	LR	10f-CV	34	87.5% 80% 72.5% 91.6% 54.5-74.2%
Drys2017 (27)	rTMS response: y/n	4-6 wk	rsfMRI fc rsfMRI+subtyping	lin SVM	LOO-CV (m-cen) +test	70/54 70/54+30	78.3% 89.6%/87.4%
Cran2017 (28)	Antidepres. Resp.: HDRS change	10 wk	fMRI+HDRS,educ.	RF	LOO-CV	29	82%
Psychosis							
Kout2009 (29)	ARMS: Transition: y/n	4 y	sMRI	RBF SVM	5f-CV (+HCtest)	15/18	87%
Khod2010 (30)	SZ: Clozapine resp.: y/n	≥1 y	EEG	KPLSR	CVs+test	23+14	85%
Kout2012 (31)	At risk: Transition: y/n	3-7 y	sMRI	RBF SVM	nCV	16/21	84.2%
Mour2012 (32)	FE-psych.: Illn. course: cont/epis.	6.2 y	sMRI	lin SVM	L2O-CV	28/28	69.5%
Kout2015 (33)	HR: Transition: y/n	≥4 y	sMRI	lin νSVM	CV (2-cen)+test	33/33+7n	80.4%
Kamb2016 (34)	ARMS: GAF outc. good/poor	4 y	sMRI cortSA	L1-reg. LR	nCV	14/13	82%

Ramy2016 (35)	CHR: Transition: y/n	≥3 y	EEG CSD EEG LPS	L1-reg. LR	nCV	18/35	70.5% 58%
Nieu2017 (36)	FE-psych.: Illn. course: cont/rem.	3–7 y	sMRI	lin SVM	CV (5-cen)	94/118	52% (n.s.)
deWi2017 (37)	UHR: Resilience: y/n Resilience: mGAF continuous	(at) 6 y	sMRI LGI sMRI subcVol LGI+subcVol+SIPS-D	lin SVM	CV	17/24	73% r=0.42 82%
Kout2017 (38)	SZ: rTMS response active: y/n rTMSresponse sham: y/n	3 wk	sMRI gm-VBM	lin SVM	nCV/LSO (3-cen) nCV	45 47	85% / 71% 51% (n.s.)
Bipolar disorder							
Flec2016 (39)	FE-bipo.mania: Lithium resp.: y/n	8 wk	fRMI+^{1}H-MRS	GFT 8 other meth.	4f-CV	15/5	80% 61–83.9%
Attention deficit hyperactivity disorder (ADHD)							
Ahma2012 (40)	Neurofeedback response: y/n	12-13 wk	EEG	LDA	100x60/40%t/t	15/15	84.2%
Kim2015 (41)	methylphenidate resp.: y/n	8 wk	rsfMRI+ genetic,environm., neuropsych., demo.,clinical	SVM LRR decis. tree RF	10f-CV	48/30	84.6% 76.9% 69.2% 73.1%
Ishi2015 (42)	methylphenidate resp: CGI ≤3/≥4	4-8 wk	NIRS Δ[oxy-Hb]	LDA	LOO-CV	16/5	81%
Autism spectrum disorder (ASD)							
Plit2015 (43)	adaptive behaviors: ΔABAS autistic traits: ΔSRS	>1 y, mean: 2.8 y	rsfMRI fc	RR	nLOO-CV	27 29	r^2=0.32 r^2=0.34
Yang2016 (44)	PRT response: ΔSRS	16 wk	fMRI	KRR	LOO-CV	20	R^2=0.72

For clarification of the abbreviations, see Supplemental Materials. The numbers (N) are the numbers of patients in the categories reported in the 'Output' column. The main performance (accuracy, correlation coefficient) displayed in the last column is the performance presented (in the article's Abstract) as the most important result; in case of models that combine imaging with non-

imaging data, the model incorporating the least amount of non-imaging features is chosen. If results of more than one model are available, these are presented below the main model's performance. Redl2016: Reported accuracies (resp. 78.3 and 73.9%) were converted to balanced accuracies: 75 and 70%, resp. Ramy2016: Accuracies (70.5% and 0.58%) were estimated from the reported AUCs (0.77 and 0.56, resp.).

Figure 1

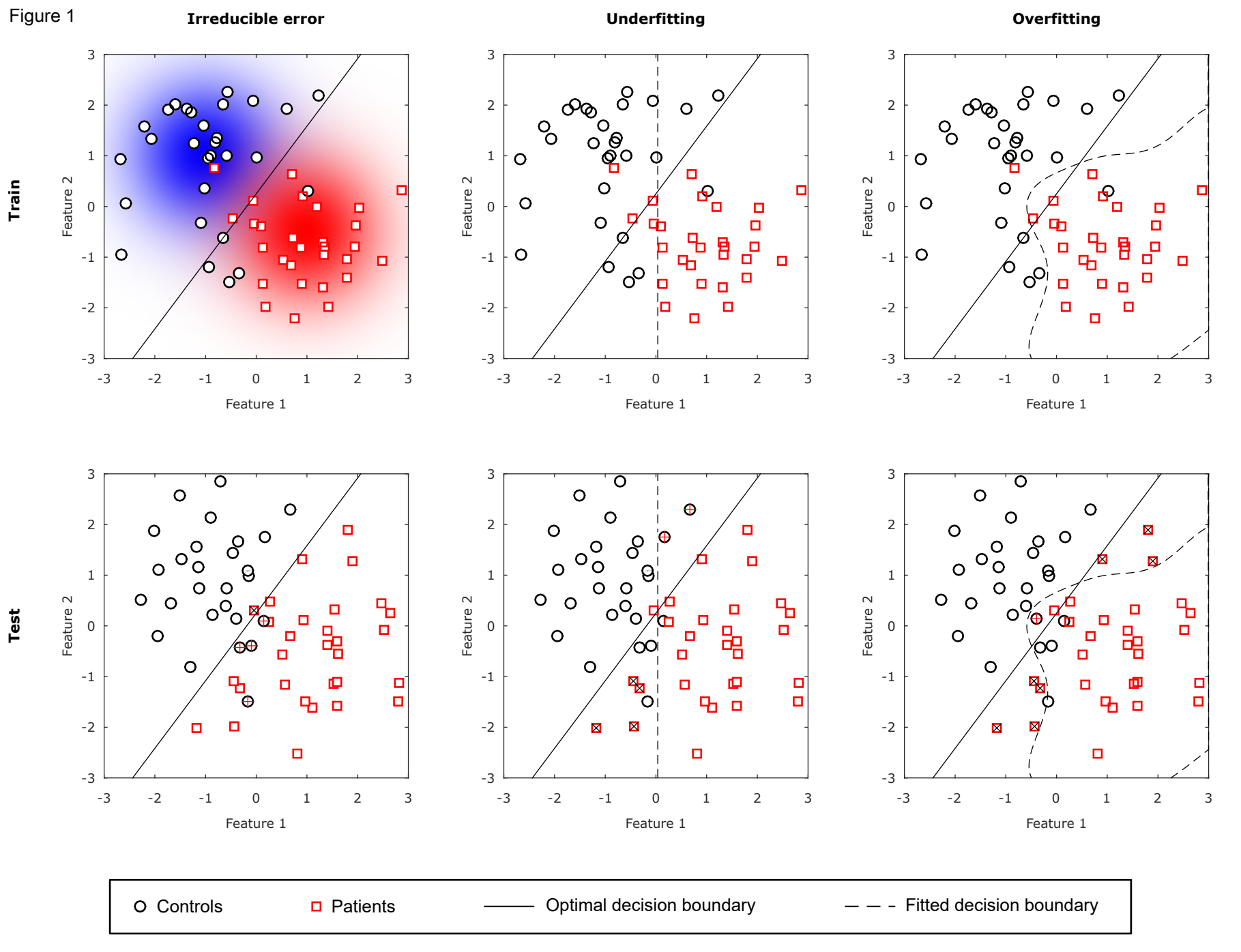

Irreducible error
Underfitting
Overfitting
Train
Test
Feature 1
Feature 2
Controls
Patients
Optimal decision boundary
Fitted decision boundary

Figure 2

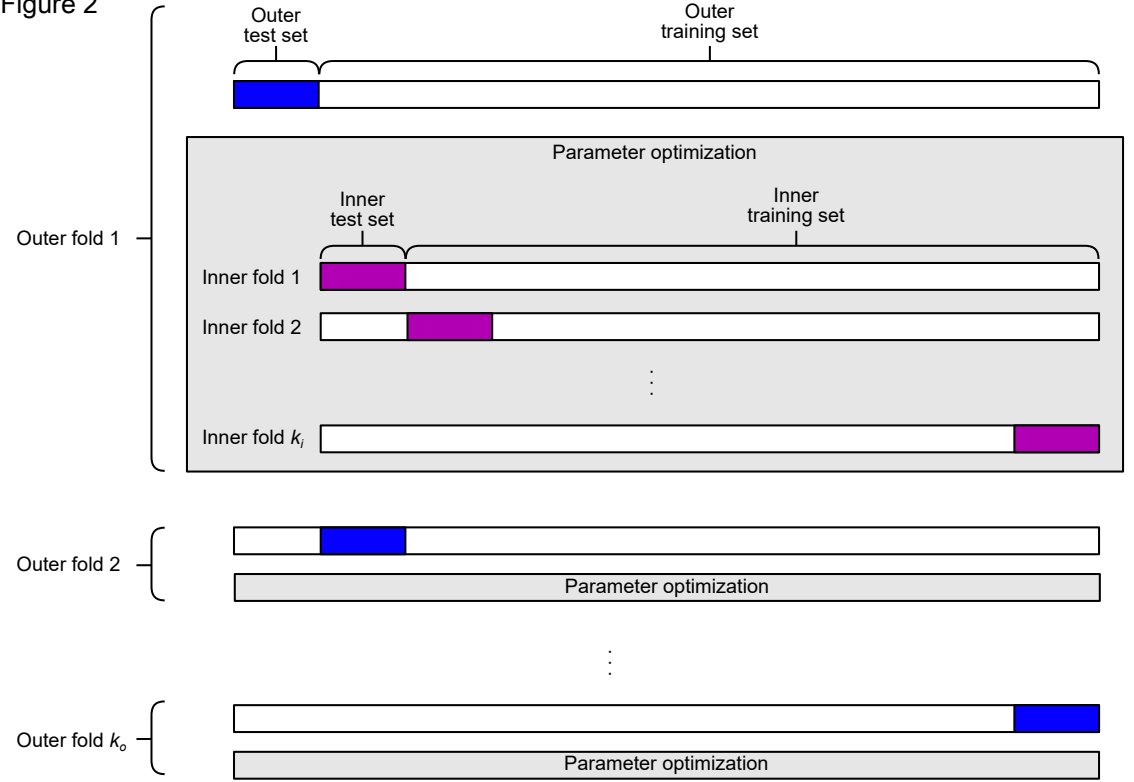

Outer test set
Outer training set
Parameter optimization
Inner test set
Inner training set
Outer fold 1
Inner fold 1
Inner fold 2
Inner fold k_i
Outer fold 2
Parameter optimization
Outer fold k_o
Parameter optimization

Figure 3

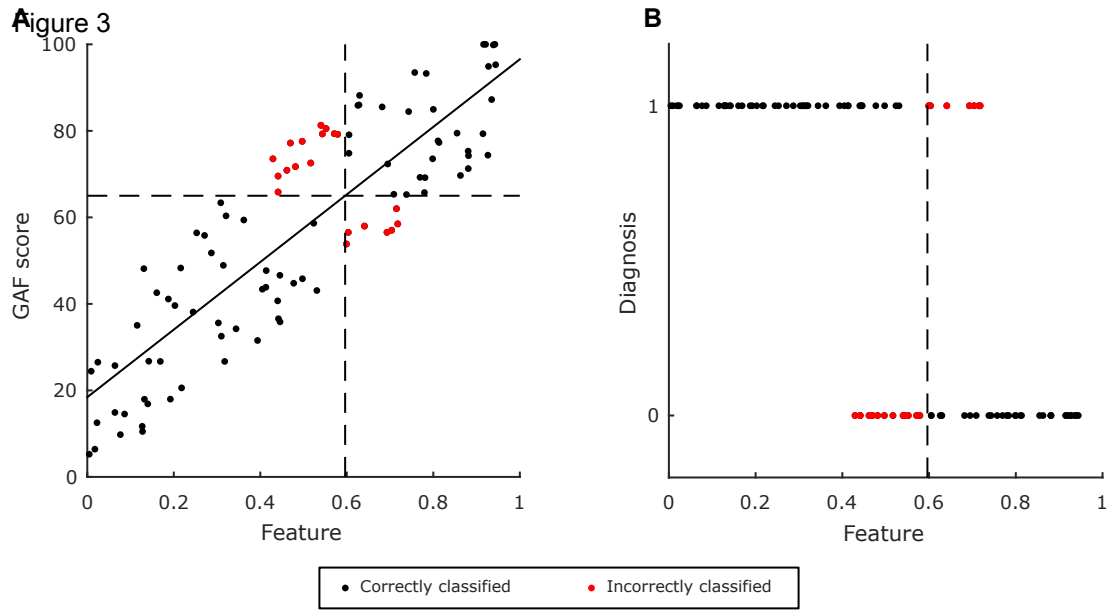
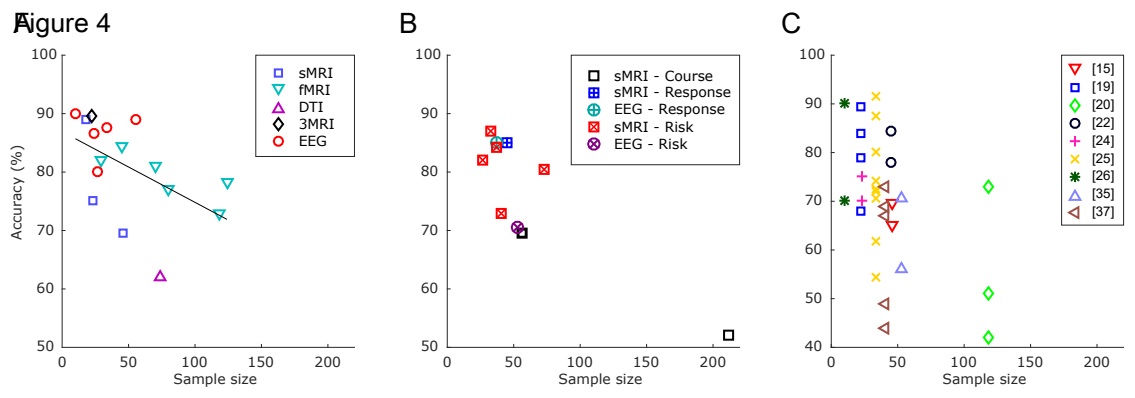

A
B
GAF score
Feature
Diagnosis
Feature
Correctly classified
Incorrectly classified

Figure 4


A
B
C
Accuracy (%)
Sample size
sMRI
fMRI
DTI
3MRI
EEG
sMRI - Course
sMRI - Response
EEG - Response
sMRI - Risk
EEG - Risk
[15]
[19]
[20]
[22]
[24]
[25]
[26]
[35]
[37]

Making Individual Prognoses in Psychiatry Using Neuroimaging and Machine Learning

Supplementary Information

In this supplementary material, we provide some additional background to concepts discussed in the paper.

Measuring performance

Numerous performance measures exist for either regression- or classification-type models. Some of the more common ones are presented in Supplementary Table S1. In this section, we briefly discuss these performance measures.

Classification

Sensitivity and specificity represent the probability of correctly classifying a participant as positive or negative, respectively. Positive predictive value (PPV) and negative predictive value (NPV) are more useful for the clinician, however, as they represent the probability of a given positive or negative prediction being true. PPV and NPV do come with the caveat that prevalence *in the population that would receive the test* matters a great deal. The nature of prognostic studies helps in this regard, but one should be aware that dropout might affect this.

While the abovementioned performance measures are useful in deciding on applicability/generalizability of a prediction model, parameter optimization requires a summary of overall performance. Accuracy does so by computing the proportion of correctly classified cases. Intuitively, with a binary classifier, one would expect 50% accuracy to reflect

chance level. This is no longer true if the classes are not balanced, which is exceedingly likely in prospective studies. Hence, it is necessary to either adjust what one considers chance level or correct the accuracy estimate. Balanced accuracy is an example of the latter option.

Regression

Performance for regression-type models amounts to computing the (scaled) distance between prediction and observation. The variety in performance measures is in part due to different choices for the distance metric.

Mean-Squared-Error (MSE) and Root-Mean-Square-Error (RMSE) are error metrics often employed in the machine learning literature. In the absence of systematic error, they correspond to the variance and standard deviation of the residuals, respectively.[1] Mean-Absolute-Error (MAE) is based on the taxicab distance metric, as a result MAE is less sensitive to outliers. The scale of these three measures depends on the scale of the observations; this means that they cannot be compared between studies without correcting for this scale.

Another common metric for accuracy is the coefficient of determination (R^2). As this metric takes the variance of the observations into account, it is more easily compared across studies. There are, however, various ways of computing this coefficient and these are only equivalent for certain cases (1). One common way of calculating R^2 is by computing the squared correlation coefficient between predictions and observations. In the definition in Supplementary Table S1, we follow the recommendation of Kvalseth (1) and use a definition based on the residual variance. A negative R^2 with this definition, corresponding to a negative

[1] They differ in that the sample variance scales the sum-of-squares by $N - 1$, whereas MSE scales by N. Systematic errors are reflected in (R)MSE, but not in residual variance. In many regression problems, however, this is not an issue.

correlation between prediction and observation, can be indicative of serious methodological problems. In classification such a phenomenon is known as anti-learning (2).

Supplementary Table S1: Performance measures

	Definition	Notes
Classification		
Sensitivity (Se)	$\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}$	
Specificity (Sp)	$\frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}}$	
Positive Predictive Value (PPV)	$\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}$	Reliability of positive predictions
Negative Predictive Value (NPV	$\frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FN}}$	Reliability of negative predictions
Accuracy (A)	$\frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}}$	
Balanced Accuracy (BA)	$\frac{\mathrm{Se}+\mathrm{Sp}}{2}$	Accuracy corrected for class imbalance
Regression		
Mean Absolute Error (MAE)	$N^{-1}\sum_{i=1}^{N}\lvert y_i-\hat{y}_i\rvert$	Penalizes outliers
Mean Squared Error (MSE)	$N^{-1}\sum_{i=1}^{N}(y_i-\hat{y}_i)^2$	Proportional to residual variance
Root MSE (RMSE)	$\sqrt{\mathrm{MSE}}$	
Coefficient of determination (R^2)	$1-\frac{\sum_{i=1}^{N}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{N}(\hat{y}_i-\bar{y})^2}$	

Abbreviations not defined in the table: TP, number of true positives; TN, number of true negatives; FP, number of false positives; FN, number false negatives. In the equations pertaining to regression: N is the number of participants, y_i is the outcome of the i^{th} participant, $\hat{y}_i$ is the predicted outcome and $\bar{y}$ is the mean outcome.

On model selection

Optimizing hyperparameters using cross-validation (CV) involves training models with various settings and selecting a setting that maximizes performance. This can be seen as fitting a model on the combined train/test set. As a result, this procedure is itself also vulnerable to overfitting. While the effect on performance estimates can be corrected for using nested CV, it might still hamper the optimization itself.

Breiman (3) proposed the 1-SE rule to ameliorate this issue. Under this rule, one would select the least-complex model within one standard error of the highest performance score. While this was proposed in the context of decision trees, it can be applied equally well to any model selection setting where models can be ranked by complexity. This might be especially interesting when performing feature selection.

It should be noted, however, that the 1-SE bound is arbitrary. If the training algorithm is particularly vulnerable to overfitting, a looser bound may be required. On the other hand, if the overfitting is minimal, a tighter bound might be required to avert underfitting.

Simulations Fig. 1

To illustrate the effects of under- and overfitting, a toy example was simulated. The simulated dataset consisted of train and test sets, each containing 20 cases labeled as controls and another 20 as patients. For each case, the observations for two features $[x_1, x_2]$ were drawn from a bivariate Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}, \mathbf{I})$, where $\mathbf{I}$ is the identity matrix and $\boldsymbol{\mu} = [-1, 1]$ for controls and $\boldsymbol{\mu} = [1, -.5]$ for patients.

The optimal classifier bound for the ground truth is defined by those values of $\boldsymbol{x} = [x_1, x_2]$ for which $P(\boldsymbol{x}|control) = P(\boldsymbol{x}|patient)$ holds. The underfitted model was

generated by training a linear SVM on the first feature dimension. The overfitted model was produced by training an SVM with a radial basis function.

Simulations Fig. 3

Data were simulated by drawing feature observations from the uniform distribution $\mathcal{U}(0,1)$. Global Assessment of Functioning (GAF) scores were simulated by adding Gaussian noise ($\sigma = 0.5$) and multiplying by 80. Any scores above 100 were set to 100. Diagnostic threshold was set to a GAF-score of 65.

PubMed search criteria

To obtain an overview of the studies that apply machine learning (ML) to neuroimaging data to make prognoses for individual patients with (risk of) a psychiatric disorder, we carried out the following search in PubMed on November 6th, 2017: (outcome OR course OR transition OR response) AND (prediction OR predicting OR machine learning) AND (imaging OR MRI OR EEG OR PET OR SPECT OR NIRS) AND (psychosis OR schizophrenia OR depression OR bipolar OR ADHD OR autism OR ASD). We removed studies that did not make predictions about patients' future status. Studies that did not test their prediction model in unseen cases (either by using an independent test set or, most often, by using cross-validation within the training sample – see section 2.2 of this review) were also excluded. This search resulted in 33 studies: 17 studies for depression (MDD), 10 for psychosis, 1 for bipolar disorder (BD), 3 for ADHD and 2 for ASD (see Table 1).

List of abbreviations

*-cen	multicenter
^{1}H-MRS	proton magnetic resonance spectroscopy
ABAS	adaptive behavior assessment system
ADTree	alternating decision trees
ANN	artificial neural network
CBT	cognitive behavioral therapy
CGI	clinical global impression
clas.	classifier
cortSA	cortical surface area
CSD	current-source density
DTI	diffusion tensor imaging
demo.	demographic
decis. tree	decision tree
educ.	education level
ELM	extreme learning machine
ELS	early life stress
EMD	empirical mode decompositions
emo. fac. expr.	emotional facial expressions
emo. wrd proc.	emotional word processing
environm.	environmental
exec. function.	executive functioning
FA	fractional anisotropy
Fc	functional connectivity
FE	first-episode
gm	gray matter
GFT	genetic fuzzy tree
GPC	Gaussian process classifier
HDRS	Hamilton depression rating scale
hierar.	hierarchic
KPLSR	kernelized partial least squares regression
(K)RR	(kernel) ridge regression
L1-reg.	L1-regularized
LDA	linear discriminant analysis
LGI	local gyrification index
lin	linear
lit.based	literature based
LPS	lagged phase synchronicity
L(R)R	logistic (ridge) regression
LSO	leave-site-out
meth.	method(s)
MFA	mixture of factor analysis
neuropsych.	neuropsychological
NIRS	near-infrared spectroscopy
n.s.	not significant
PRT	pivotal response treatment
RBF	radial basis function

resp.	response
RF	random forests
(rs)fMRI	(resting state) functional MRI
SIPS-D	Structured Interview for Prodromal Symptoms - Disorganization
sMRI	structural MRI
SNRI	serotonin-norepinephrine reuptake inhibitors
SRS	social responsiveness scale
SSRI	Selective serotonin reuptake inhibitor
STFT	short-time Fourier transform
subcVol	subcortical volumes
SVM	support vector machine
SVR	support vector regression
SZ	schizophrenia
t/t	train/test
VBM	voxel-based morphometry
wk	week
wm	white matter
WT	wavelet transform
y	year
y/n	yes/no

Supplemental References

1. Kvalseth TO (1985): Cautionary Note about R2. *Am Stat*. 39: 279–285.

2. Kowalczyk A, Chapelle O (2005): An Analysis of the Anti-learning Phenomenon for the Class Symmetric Polyhedron. In: Jain S, Simon HU, Tomita E, editors. *Algorithmic Learn Theory*. (Vol. 3734), Berlin, Heidelberg: Springer Berlin Heidelberg, pp 78–91.

3. Breiman L, Friedman JH, Olshen RA, Stone CJ (1984): Classification and regression trees. *CERN Doc Serv*. Retrieved November 28, 2017, from http://cds.cern.ch/record/2253780.