HHS Public Access

Author manuscript

Neuroimage. Author manuscript; available in PMC 2020 October 01.

Published in final edited form as:

Neuroimage. 2019 October 01; 199: 351–365. doi:10.1016/j.neuroimage.2019.05.082.

Quantifying performance of machine learning methods for neuroimaging data

Lee Jollans, PhD[1,2], **Rory Boyle, MSc**[1], **Eric Artiges, MD, PhD**[3], **Tobias Banaschewski, MD, PhD**[4], **Sylvane Desrivières, PhD**[5], **Antoine Grigis, PhD**[6], **Jean-Luc Martinot, MD, PhD**[7], **Tomáš Paus, MD, PhD**[8], **Michael N. Smolka, MD**[9], **Henrik Walter, MD, PhD**[10], **Gunter Schumann, MD**[5], **Hugh Garavan**[11], **Robert Whelan**[1,12,*]

[1]School of Psychology, Trinity College Dublin, Dublin, Ireland [2]Max-Planck Institute of Psychiatry, Munich, Germany [3]Institut National de la Santé et de la Recherche Médicale, INSERM Unit 1000 “Neuroimaging & Psychiatry”, University Paris Sud, University Paris Descartes - Sorbonne Paris Cité; and Psychiatry Department 91G16, Orsay Hospital, France [4]Department of Child and Adolescent Psychiatry and Psychotherapy, Central Institute of Mental Health, Medical Faculty Mannheim, Heidelberg University, Square J5, 68159 Mannheim, Germany [5]Medical Research Council - Social, Genetic and Developmental Psychiatry Centre, Institute of Psychiatry, Psychology & Neuroscience, King’s College London, United Kingdom [6]NeuroSpin, CEA, Université Paris-Saclay, F-91191 Gif-sur-Yvette, France [7]Institut National de la Santé et de la Recherche Médicale, INSERM Unit 1000 “Neuroimaging & Psychiatry”, University Paris Sud, University Paris Descartes - Sorbonne Paris Cité; and Maison de Solenn, Paris, France; [8]Bloorview Research Institute, Holland Bloorview Kids Rehabilitation Hospital and Departments of Psychology and Psychiatry, University of Toronto, Toronto, Ontario, M6A 2E1, Canada; [9]Department of Psychiatry and Neuroimaging Center, Technische Universität Dresden, Dresden, Germany [10]Charité – Universitätsmedizin Berlin, corporate member of Freie Universität Berlin, Humboldt-Universität zu Berlin, and Berlin Institute of Health, Department of Psychiatry and Psychotherapy, Campus Charité Mitte, Charitéplatz 1, Berlin, Germany [11]Department of Psychiatry, University of Vermont, Burlington, USA [12]Global Brain Health Institute, Trinity College Dublin, Dublin, Ireland

Abstract

Machine learning is increasingly being applied to neuroimaging data. However, most machine learning algorithms have not been designed to accommodate neuroimaging data, which typically has many more data points than subjects, in addition to multicollinearity and low signal-to-noise.

*corresponding author.

Disclosures: Dr. Banaschewski has served as an advisor or consultant to Bristol-Myers Squibb, Desitin Arzneimittel, Eli Lilly, Medice, Novartis, Pfizer, Shire, UCB, and Vifor Pharma; he has received conference attendance support, conference support, or speaking fees from Eli Lilly, Janssen McNeil, Medice, Novartis, Shire, and UCB; and he is involved in clinical trials conducted by Eli Lilly, Novartis, and Shire; the present work is unrelated to these relationships. The other authors report no biomedical financial interests or potential conflicts of interest.

Consequently, the relative efficacy of different machine learning regression algorithms for different types of neuroimaging data are not known. Here, we sought to quantify the performance of a variety of machine learning algorithms for use with neuroimaging data with various sample sizes, feature set sizes, and predictor effect sizes. The contribution of additional machine learning techniques – embedded feature selection and bootstrap aggregation (bagging) – to model performance was also quantified. Five machine learning regression methods – Gaussian Process Regression, Multiple Kernel Learning, Kernel Ridge Regression, the Elastic Net and Random Forest, were examined with both real and simulated MRI data, and in comparison to standard multiple regression. The different machine learning regression algorithms produced varying results, which depended on sample size, feature set size, and predictor effect size. When the effect size was large, the Elastic Net, Kernel Ridge Regression and Gaussian Process Regression performed well at most sample sizes and feature set sizes. However, when the effect size was small, only the Elastic Net made accurate predictions, but this was limited to analyses with sample sizes greater than 400. Random Forest also produced a moderate performance for small effect sizes, but could do so across all sample sizes. Machine learning techniques also improved prediction accuracy for multiple regression. These data provide empirical evidence for the differential performance of various machines on neuroimaging data, which are dependent on number of sample size, features and effect size.

Keywords

Machine learning; Neuroimaging; Regression algorithms; reproducibility

1. Introduction

An increasing number of projects and consortia are now collecting large neuroimaging datasets. These include IMAGEN (Schumann et al., 2010), , the Alzheimer's Disease Neuroimaging Initiative (ADNI, Jack et al., 2008), the Human Connectome project (Van Essen et al., 2012), ENIGMA (Thompson et al., 2017), the 1000 Functional Connectomes project (Biswal et al., 2010) and the Adolescent Brain Cognitive Development study (ABCD; https://abcdstudy.org/, see Vol. 32 of Developmental Cognitive Neuroscience, which is dedicated to the ABCD study). In addition, there are data-sharing facilities such as NeuroVault (neurovault.org, Gorgolewski et al., 2015), OpenNeuro (openneuro.org, Gorgolewski et al., 2017), and the Neuroimaging Informatics Tools and Resources Clearinghouse (NITRC; Kennedy, Haselgrove, Riehl, Preuss, & Buccigrossi, 2016). These sources of high-dimensional imaging data offer exciting opportunities to produce generalizable and reproducible research findings in arenas such as predicting disease trajectories, or linking behavioral and personality factors to functional and structural imaging data.

As large samples become more commonplace in neuroimaging, analytical tools developed for data science, such as machine learning, are more frequently applied to neuroimaging data (Jollans & Whelan, 2017; Woo et al., 2017). A wide variety of studies have used machine learning algorithms to classify individuals based on structural or functional imaging data using, among other algorithms, Support Vector Machines (e.g. Costafreda et al., 2009;

Davatzikos et al., 2011; Koutsouleris et al., 2012), Random Forest (e.g. Ball et al. 2014, Ramirez et al., 2010), and Naïve Bayes classifiers (e.g. Adar et al., 2016; Wang, Redmond, Bertoux, Hodges & Hornberger, 2016; Zhou et al., 2015). There have also been successful efforts to predict continuous outcome variables, mostly using Relevance or Support vector regression, such as age (Dosenbach et al., 2010; Franke et al., 2010; Mwangi et al., 2013), cognitive ability (Stonnington et al.,2010), language ability (Formisano et al., 2008), and disease severity in patients with major depression (Mwangi et al., 2012). While these algorithms have been increasingly used in neuroimaging research, none of them were specifically developed for neuroimaging data, which have high dimensionality, inherent multicollinearity, and typically small signal-to-noise ratios. Below we briefly review important considerations when analysing large neuroimaging datasets, and how machine learning methods may address these issues.

1.1. Outcome prediction

Several authors have emphasized the importance of moving away from explanatory and univariate analysis procedures and towards multivariate outcome prediction in psychology and neuroscience (Gabrieli et al., 2015; Jollans & Whelan, 2016; Poldrack, 2011; Westfall & Yarkoni, 2016). Using regression approaches, effective outcome prediction requires that accurate outcome estimations can be achieved for new (i.e., unseen) cases. Prediction models exploit between-subject heterogeneity to make individual-level predictions rather than utilizing differences in group means (Lo et al., 2015). Embracing machine learning for outcome prediction would significantly contribute to the generalizability and reproducibility of neuroimaging research, and improve the ability of neuroimaging to explore individual differences (Dubois & Adolphs, 2016). There are several methods used to estimate and improve the generalizability of a regression model. Most common among these is cross-validation (CV) in which the data are split into ‘training’ and ‘test’ sets. Models are developed using only the training set, and model performance is assessed using the test set. The training and test set must be kept separate for all analysis steps (Cawley & Talbot, 2010). Typically, this split is carried out multiple times, alternating the data that are included in the test set. While some neuroimaging studies use test sets comprised of only one observation (leave-one-out CV; e.g. Brown et al., 2012; Clark et al., 2014; Duff et al., 2012; Niehaus et al., 2014), larger test sets (leave-k-out, where k typically equals 5 or 10; e.g. Wang et al., 2013; Whelan et al., 2014) are preferable as they provide more accurate model performance estimates (Kohavi, 1995). CV can also be used to provide an out-of-sample estimate of model performance within the analysis pipeline itself, in order to optimize parameters for the regression model. When multiple layers of CV are used for internal and external validation of model performance this is referred to as ‘nested’ CV (see Figure 1). By embedding a layer of CV within the training set of the ‘outer layer’ of CV it is made possible to train and validate a model within the training set itself. The test set remains a separate dataset used to carry out a final validation of model performance, removing the need for a separate validation set. The reader is referred to Varoquaux et al., 2017 for an empirical investigation of CV on neuroimaging data. In large datasets comprised of data from multiple sites, generalizability can be further quantified by using leave-site-out CV, where data from one site is withheld as a test set and the model is developed using data from the remaining sites (Dwyer, Falkai, & Koutsouleris, 2018). Assessing model performance on

the withheld test set is then a more rigorous test of generalizability and the obtained measure of generalizability can be further optimized using nested leave-site-out CV (Dwyer et al., 2018). Such complex CV techniques have been used to build generalizable and accurate prediction models of treatment outcomes in psychosis using multisite psychosocial, sociodemographic, and psychometric data (Koutsouleris et al., 2016) and a combination of clinical and neuroimaging data (Koutsouleris et al., 2018).

1.2. Prediction with neuroimaging data

Depending on the voxel size, single MRI images can contain from 100,000 to a million voxels. As sample sizes in neuroimaging are often modest even very large studies will have more voxels than participants. A higher ratio of features to cases increases the tendency of the model to fit to noise in that sample (i.e., *overfitting*; see Whelan & Garavan, 2014 for a discussion specific to neuroimaging). Overfitting will result in the model fitting poorly when it is applied to a new dataset. Even when using a smaller number of regions of interest (ROIs) instead of voxels, combining multiple data sources (such as neuroimaging data and cognitive data or demographics), imaging modalities, or conditions will result in a large number of features. Feature selection and regularization are two approaches that are commonly adopted for dealing with high-dimensional data. A further potentially useful method for neuroimaging data is bootstrap aggregation (bagging).

1.2.1. Dimension reduction.—Reducing the number of features in a regression model (i.e., dimension reduction), will almost always be beneficial for attenuating overfitting when working with neuroimaging data. There exists a wide array of methods for reducing the number of input variables in neuroimaging data (Mwangi, Tian & Soares, 2014). These methods work by selecting a subset of features or by summarizing features in new variables. Some of these methods, such as principal and independent component analysis (PCA and ICA), have long been standard tools in neuroscience. Dimension reduction techniques can be separated by whether they preserve the original values of features (this is not the case for ICA and PCA), whether they consider each feature in isolation or not, and whether they are unsupervised (using only the feature values) or supervised (using the feature and dependent variable values). *Feature selection* is a dimension reduction technique that is often favored in neuroimaging studies that use machine learning approaches. Feature selection is an umbrella term for supervised methods that do not alter the original feature values. Feature selection methods can broadly be categorized into ‘filter’ methods, ‘wrapper’ methods, and embedded methods (see Chandrashekar and Sahin, 2014). Filter methods are unimodal, considering each feature individually. The application of filter methods involves evaluating the outcome of each feature on some statistical test (e.g., a t-test or Pearson’s correlation with the outcome variable), and only retaining those features with the highest values. A key benefit of filter methods is low computational cost when compared to much more computationally expensive wrapper methods (Nnamoko et al., 2014), which are multimodal and consider subsets of features. Popular wrapper methods include forward selection, backward elimination, and recursive feature elimination, all of which carry out step-wise search procedures including or excluding features in each step to arrive at the feature set which maximizes algorithm performance. Wrapper methods lend themselves well to embedding within optimization of the regression model (e.g. an adaptive forward-backward greedy

algorithm integrated within a model; Jie et al., 2015). Embedded methods integrate feature selection directly into optimization of the regression model by choosing the feature selection criterion through hyperparameters. The most widely used embedded feature selection methods use regularization (discussed further below). A key advantage of embedded methods is that any researcher input regarding minimum effect sizes or desired feature set size which is typically necessary in filter and wrapper methods is eliminated, reducing the possible bias introduced to the model at this step. Novel wrapper methods, such as an adaptive forward-backward greedy algorithm, can also be integrated within models (Jie et al., 2015). Sophisticated pipelines can combine feature selection techniques and other dimension reduction methods. For example, Koutsouleris et al. (2018) implemented principal components analysis to reduce the dimensionality of MRI data and then used a wrapper method, specifically a greedy sequential backward elimination algorithm, to identify the principal components that optimally predicted the outcome.

In neuroimaging, good outcome predictions may rely on large feature sets, as any cognitive or behavioral variable of interest will most likely be best explained by a network of spatially correlated brain regions. Good regression models with neuroimaging data may therefore include interaction effects between features. To account for this, the feature selection methods that should be used with neuroimaging data will consider feature sets rather than individual features. Accordingly, previous work has shown that both wrapper methods (Tangaro et al., 2015) and embedded methods (Tohka, Moradi, Huttunen & ADNI, 2016) are preferable to filter methods with neuroimaging data. However, wrapper methods are sometimes prone to overfitting and are typically more computationally intense than embedded methods (Saeys, Inza, & Larrañaga, 2007). Furthermore, as neuroimaging data have an inherently low signal-to-noise ratio, the individual predictive power of each voxel or ROI is likely to be quite small. It may therefore be advantageous to consider complex regression models that allow for the inclusion of some predictors with low effect sizes. Due to the amount of unknown factors relevant to the selection of a feature selection method (such as the unknown ideal number of features and the optimal threshold for inclusion of features with low effect sizes), the focus of this paper with regard to dimension reduction will be on embedded methods, which can be implemented without much researcher input.

1.2.2. Regularization.—Regularization is a method that attenuates overfitting by penalizing the size of the regression weights as model complexity increases. Regularization is often achieved through the L1- or the L2-norm. The L1-norm, as implemented in the Least Absolute Shrinkage and Selection Operator (LASSO), penalizes regression weights based on their absolute size, and results in sparse models (i.e., some regression weights can be set to zero). The L2-norm (also known as Ridge Regression or Tikhonov Regularization) penalizes regression weights based on their squared values, and does not result in sparse models. However, with highly multicollinear data (such as neuroimaging data) neither L1- nor L2-norm regularization are ideal because the large number of non-zero coefficients in models using the L2-norm is unable to produce parsimonious solutions, and the L1-norm is inadequate in accounting for highly correlated groups of predictors (Ogutu, Schulz-Streeck & Piepho, 2012; Mwangi, Tian & Soares, 2014). The Elastic Net (EN; Zou & Hastie, 2005) combines L1-norm and L2-norm regularization, and has the advantage of being an

embedded feature selection algorithm, and thus produces a sparse solution in which groups of correlated features are included or excluded. The Elastic Net has gained popularity among neuroimaging researchers in recent years, and has been successfully used in several studies with large samples (e.g. Chekroud et al., 2016; Whelan et al., 2014).

1.2.3. Bootstrap aggregation (bagging).—

The low signal-to-noise ratio of neuroimaging data calls for a tool to increase the stability of findings and reduce error in outcome estimates. Stability can be estimated using *bootstrapping* (Efron & Tibshirani, 1997), where the dataset is randomly sampled with replacement many times to minimize the effect of outliers and estimate the true population mean (Hall & Robinson, 2009). Like CV, bootstrapping serves a purely descriptive purpose when used to estimate population metrics. However, a related approach termed bootstrap aggregation (*bagging*; Breiman, 1996), uses bootstrapping to improve stability within the model optimization framework. Bagging uses bootstrapped samples to generate multiple estimates of a calculation or metric, and an aggregate of these estimates is created. These aggregated estimates can be used instead of singular outcome estimates at every step of the analysis. An important application of this method is in unsupervised learning, where stability of clustering applications with neuroimaging data can be greatly improved through bagging (Bellec et al., 2010). Bagging has also previously been used for embedded feature selection with large genetic datasets and showed significant improvements over standard non-bagged embedded methods in terms of model accuracy and stability (Abeel, Helleputte, Peer, Dupont & Saeys, 2010). Bagging is an effective way to decrease error, particularly with datasets that have a low signal-to-noise ratio and high multicollinearity (Zahari, Ramli & Mokhtar, 2014).

1.3. Researcher degrees-of-freedom

Another important consideration in neuroimaging is that flexible or ‘exploratory’ analysis introduces a high risk of false positive results or overestimated effect sizes (Button et al., 2013). Therefore, predetermined analysis pipelines and analytical decisions aid in producing reproducible results. The tendency for researchers to screen data before data collection is completed, to carry out multiple iterations of analyses without reporting the findings (e.g., with and without covariates), or to tweak parameters for group inclusion to better represent the problem has been termed ‘researcher degrees of freedom’ (Simmons, Nelson, & Simonsohn, 2011; Loken & Gelman, 2017; Westfall & Yarkoni, 2016). In the case of machine learning frameworks, the researcher input can potentially be greatly reduced, limiting the room for subjectivity and reducing the researcher degrees of freedom. To enhance objectivity, the role of the researcher should be confined to collecting and preparing the best data possible to describe the problem of interest, based on domain knowledge (Dubois & Adolphs, 2016). Ideally, dimension reduction, model building, and parameter optimization should be data-driven.

1.4. Effect of study design and ML method on model performance

The choice of which ML methods to use, or whether to use them at all, is an important consideration but it is not clearly defined in the literature. A parameter to be defined before commencing any machine learning analysis is the CV framework to use. While leave-one-out CV yields more accurate predictions from neuroimaging data than split-half or two-fold

CV (Price, Ramsden, Hope, Friston, & Seghier, 2013), it also generates more variable and biased estimates of out of sample accuracy (Varoquaux et al., 2017). Ten-fold CV produces more stable accuracy estimates and is recommended (Kohavi, 1995).

The use of feature selection, and the method used, can also impact model performance with neuroimaging data. Some feature selection techniques have little impact on model performance and may only increase computational expense (e.g., classification of individuals with mild cognitive impairment or Alzheimer's disease in Chu, Hsu, Chou, Bandettini, & Lin, 2012). The Elastic Net (embedded method), yields more accurate predictions than filter and wrapper methods for some classification problems (Tohka, Moradi, Huttune & ADNI, 2016). When using embedded methods like the Elastic Net, additional prior dimension reduction steps routinely employed with neuroimaging data (such as initial ROI selection or PCA) likely also become redundant, although this remains to be empirically investigated.

Bagging has been used with neuroimaging data for Alzheimer's disease detection (Shen et al., 2012), discrimination between Alzheimer's disease and mild cognitive impairment (Ramirez et al., 2018) and between Parkinson's disease and atypical Parkinsonian syndrome (Garraux et al., 2013). Bagging outperforms boosting algorithms, another class of sophisticated ensemble technique (Ramírez, Górriz, Ortiz, Padilla, & Martínez-Murcia, 2016). As these studies did not specifically investigate the effect bagging had on the analysis results, and how bagging interacts with algorithm and dimension reduction choices, the use of bagging with neuroimaging data remains to be further examined. Moreover, Munson and Caruana (2009) demonstrated that while bagging can continue to improve model performance with increasing feature set size, performance does eventually plateau for most data. The relationship between feature set size and model performance with bagging has not been formally tested using neuroimaging data. Additionally, Munson and Caruana (2009) reported that feature selection can also reduce model performance when combined with bagging. This potentially negative interaction between feature selection and bagging is also unclear for neuroimaging data.

A final crucial parameter known to impact model performance is sample size. Sample size greatly affects prediction accuracy in ML models using neuroimaging data (Arbabshirani, Plis, Sui, & Calhoun, 2017). The accuracy of ML models in predicting age (Franke et al., 2010) and identifying schizophrenia (Schnack & Kahn, 2016) from neuroimaging data increases with training set size. This is likely because smaller training sets are more heterogenous (Schnack & Kahn, 2016). While some models, such as the Elastic Net (Zou & Hastie, 2005) are relatively robust to smaller sample sizes where the number of predictors is far bigger than the number of observations, it is unclear how changes in the ratio of features to observations impact performance of the Elastic Net and other algorithms with neuroimaging data.

1.5. ML algorithms and neuroimaging studies

Here we have selected a number of machine learning algorithms (see Bzdok, Altman,& Krzywinski, 2018 for a treatment of the overlap between statistics and machine learning) as the target of a structured quantitative examination of their performance on the same

neuroimaging datasets. The selected algorithms have been applied to linear regression problems in neuroimaging research to date and are implemented in machine learning toolboxes intended for use with neuroimaging data. The statistical tool historically used most often for linear regression and prediction problems in psychological and biological science – multiple regression (MR) - is used as a 'baseline' against which to compare the machine learning algorithms. In MR, it is assumed that the output variable is a linear combination of all input variables, and regression weights are determined for each variable based on this assumption. MR is a non-sparse method and may thus not be suitable for very high-dimensional data. A non-sparse machine learning algorithm evaluated here is Gaussian Process regression (GPR). GPR is a non- parametric probabilistic Bayesian method that uses a predefined covariance function ('kernel') to optimize the function of input values describing the output. While GPR has been applied with some success to various prediction problems using MRI data (e.g. Momte-Rubio et al., 2018), choosing the kernel in GPR appropriately for neuroimaging data may prove challenging. A Multiple Kernel Learning (MKL) approach implemented here uses the L1 norm to create a sparse combination of multiple kernels (Rakotomamonjy et al., 2008). MKL was previously found to outperform support vector machine models when using fMRI data to classify stimulus types (Schrouff et al., 2018). Another kernel method is Kernel Ridge Regression (KRR), which uses a kernel to make ridge regression (regularization via the L2 norm) non-linear (Shawe-Taylor & Cristianini, 2004). KRR can be thought of as a specific case of GPR but lacks the ability to give confidence bounds. KRR has been used to predict treatment outcomes in children with autism spectrum disorders based on fMRI to biological motion stimuli (Yang et al., 2016). The Elastic Net (EN) combines the L1 and L2 penalties to arrive at a linear solution and has been used to predict substance use outcomes in a large sample of adolescents based on functional and structural MRI (Whelan et al., 2014). For MR, GPR, MKL, KRR, and EN, each input feature is assigned a weight, which may be zero when regularization is used (EN and KRR). This is not the case for Random Forest (RF) models. Rather, a number of decision trees are grown based on the input features and the output. The predicted outcomes from multiple trees are aggregated using bootstrap aggregation, and in this way the tendency to overfit is greatly attenuated using RF. RF has been used in many neuroimaging studies for a variety of applications such as classification of patients (e.g. Fredo et al., 2018; Zhu et al., 2018).

1.6. The current study

Here, the efficacy of the five machine learning tools outlined above for use with large neuroimaging datasets was assessed. The performance of a number of machine learning algorithms used for linear regression problems in neuroimaging was compared to standard multiple regression as a baseline to evaluate the added value of choosing each machine learning algorithm. We conducted an empirical evaluation of the extent to which feature selection and resampling procedures influenced results. The effect that data dimensionality has on accuracy was quantified by varying both sample size and number of features. Using simulated neuroimaging data with varying predictor effect sizes as well as real neuroimaging data, this study first compared performance of the Elastic Net, standard multiple regression, a state-of-the-art machine learning toolbox for imaging data (PRoNTo, Schrouff et al., 2013), and an implementation of the Random Forest method available in Matlab.

Furthermore, we examined how the addition of bagging and feature selection affected the accuracy of results from simulated and real data, using an embedded feature selection approach developed with the intention of minimizing researcher degrees of freedom. Based on previous work, it was anticipated that both feature selection and regularization would improve predictions for datasets with large feature sets by creating less complex models, and that bagging would reduce overfitting for small samples by reducing the effect of outliers.

2. Methods

2.1. Machine Learning protocol

The analysis steps outlined below were implemented in MATLAB 2016b using custom analysis scripts for EN, MR, and RF, and the PRoNTo Toolbox for GPR, MKL, and KRR. Analysis scripts used are available at github.com/ljollans/RAFT. Specific aspects of the steps are described below.

2.1.1. Nested cross-validation.—The dataset was initially divided into 10 CV folds, using 90% of the dataset (the training set) to create a regression model which was then tested on the remaining 10% of the data (the test set). This step was performed 10 times, with each fold serving as the test set once. Within the training set, an additional nested CV with 10 partitions was used for feature selection, and for optimization of model parameters. The final (optimized) model from each CV fold was used to make outcome predictions for the test set (10% of the data) and the accuracy of predictions for the entire dataset was used to quantify model fit (see Figure 1).

2.1.2. Feature selection.—An embedded feature selection method was tested; this used prediction accuracy and the stability of model performance across subsets of the sample to learn and to adapt the prediction model. Initial feature ranking, based on the mean squared error of univariate regressions of each feature with the outcome (I.e. the individual 'prediction strength'), was used to define feature subsets, and nested CV was used to assess the stability of findings across different subsets of the data. The key element of this method is an embedded thresholding step, which adjusts the criterion for feature selection according to the performance of feature subsets. This thresholding considered 1) the prediction strength of each feature in each cross-validation fold and 2) the stability of a feature's prediction strength across cross-validation folds. A detailed explanation of this feature selection step is provided in Supplementary information.

2.1.3. Bootstrap aggregation.—All calculations other than the final outcome prediction were validated using 25-fold bootstrap aggregation (*bagging*, see Figure 2). Instead of performing the analysis once using all data, summary datasets were created by randomly sampling data with replacement in each iteration. Results from each iteration are aggregated using the median value.

2.1.4. Model hyperparameter optimization.—Of the algorithms that were tested (other than those in the PRoNTo toolbox) only the Elastic Net has model parameters to optimize. The Elastic Net uses two parameters: λ and α. Alpha represents the weight of LASSO vs. ridge regularization which the Elastic Net uses, and λ is the regularization

coefficient. Both LASSO and Ridge regression apply a penalty for large regression coefficient values, but LASSO regularization favors models with fewer features, making it more prone to excluding features. Parameter values between 0.1 and 1 for α and between 0.032 ($10^{-1.5}$) and 1 for λ were tested. Alpha values were chosen on a linear scale (0.1, 0.325, 0.55 0.775, 1) and λ values on a logarithmic scale (0.032, 0.075, 0.177, 0.422, 1). The lower thresholds for the parameter values were chosen based on observations of models with smaller parameter values failing to converge with neuroimaging models (including the data used here). Here, five values of λ and α were chosen primarily to manage computational expense while also maintaining an adequate range of values as indicated by exploratory comparisons of analyses with a larger range of parameter values. For each model, features excluded by the Elastic Net were noted, and features were removed after the model optimization step if the Elastic Net removed them in more than half of all bagging iterations.

2.1.5. Model validation.—After the nested CV step, the combination of parameters (where applicable) that resulted in the model with the lowest prediction error was identified for each nested CV partition. Prediction error was quantified using root mean squared error. The optimal model parameters from each nested CV partition were used to identify the parameters to be used for the final prediction model in each main CV fold. The evaluation of model fit was carried out using the complete vector of outcome predictions from all CV folds.

2.2. Simulated and Real Data

2.2.1. Constructing simulated data.—The analysis methods were tested on simulated datasets, built to resemble real neuroimaging datasets in terms of the between-feature correlations, and the range of correlations between features and the outcome variable. In order to ensure that the simulated data were reflective of the between- feature correlations and effect size of real data, neuroimaging data from the IMAGEN study (Schumann et al., 2010) were used as a guide (see Supplementary Information). Correlation coefficients for correlations between features (within and between contrasts) and between features and the continuous outcome variable were calculated (see Figure 3). Simulated data were constructed to mirror these correlation strengths as closely as possible, while achieving variation in predictor strength between data types. There are a number of different neuroimaging atlases that parcellate the brain into various numbers of ROIs. We sought to capture differences in the covariance structure of neuroimaging datasets by simulating atlases with either 278 or 97 ROIs. Simulated data were constructed by combining three layers of data matrices to capture the following elements of real neuroimaging data: (1) a non-random relationship to an outcome variable, (2) a cluster structure within the data with increased collinearity within subgroups of features or “clusters”, (3) a low baseline correlation level between all features of the dataset. The construction of simulated data occurred as follows:

Step 1: Predictor and outcome creation. A random matrix X was created with 2000 observations by 1000 features. A vector b representing beta weights, and a vector Y representing the continuous outcome variable were created such that $X*b=Y$. Here the

vector b was created using a range of predefined weightings similar to the range of effect sizes for features in X in relation to Y based on the real neuroimaging reference data.

Step 2: Inter-region of interest (ROI) correlation clusters. A covariance matrix was created that was used to create 30 small clusters of features that were strongly correlated with each other using the *mvnrnd* function in MATLAB. The correlation coefficients for these Inter-ROI correlations were between r=.2 and r=.8, peaking at r=.6. The 1000 features created in Step 1 were assigned to either one of 20 'clusters' of 33 features or 10 clusters of 34 features (i.e., analogous to correlated networks of extent = 33 or 34 ROIs). In this way, a feature within a cluster was more strongly correlated with other features within the same cluster, and was only weakly correlated with features outside the cluster. This cluster size was chosen to best approximate the correlation structure in the real neuroimaging reference data. While the subsets of the simulated feature sets were selected without regard for this cluster structure, the random nature of subset selection was expected to result in a proportional number of features from each cluster being selected, in line with what would be expected when selecting a subset of ROIs.

Step 3: Whole-brain correlations. Similar to the process used in Step 2 the *mvnrnd* function was used to additionally create one matrix the size of X with features that were all correlated with each other at r=.25 on average.

Step 4: Dataset creation. The layers of data created in Steps 1, 2, and 3 were combined using different weighting for each layer to achieve some variation in predictor strength (i.e., the final dataset was a weighted summation of all three data layers). The range of correlations 1) between features and 2) between features and the outcome was manipulated to produce datasets with small to moderate predictor effect sizes ($\text{Simulated}_{\text{small}}$), and datasets with strong predictor effect sizes ($\text{Simulated}_{\text{Large}}$; see Figure 3).

2.2.2. Real MRI data.—In order to test if findings transfer to real-world imaging data two real neuroimaging datasets were selected. First, a dataset from the IMAGEN study (Schumann et al., 2010) that included data from 967 participants was selected. The linear outcome variable used was the score on the block design subscale of the WISC-IV (Wechsler, 2003). Data drawn from grey matter volume (GMV) and the Global Cognitive Assessment Task (GCA, Pinel et al., 2007) were used. In the GCA task participants were presented with visual and auditory stimuli for short sentences (e.g. '*We easily found a taxi in Paris*'), subtractions (e.g. '*Subtract nine from eleven*'), and motor instructions (e.g. '*Press the left button three times*'). Maps for subtractions and sentence presentations were used. Data from these two GCA contrasts and for GMV were extracted using the same functionally defined atlas used to create simulated data (Shen et al., 2013), as outlined in the Supplementary Information. A total of 834 ROIs were used. Note that the data from the GCA task were not used to establish the correlation coefficients to construct simulated datasets (see above). Based on previous work examining the relationship between intelligence and neuroimaging findings (Deary, Penke & Johnson, 2010) this dataset was presumed to have low-moderate effect sizes and was thus termed $\text{Imaging}_{\text{small}}$. The IMAGEN project was approved by all local ethics research committees, and informed consent was obtained from participants and their parents/guardians. A detailed description of

the study protocol and data acquisition has been previously published (Schumann et al., 2010).

The second real neuroimaging dataset was comprised of 1360 structural T1 MRI images drawn from a number of sources: the Autism Brain Imaging Data Exchange II (ABIDE II; Di Martino et al., 2017); the Neuro Bureau – Berlin: Mind and Brain dataset (http://fcon_1000.projects.nitrc.org/indi/pro/Berlin.html); the Beijing Normal University Enhanced Sample (BNU, http://fcon_1000.projects.nitrc.org/indi/retro/BeijingEnhanced.html); the Centre for Biomedical Research Excellence (COBRE; http://fcon_1000.projects.nitrc.org/indi/retro/cobre.html) dataset; the Dallas Lifespan Brain Study (DLBS; http://fcon_1000.projects.nitrc.org/indi/retro/dlbs.html) dataset; the Washington University in St. Louis (WUSL; Power, Barnes, Snyder, Schlaggar, & Petersen, 2012) sample; the Nathan Kline Institute Rockland Sample - Release 1 (NKI; Nooner et al., 2012), the Information eXtraction from Images dataset (IXI; http://www.brain-development.org), and the Southwest University Adult Lifespan Dataset (SALD; Wei et al., 2017). These data were freely available online through either NITRC.org or http://www.brain-development.org. Data from 97 grey matter ROIs based on the AAL atlas (Tzurio-Mazoyer et al., 2002) were extracted. These ROIs were also the same used when determining parameters for construction of the simulated data (see supplementary information). The linear outcome variable used was participants' age, which has been shown to have a moderate-large effect size (Cole et al., 2017). This dataset was thus termed $\text{Imaging}_{\text{large}}$.

2.2.3. Evaluation of dataset size and number of features.—For each analysis, simulated datasets ($\text{Simulated}_{\text{small}}$ and $\text{Simulated}_{\text{large}}$ to function as a comparison to the two real neuroimaging datasets with presumed differences in effect size) were generated with 2000 observations and 1000 features, and subsets of these data were randomly sampled. Simulated data were sampled with the following sample sizes: 75, 200, 400, 750, 1000, and 2000. The size of the input feature set (regions of interest) was sampled using the following number of features: 75, 200, 400, 750, or 1000. Therefore, analyses were carried out across 30 dataset sizes. The maximum number of features and observations for simulated data was chosen to be comparable in dimensionality to the real neuroimaging data while also offering some insight into how an increase in sample size may affect findings. For $\text{Imaging}_{\text{small}}$, random subsampling of the dataset (N=967 and 834 ROIs) at the following sample sizes was carried out: 75, 200, 400, 750, and 967. The features were subsampled at 75, 200, 400, 750 and 834 features. Therefore, analyses were carried out at 25 dataset sizes. $\text{Imaging}_{\text{large}}$ (N=1360 and 97 ROIs) was subsampled only in the domain of sample size, using the following sample sizes: 75, 200, 400, 750, 1000, and 1360. This resulted in analyses being carried out at 6 dataset sizes.

2.2.4. Regression machine performance.—Analyses for each approach at each cell (i.e., each sample and feature set size) and for each data type were carried out 10 times, with 10 different CV assignments. To directly compare performance of different machines for each data type, the results of all analysis iterations for all algorithms within each cell were combined, and the quintiles of this distribution were calculated. Based on the median prediction accuracy of each algorithm within that cell we determined the quintile of an

algorithm's performance, thereby determining a ranking of algorithms on a scale of 1 to 5 for each cell. For a clearer representation of rank, those algorithms that had negative median prediction accuracy (i.e., zero results) were assigned rank zero within each cell. This approach was carried out in lieu of null hypothesis significance testing due to the large number of analyses carried out here: 546 results are displayed in Fig. 4. The examination of quintile rank shown across sample and feature set sizes is an alternative method of demonstrating differences in performance that takes into account the entire spectrum of prediction performance across all approaches and the volatility of rankings across dataset sizes. When considered in conjunction with a representation of median prediction performance of each approach the quintile ranking is an intuitive representation of performance comparison. As no significance levels were established using this method the observed effects will be described in terms of the quintile ranking and median model performance.

2.2.5. Bagging and Feature Selection.—MR and EN are variants of generalized linear regression, which are relatively simple in comparison to GPR, KRR, and MKL. It is plausible that the performance of MR and EN could be improved through the addition of bagging and/or embedded feature selection. We also examined RF with feature selection (RF already incorporates bagging). A series of t-tests at each sample and feature set size were conducted to examine if embedded feature selection and/or bagging significantly changed results.

3. Results

3.1. Machine comparison

Median out-of-sample model performance (i.e., correlation between prediction for the test set and truth) for all regression algorithms is shown in Figures 4 and 5. There was a clear effect of predictor effect sizes on prediction accuracy, with more accurate predictions for both $Simulated_{Large}$, *and* $Imaging_{Large}$, relative to $Simulated_{Small}$, *and* $Imaging_{Small}$, for all analysis methods.

RF had the least amount of variation between data types, although it produced poorer predictions for datasets with large sample and feature set sizes relative to the other algorithms with all data types except $Imaging_{Large}$. The strongest variation in prediction accuracy between data types was observed for GPR, KRR, and MKL. These methods produced lower predictions than other approaches for $Imaging_{Large}$ and failed to produce significant results at any sample and feature set size for $Simulated_{Small}$ and $Imaging_{Small}$. However, KRR and GPR produced predictions similar to other approaches for $Simulated_{Large}$.

The extent to which increases in sample and feature set size affected accuracy varied by analysis method and data type, but except for MR and MKL, the highest prediction accuracy was always achieved for datasets with the largest sample size and highest feature set sizes within each data type. For MR, the *'curse of dimensionality'* was observed for $Simulated_{Large}$, $Simulated_{Small}$, and $Imaging_{Small}$, such that models with large numbers of features relative to observation failed due to overfitting. This effect was also observed for

feature sets up to 400 features with RF for $Imaging_{small}$. For MKL, $Simulated_{Large}$ data indicated that prediction accuracy declined when the sample size exceeded 400 and more than 200 features were included. While some predictions at small sample sizes reached significance with $Simulated_{Large}$ and $Imaging_{Large}$, predictions were generally most successful if the sample size was at least N=200, and ideally more than N=400.

3.1.1. Dataset: $Simulated_{Small}$.—

RF had a high ranking across all cells with sample size (N) over 200. EN ranked highest for datasets with N≥750. For datasets with 400 or more features and N between 200 and 750, RF and EN performed similarly. EN performed very poorly with small samples, particularly when the feature set was small. While MR ranked below RF and EN for almost all sample and feature set sizes, accuracy for ML for N=75 and up to 200 features was higher than for RF and EN. MKL, KRR, and GPR ranked below the other approaches in all cells, except for MKL at sample size equal to 400 and 75 features.

3.1.2. Dataset: $Imaging_{Small}$.—

Quintile ranks for $Imaging_{Small}$ were very similar to results for $Simulated_{Small}$, lending plausibility to the simulated data findings. EN ranked highest N≥400, but performed poorly with small samples. Ranks for MKL, KRR, and GPR were zero for all dataset sizes. There was a trend toward higher performance of MR with smaller feature sets and higher performance of RF with larger feature sets.

3.1.3. Dataset: $Simulated_{Large}$.—

GPR showed the highest average ranking overall. In comparison to other methods, RF ranked lowest across cells. The *'curse of dimensionality'* effect was evident in the rankings for MR, which performed broadly similar to KRR and EN when the sample size exceeded the feature set size, but showed distinctly poor performance (comparable to RF) when the number of features exceeded the sample size. EN, KRR, and GPR ranked very similarly for datasets with N≥400, but EN ranked lowest with small feature sets. KRR and MKL both ranked above other approaches for small datasets with more features than observations, and performed better with small sample sizes than EN. Both GPR and EN performed poorly for datasets with small samples and small feature sets. MKL performed very poorly for datasets with large samples, particularly when the number of features was also large.

3.1.4. Dataset: $Imaging_{Large}$.—

Although only one feature set size was examined, data from $Imaging_{Large}$ repeated the finding of low performance of MKL, KRR and GPR compared to the other approaches. Unlike with other data types, both RF and MR outperformed EN at larger sample sizes. Given the similarity in median performance at larger sample sizes for EN, MR, and RF, this was due to only very small differences in accuracy (see Figure 4). Furthermore, RF performed equal to or better than all other algorithms for datasets with N<1000, while MR performed best for datasets with N≥1000.

3.2. Change in prediction accuracy from Feature Selection and bagging

Changes in prediction accuracy from adding embedded feature selection, bagging, or both in combination were evaluated (see Figure 6). Mean performance of RF, MR and EN with feature selection and/or bagging (see Figure 7) and quintile ranks recalculated to include

analyses with feature selection and/or bagging (see Figure 8) showed considerable effects of feature selection and bagging on algorithm performance. Ranks for the original six algorithms (see Figure 5) showed little change for $Imaging_{Small}$, $Simulated_{Large}$ and $Imaging_{Large}$. For $Simulated_{Small}$ ranks for RF, MR, and EN were reduced as MR and EN with bagging and/or feature selection ranked equal to or higher than the original approaches. Across data types, the rank of RF improved as RF with feature selection ranked very low for all data types except $Imaging_{Large}$, and MR with FS and bagging ranked very low for $Simulated_{Large}$ and $Imaging_{Large}$.

3.2.1. Random Forest (RF)

3.2.1.1. Feature selection.: The addition of embedded feature selection did not improve prediction accuracy of RF for any dataset size or data type. Significant decreases in prediction accuracy were observed for $Simulated_{Small}$ when at least 200 features and $N \geq 750$ were used, and for $Imaging_{Small}$ with 750 or more features and $N \geq 400$ and above. In the quintile ranking of all analysis approaches RF with feature selection ranked very highly for $Imaging_{Large}$, in the absence of any significant changes in prediction accuracy. In contrast, RF with feature selection ranked very low for all other data types.

3.2.2. Multiple Regression (MR)

3.2.2.1. Feature selection.: There were some small improvements in prediction accuracy for MR as a result of adding embedded feature selection with all data types. For $Simulated_{Small}$ and $Imaging_{Small}$ improvements occurred with $N \geq 750$, and for $Simulated_{Large}$ and $Imaging_{Large}$ improvements occurred with $N \geq 75$ with additional small improvements up to a sample size of 400 for $Imaging_{Large}$. For $Simulated_{Small}$, MR with feature selection ranked higher than MR in the quintile ranking for almost all dataset sizes with more than 200 observations and features, and for most datasets with 400 or more features with $Imaging_{Small}$. Examination of the relationship between feature set size and accuracy at each sample size revealed that these differences in accuracy were due to a reduction of the 'curse of dimensionality' effect observed with MR, evidenced by non-negative correlations between number of features and accuracy (see Figure 9). Quintile ranks for $Simulated_{Small}$ also showed that rank of MR with feature selection was higher than rank of MR for datasets with N=75 and more than 75 features. At larger sample sizes, rankings and observed correlations between feature set size and accuracy were very similar, indicating no effect of the feature selection step on performance. With $Imaging_{Large}$, ranking of MR with feature selection was higher than ranking for MR for $N<400$, and lower for larger samples.

3.2.2.2. Bagging.: When bagging was used, prediction accuracy for MR also showed improvements for all data types except $Imaging_{Small}$. For $Simulated_{Small}$ there were some improvements for sample sizes over 400 and 1000 features and when N=400 and 75 features, and higher quintile ranks for MR with bagging compared to MR without bagging at almost all dataset sizes. For $Simulated_{Large}$ improvements occurred for datasets with sample sizes over 75 and at least 400 features when the number of features was equal to or larger than the sample size. These cells overlap to a large extent with the dataset sizes for which the *'curse of dimensionality'* effect was observed (see Figure 7). Examination of the

correlations between feature set size and accuracy revealed that bagging drastically increased this correlation for $Simulated_{Large}$, resulting in an almost complete disappearance of the '*curse of dimensionality'* effect when evaluating algorithm performance (see Figure 9). For $Imaging_{Large}$ improvements as a result of bagging occurred at sample size equal to 75 and were thus similar to those seen for *feature selection*.

3.2.2.3. Feature selection and bagging.: When both feature selection and bagging were used performance of MR for $Simulated_{Small}$ showed some small improvements for datasets with N=400 to N=1000 and 200 or more features, and quintile rank for MR with feature selection and bagging was higher than rank for MR at almost all dataset sizes with N>75. Performance of $Imaging_{Small}$ was also improved at the largest dataset size (N=967 and 834 features), while performance was reduced at N=750 and 75 features. As with $Simulated_{Small}$, quintile ranks for MR with feature selection and bagging were higher than ranks for MR for most cells with N>75, when 400 or more features were used. For $Simulated_{Large}$ performance was improved for N=400 and 400 to 750 features, but performance decreased for datasets for which the sample size was larger than the number of features with 200 or more features and N>400. Similarly, performance for $Imaging_{Large}$ was reduced for datasets with N>200, and quintile ranks for MR with feature selection and bagging were lower than those of MR in most cells for both $Simulated_{Large}$ and $Imaging_{Large}$, although ranks for datasets with N<400 were higher in some cells. For all data types the number of features showed a reduced correlation with prediction accuracy when MR was combined with both feature selection and bagging (see Figure 9). For $Simulated_{Large}$ and $Imaging_{Large}$ this caused reduced accuracy compared to MR alone when sample sizes exceeded feature set sizes.

3.2.3. Elastic Net (EN)

3.2.3.1. Feature selection.: For $Simulated_{Small}$ and $Imaging_{Small}$ the addition of feature selection to EN resulted in significant reductions in accuracy for datasets with N>400 and 400 or more features. For $Simulated_{Small}$ there was a small improvement from feature selection at N=75 and 75 features. While quintile ranks for both $Simulated_{Small}$ and $Imaging_{Small}$ were reduced for EN with feature selection compared to EN for N>400, quintile ranks at small sample sizes were higher for EN with feature selection than for EN in some cells. Examination of the relationship between feature set size and accuracy revealed that the addition of feature selection reduced the positive correlation between number of features and accuracy, which accounts for reduced EN performance with large datasets when FS was used (see Figure 9). For $Simulated_{Large}$ there was a small improvement in accuracy from feature selection at N=1000 and 750 features. Despite only a small significant change in prediction accuracy, quintile ranks indicated that EN with feature selection outperformed EN in almost all cells for $Simulated_{Large}$, with EN with feature selection ranking highest among all analysis approaches for almost all cells with N≥400 and 200 or more features. While feature selection also reduced the correlation between feature set size and accuracy for $Simulated_{Large}$, the correlation remained at $r \tilde{} .5$ for N≥400, which is comparable to the correlations observed for $Simulated_{Small}$ and $Imaging_{Small}$ without feature selection. No significant differences were observed for $Imaging_{Large}$, and quintile ranks for EN with FS and EN were largely the same for this data type.

3.2.3.2. Bagging.: The addition of bagging to EN only resulted in a significant change in accuracy for $Imaging_{Small}$ at N=967 and 400 features, where accuracy was reduced. While quintile ranks for EN with bagging were lower than ranks for EN in most cells for $Imaging_{Small}$, ranks for the other data types were similar between EN and EN with bagging. However, for both $Simulated_{Small}$ and $Imaging_{Small}$ EN with bagging ranked highest and equal to EN alone for N≥750 and large feature set sizes (400 or more for $Simulated_{Small}$ and 834 for $Imaging_{Small}$). Examination of the relationship between feature set size and accuracy revealed only a very small difference in correlations for EN and for EN with bagging (see Figure 9).

3.2.3.3. Feature selection and bagging.: When both FS and bagging were used performance of EN with $Simulated_{Small}$ and $Imaging_{Small}$ was again significantly reduced for datasets with N>400 and 400 or more features, as was the case for EN with feature selection only. Similarly, the correlation between feature set size and accuracy was also reduced for $Simulated_{Small}$ and $Imaging_{Small}$ when both feature selection and bagging were used (see Figure 9). With $Simulated_{Small}$ quintile ranks for EN with bagging and feature selection were higher than for EN with just feature selection and lower than for EN alone. Ranks at large dataset sizes were higher for EN with only bagging than for EN with bagging and feature selection. For $Imaging_{Small}$ quintile ranks of EN with bagging and feature selection were lower than ranks for EN only and EN with feature selection. However, ranks for EN with bagging and feature selection were higher than for EN with only bagging in most cells. As with both bagging and feature selection individually, the combination of both bagging and FS did not result in any significant changes in accuracy for $Imaging_{Large}$. However, the quintile ranking showed that for $Imaging_{Large}$ EN with both bagging and feature selection ranked very poorly at the largest sample size (N=1000). While $Simulated_{Large}$ had shown a small improvement in accuracy for large datasets with feature selection, and no significant change for bagging, the addition of both bagging and feature selection resulted in a decrease in accuracy for the largest dataset sizes (i.e. N=2000 and 1000 features). The quintile ranking for $Simulated_{Large}$ indicated lower rank for EN with bagging and feature selection in almost all cells compared to EN with feature selection, lower performance in some cells than EN with bagging, and some improvement at small feature set sizes compared to EN alone. Unlike feature selection alone, feature selection in combination with bagging did not result in a reduction of the correlation between feature set size and accuracy for $Simulated_{Large}$ (see Figure 9), which accounts for the higher quintile rank of analyses with FS in many cells with N≥400.

An additional analysis reporting the association of regularization and prediction accuracy is contained in Supplemental Information.

4. Discussion

Analytical tools developed for data science have become frequently used in neuroimaging (Woo et al., 2017), but none of these tools were specifically developed for neuroimaging data. With the small samples, large feature sets, and low signal-to-noise that are characteristic of neuroimaging data, prediction models built using neuroimaging data are at a high risk of overfitting. In this paper, the merit of six different regression approaches for

prediction analysis was empirically evaluated and compared using simulated and real neuroimaging data for the first time. Results showed that GPR, MKK, and KRR could produce good predictions, but failed when effect sizes were small regardless of sample size. The Elastic Net on the other hand emerged as the most flexible and reliable regression machine. The Elastic Net created the most accurate prediction models independent of absolute predictor effect sizes, and across many sample and feature set sizes. Predictions were always improved when sample size was increased, but across all analyses a minimum sample size of about 400 emerged as necessary to achieve reliable results. At smaller sample sizes and for datasets with weak effect sizes modest improvements in accuracy could be made using an embedded feature selection method. Another approach designed to increase model performance – bootstrap aggregation – could counteract the decline in standard Multiple Regression model accuracy with more predictor variables than observations. However, given adequate dataset sizes and using the Elastic Net, neither feature selection nor bootstrap aggregation improved findings significantly, and indeed resulted in substantially increased computational time for all analyses and reduced accuracy for some models. A visual summary of findings with regard to the best analysis method based on sample size, number of features, and effect size is given in Figure 10.

The central observation of this study was that different types of regression approaches provide widely different results, and that these results are differently affected by sample size, number of predictors, and the ratio of signal to noise in the data. Previous meta-analyses by Kambeitz and colleagues (2015, 2016) have shown that not only the outcome to be predicted, but also the type of neuroimaging data that is used has a strong effect on the maximum performance of a model. Findings in the present study confirmed that when using multivariate regression methods, the expected size of the effect and effect sizes for individual predictor variables are the most important criteria for selection not only of minimum sample size, but also selection of the analysis approach. However, across simulated and real neuroimaging data of varying effect sizes the Elastic Net had the highest median prediction accuracy for datasets with 400 or more features and observations. For smaller feature sets, variations of Multiple Regression resulted in better model fit.

When both the sample and feature set size were small, the MATLAB implementation of random forest (Treebagger) also showed some promise. A key difference between Random Forest and many other regression methods is that the contribution of individual predictors is not easily, or at all, determinable from a completed model. While it has been debated in the literature whether the main goal of neuroimaging prediction should be predicting an outcome as accurately as possible, or identifying when and where data contain information about an outcome (Paulus, 2015; Pine & Leibenluft, 2015), the readability of neuroimaging prediction models is an important aspect of model development. The ability to scrutinize the contribution of individual neuroimaging predictors allows researchers to verify the neurophysiological plausibility of the model, while also enabling future research to consider which variables are strong or poor predictors of an outcome in the development of further experiments, studies, and prediction models (Woo et al., 2017; Jollans & Whelan, 2018). Although some methods make it possible to gain insight into the contribution of individual predictors in random forest models (e.g., Palczewska et al., 2014), these are computationally expensive. Random forest should, in theory, outperform the Elastic Net if non-linear

relationships are present in the data. However, the results from the real imaging data suggest that non-linearities are either not present in the data or that the random forest did not detect them, at least in our implementation using TreeBagger.

The Elastic Net, which had the most consistent performance across effect sizes and dataset sizes, was able to improve prediction accuracy by adding more input features. That is, given smaller sample sizes, inclusion of a larger feature set is thus one approach to improve model performance (assuming the extra data contains some signal). Crucially, preselection of variables for inclusion in the model did not improve performance, and indeed resulted in lower model accuracy in some cases. Therefore, we suggest that neuroimaging researchers do not preselect regions of interest or contrasts of interest before implementing Elastic Net models. This will allow researchers to conduct analyses that include variables not previously linked to the outcome of interest, without being unduly penalized by the inclusion of more *exploratory* variables. This is important because most neuroimaging literature to date reports only univariate and frequentist findings that may not translate to predictive utility (Lo et al., 2015). An important caveat is that our findings may only apply to ROI data, and may only hold when sample sizes exceed a certain minimum threshold as determined by the smallest sample sizes examined in this study (i.e. N=75–200). Voxelwise analyses and analyses with very small sample sizes are likely to benefit from some additional dimension reduction, as we showed in our findings regarding accuracy for very small samples using the embedded feature selection approach.

There was evidence for a beneficial effect of embedded feature selection at small sample sizes for both the Elastic Net and multiple regression. Through feature selection, the association between the number of features and model performance tended to shift toward zero, reducing the *'curse of dimensionality'* effect for multiple regression, but also counteracting the positive relationship between feature set size and model performance at large sample sizes for the Elastic Net. For the Elastic Net the feature selection step greatly reduced the need for regularization, as seen by very small regularization weights for analyses after feature selection. Any significant improvements in model performance because of embedded feature selection were not consistent or strong enough to recommend use of this approach, particularly considering the computational expense. Time needed to run Elastic Net analyses with N=400 and 1000 features was approximately 18 seconds (see Figure 11) for the standard Elastic Net (r=.25) compared to 225 minutes for the Elastic Net with the embedded feature selection approach (r=.15). While the feature selection method utilized here did not result in consistent improvements in model performance, innovations in dimension reduction for neuroimaging studies are forthcoming (e.g. Koutsouleris et al. 2018), and the possible benefit of these new tools should not be discounted based on the results obtained using one specific method in this paper.

In contrast to embedded feature selection, there was strong evidence for the utility of bootstrap aggregation to improve prediction accuracy with Multiple Regression. This approach strongly counteracted the *'curse of dimensionality'* effect for multiple regression. Indeed, for half of all cells with fewer than 400 features or a sample size of N<400 Multiple Regression paired with 25-fold bootstrap aggregation performed in the highest quintile for the simulated and real neuroimaging data with weak effect sizes. There was no significant

effect of bootstrap aggregation on performance of the Elastic Net, and quintile ranks for analyses with and without this method were largely similar. Given the relatively small increase in time required for computations when bootstrap aggregation was used (see Figure S2) it may then be worthwhile including this method with a view to increasing model stability.

There are some important limitations to the generalizability of findings in this study. While there were strong commonalities across results for the real neuroimaging dataset examined here and results achieved using data simulations, there was some indication that not all characteristics of real neuroimaging data were sufficiently accounted for in the simulations. In particular, there was higher accuracy for analyses with Random Forest for the real compared to the simulated datasets. Further examination of Random Forest and other regression methods such as Support Vector Machines for neuroimaging data are therefore warranted. Furthermore, only ROI data rather than voxelwise analyses were considered in this study. While this decision was based on the intention of creating models that are easily interpretable, findings also do not necessarily translate to models with a strongly increased feature set size, and the characteristics of voxelwise as compared to ROI data are likely quite different in terms of the between-feature correlations and predictor strengths. Finally, based on previous findings that non-brain variables are much better predictors of phenotypic outcomes than neuroimaging data (Whelan et al., 2014), identifying the best methods to integrate imaging and non-imaging data in prediction analyses is an important step.

Conclusion

A number of recommendations for future machine learning studies using neuroimaging data can be made based on these findings. Datasets with at least 400 observations have the highest likelihood of uncovering meaningful findings. When at least 400 observations and 400 or more predictor variables are included in the analysis, regularized regression via the Elastic Net was shown to be the best analysis approach for ROI data. When the sample or feature set size is smaller, standard Multiple Regression supported by bootstrap aggregation showed the best performance in this study. Furthermore, when the analysis framework was chosen appropriately - based on number of observations and presumed effect size - increasing the number of ROI variables for inclusion in a model improved results, eliminating the need for the researcher to preselect variables for inclusion.

Here, we have shown that the choice of analysis approach for linear regression analyses has a large impact on the accuracy of the resulting regression model. The sample size and number of predictors are important factors that determine the analysis approach that will have the greatest success in extracting meaningful information from a neuroimaging dataset. Data-driven machine learning approaches have great potential for increasing reproducibility in neuroimaging. Understanding the boundaries of what machine learning can achieve with neuroimaging data will help the field make informed choices.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Funding Acknowledgements:

LJ and RB are supported by the Irish Research Council under Grant Number GOIPG/2014/418 and EPSPG/2017/277 respectively. This work received support from the following sources: the European Union-funded FP6 Integrated Project IMAGEN (Reinforcement-related behaviour in normal brain function and psychopathology) (LSHM-CT-2007-037286), the Horizon 2020 funded ERC Advanced Grant 'STRATIFY' (Brain network based stratification of reinforcement-related disorders) (695313), ERANID (Understanding the Interplay between Cultural, Biological and Subjective Factors in Drug Use Pathways) (PR-ST-0416–10004), BRIDGET (JPND: BRain Imaging, cognition Dementia and next generation GEnomics) (MR/N027558/1), the FP7 projects IMAGEMEND(602450; IMAging GEnetics for MENtal Disorders) and MATRICS (603016), the Innovative Medicine Initiative Project EU-AIMS (115300–2), the Medical Research Council Grant 'c-VEDA' (Consortium on Vulnerability to Externalizing Disorders and Addictions) (MR/N000390/1), the Swedish Research Council FORMAS, the Medical Research Council, the National Institute for Health Research (NIHR) Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's College London, the Bundesministeriumfür Bildung und Forschung (BMBF grants 01GS08152; 01EV0711; eMED SysAlc01ZX1311A; Forschungsnetz AERIAL 01EE1406A, 01EE1406B), the Deutsche Forschungsgemeinschaft (DFG grants SM 80/7–2, SFB 940/2), the Medical Research Foundation and Medical research council (grant MR/R00465X/1). Further support was provided by grants from: ANR (project AF12-NEUR0008–01 - WM2NA, and ANR-12-SAMA-0004), the Fondation de France, the Fondation pour la Recherche Médicale, the Mission Interministérielle de Lutte-contre-les-Drogues-et-les-Conduites-Addictives (MILDECA), the Assistance-Publique-Hôpitaux-de-Paris and INSERM (interface grant), Paris Sud University IDEX 2012; the National Institutes of Health, Science Foundation Ireland (16/ERCD/3797), U.S.A. (Axon, Testosterone and Mental Health during Adolescence; RO1 MH085772–01A1), and by NIH Consortium grant U54 EB020403, supported by a cross-NIH alliance that funds Big Data to Knowledge Centres of Excellence. HG is supported by the National Institute on Drug Abuse (T32DA043593 and R01DA047119).

References

Abeel T, Helleputte T, Van de Peer Y, Dupont P, & Saeys Y (2010). Robust biomarker identification for cancer diagnosis with ensemble feature selection methods. Bioinformatics, 26(3), 392–398. 10.1093/bioinformatics/btp630 [PubMed: 19942583]

Adar N, Okyay S, Özkan K, Şaylısoy S, Adapınar BDÖ, & Adapınar B (2016). Feature Selection on MR Images Using Genetic Algorithm with SVM and Naive Bayes Classifiers. Retrieved from http://dergipark.gov.tr/download/article-file/236999

Arbabshirani MR, Plis S, Sui J, & Calhoun VD (2017). Single subject prediction of brain disorders in neuroimaging: Promises and pitfalls. NeuroImage, 145, 137–165. 10.1016/J.NEUROIMAGE.2016.02.079 [PubMed: 27012503]

Ball TM, Stein MB, Ramsawh HJ, Campbell-Sills L, & Paulus MP (2014). Single-subject anxiety treatment outcome prediction using functional neuroimaging. Neuropsychopharmacology, 39(5), 1254–1261. [PubMed: 24270731]

Bellec P, Rosa-Neto P, Lyttelton OC, Benali H, & Evans AC (2010). Multi-level bootstrap analysis of stable clusters in resting-state fMRI. Neuroimage, 51(3), 1126–1139. [PubMed: 20226257]

Biswal BB, Mennes M, Zuo X-N, Gohel S, Kelly C, Smith SM, … Milham MP (2010). Toward discovery science of human brain function. Proceedings of the National Academy of Sciences, 107(10), 4734–4739. 10.1073/pnas.0911855107

Breiman L (1996). Bagging predictors. Machine Learning, 24(2), 123–140.

Brown TT, Kuperman JM, Chung Y, Erhart M, McCabe C, Hagler DJ Jr., … Dale AM (2012). Neuroanatomical Assessment of Biological Maturity. Current Biology, 22(18), 1693–1698. 10.1016/j.cub.2012.07.002 [PubMed: 22902750]

Bzdok D, Altman N, & Krzywinski M (2018). Statistics versus machine learning. Nature methods, 15(4), 233. [PubMed: 30100822]

Button KS, Ioannidis JPA, Mokrysz C, Nosek BA, Flint J, Robinson ESJ, & Munafò MR (2013). Power failure: why small sample size undermines the reliability of neuroscience. Nature Reviews Neuroscience, 14(5), 365–376. 10.1038/nrn3475 [PubMed: 23571845]

Cawley GC, & Talbot NL (2010). On over-fitting in model selection and subsequent selection bias in performance evaluation. Journal of Machine Learning Research, 11(Jul), 2079–2107.

Chandrashekar G, & Sahin F (2014). A survey on feature selection methods. Computers & Electrical Engineering, 40(1), 16–28. 10.1016/j.compeleceng.2013.11.024

Chekroud AM, Zotti RJ, Shehzad Z, Gueorguieva R, Johnson MK, Trivedi MH, … Corlett PR (2016). Cross-trial prediction of treatment outcome in depression: a machine learning approach. The Lancet Psychiatry, 3(3), 243–250. 10.1016/S2215-0366(15)00471-X [PubMed: 26803397]

Chu C, Hsu A-L, Chou K-H, Bandettini P, & Lin C (2012). Does feature selection improve classification accuracy? Impact of sample size and feature selection on classification using anatomical magnetic resonance images. NeuroImage, 60(1), 59–70. 10.1016/J.NEUROIMAGE.2011.11.066 [PubMed: 22166797]

Clark VP, Beatty GK, Anderson RE, Kodituwakku P, Phillips JP, Lane TDR, … Calhoun VD (2014). Reduced fMRI activity predicts relapse in patients recovering from stimulant dependence: Prediction of Relapse Using fMRI. Human Brain Mapping, 35(2), 414–428. 10.1002/hbm.22184 [PubMed: 23015512]

Cole JH, Poudel RPK, Tsagkrasoulis D, Caan MWA, Steves C, Spector TD, & Montana G (2017). Predicting brain age with deep learning from raw imaging data results in a reliable and heritable biomarker. NeuroImage, 163, 115–124. 10.1016/j.neuroimage.2017.07.059 [PubMed: 28765056]

Costafreda SG, Chu C, Ashburner J, & Fu CHY (2009). Prognostic and Diagnostic Potential of the Structural Neuroanatomy of Depression. PLoS ONE, 4(7), e6353 10.1371/journal.pone.0006353 [PubMed: 19633718]

Davatzikos C, Bhatt P, Shaw LM, Batmanghelich KN, & Trojanowski JQ (2011). Prediction of MCI to AD conversion, via MRI, CSF biomarkers, and pattern classification. Neurobiology of Aging, 32(12), 2322–e19.

Deary IJ, Penke L, & Johnson W (2010). The neuroscience of human intelligence differences. Nature Reviews Neuroscience, 11(3), 201–211. 10.1038/nrn2793 [PubMed: 20145623]

Di Martino A, O'Connor D, Chen B, Alaerts K, Anderson JS, Assaf M, … Milham MP (2017). Enhancing studies of the connectome in autism using the autism brain imaging data exchange II. Scientific Data, 4, 170010 10.1038/sdata.2017.10 [PubMed: 28291247]

Dosenbach NUF, Nardos B, Cohen AL, Fair DA, Power JD, Church JA, … Schlaggar BL (2010). Prediction of Individual Brain Maturity Using fMRI. Science, 329(5997), 1358–1361. 10.1126/science.1194144 [PubMed: 20829489]

Dubois J, & Adolphs R (2016). Building a Science of Individual Differences from fMRI. Trends in Cognitive Sciences, 20(6), 425–443. 10.1016/j.tics.2016.03.014 [PubMed: 27138646]

Duff EP, Trachtenberg AJ, Mackay CE, Howard MA, Wilson F, Smith SM, & Woolrich MW (2012). Task-driven ICA feature generation for accurate and interpretable prediction using fMRI. NeuroImage, 60(1), 189–203. 10.1016/j.neuroimage.2011.12.053 [PubMed: 22227050]

Dwyer DB, Falkai P, & Koutsouleris N (2018). Machine Learning Approaches for Clinical Psychology and Psychiatry. Annual Review of Clinical Psychology, 14, 91–118. 10.1146/annurev-clinpsy-032816-045037

Efron B, & Tibshirani R (1997). Improvements on Cross-Validation: The .632+ Bootstrap Method. Journal of the American Statistical Association, 92(438), 548 10.2307/2965703

Formisano E, De Martino F, & Valente G (2008). Multivariate analysis of fMRI time series: classification and regression of brain responses using machine learning. Magnetic Resonance Imaging, 26(7), 921–934. 10.1016/j.mri.2008.01.052 [PubMed: 18508219]

Franke K, Ziegler G, Klöppel S, & Gaser C (2010). Estimating the age of healthy subjects from T1-weighted MRI scans using kernel methods: Exploring the influence of various parameters. NeuroImage, 50(3), 883–892. 10.1016/j.neuroimage.2010.01.005 [PubMed: 20070949]

Fredo AJ, Jahedi A, Reiter M, & Müller RA (2018). Diagnostic Classification of Autism using Resting-State fMRI Data and Conditional Random Forest. Age (years), 12(2.76), 6–41.

Gabrieli JDE, Ghosh SS, & Whitfield-Gabrieli S (2015). Prediction as a Humanitarian and Pragmatic Contribution from Human Cognitive Neuroscience. Neuron, 85(1), 11–26. 10.1016/j.neuron.2014.10.047 [PubMed: 25569345]

Garraux G, Phillips C, Schrouff J, Kreisler A, Lemaire C, Degueldre C, … Salmon E (2013). Multiclass classification of FDG PET scans for the distinction between Parkinson's disease and atypical parkinsonian syndromes. NeuroImage: Clinical, 2, 883–893. [PubMed: 24179839]

Gorgolewski KJ, Varoquaux G, Rivera G, Schwarz Y, Ghosh SS, Maumet C, … Margulies DS (2015). NeuroVault.org: a web-based repository for collecting and sharing unthresholded statistical maps of the human brain. Frontiers in Neuroinformatics, 9 10.3389/fninf.2015.00008

Gorgolewski K, Esteban O, Schaefer G, Wandell B, & Poldrack R (2017). OpenNeuro—a free online platform for sharing and analysis of neuroimaging data. Organization for Human Brain Mapping. Vancouver, Canada, 1677.

Hall P, & Robinson AP (2009). Reducing variability of crossvalidation for smoothing-parameter choice. Biometrika, 96(1), 175–186. 10.1093/biomet/asn068

Jack CR, Bernstein MA, Fox NC, Thompson P, Alexander G, Harvey D, … Weiner MW (2008). The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods. Journal of Magnetic Resonance Imaging, 27(4), 685–691. 10.1002/jmri.21049 [PubMed: 18302232]

Jie N, Zhu M, Ma X, Osuch EA, Wammes M, Théberge J, … Calhoun VD (2015). Discriminating Bipolar Disorder From Major Depression Based on SVM-FoBa: Efficient Feature Selection With Multimodal Brain Imaging Data. IEEE Transactions on Autonomous Mental Development, 7(4), 320–331. 10.1109/TAMD.2015.2440298 [PubMed: 26858825]

Jollans L, & Whelan R (2016). The Clinical Added Value of Imaging: A Perspective From Outcome Prediction. Biological Psychiatry: Cognitive Neuroscience and Neuroimaging, 1(5), 423–432. 10.1016/j.bpsc.2016.04.005 [PubMed: 29560871]

Jollans L, & Whelan R (2018). Neuromarkers for Mental Disorders: Harnessing Population Neuroscience. Frontiers in Psychiatry, 9 10.3389/fpsyt.2018.00242

Kambeitz J, Cabral C, Sacchet MD, Gotlib IH, Zahn R, Serpa MH, … Koutsouleris N (2016). Detecting Neuroimaging Biomarkers for Depression: A Meta-analysis of Multivariate Pattern Recognition Studies. Biological Psychiatry. 10.1016/j.biopsych.2016.10.028

Kambeitz J, Kambeitz-Ilankovic L, Leucht S, Wood S, Davatzikos C, Malchow B, … Koutsouleris N (2015). Detecting neuroimaging biomarkers for schizophrenia: a meta-analysis of multivariate pattern recognition studies. Neuropsychopharmacology, 40(7), 1742–1751. [PubMed: 25601228]

Kennedy DN, Haselgrove C, Riehl J, Preuss N, & Buccigrossi R (2016). The NITRC Image Repository. NeuroImage, 124(0 0), 1069–1073. 10.1016/j.neuroimage.2015.05.074 [PubMed: 26044860]

Kohavi R (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection In Ijcai (Vol. 14, pp. 1137–1145). Stanford, CA Retrieved from https://pdfs.semanticscholar.org/0be0/d781305750b37acb35fa187febd8db67bfcc.pdf

Koutsouleris N, Borgwardt S, Meisenzahl EM, Bottlender R, Moller H-J, & Riecher-Rossler A (2012). Disease Prediction in the At-Risk Mental State for Psychosis Using Neuroanatomical Biomarkers: Results From the FePsy Study. Schizophrenia Bulletin, 38(6), 1234–1246. 10.1093/schbul/sbr145 [PubMed: 22080496]

Koutsouleris N, Kahn RS, Chekroud AM, Leucht S, Falkai P, Wobrock T, … Hasan A (2016). Multisite prediction of 4-week and 52-week treatment outcomes in patients with first-episode psychosis: a machine learning approach. The Lancet Psychiatry, 3(10), 935–946. [PubMed: 27569526]

Koutsouleris N, Kambeitz-Ilankovic L, Ruhrmann S, Rosen M, Ruef A, Dwyer DB, … Borgwardt S (2018). Prediction Models of Functional Outcomes for Individuals in the Clinical High-Risk State for Psychosis or With Recent-Onset Depression: A Multimodal, Multisite Machine Learning Analysis. JAMA Psychiatry, 75(11), 1156–1172. 10.1001/jamapsychiatry.2018.2165

Lo A, Chernoff H, Zheng T, & Lo S-H (2015). Why significant variables aren't automatically good predictors. Proceedings of the National Academy of Sciences, 112(45), 13892–13897. 10.1073/pnas.1518285112

Loken E, & Gelman A (2017). Measurement error and the replication crisis. Science, 355(6325), 584–585. 10.1126/science.aal3618 [PubMed: 28183939]

Monté-Rubio GC, Falcón C, Pomarol-Clotet E, & Ashburner J (2018). A comparison of various MRI feature types for characterizing whole brain anatomical differences using linear pattern recognition methods. NeuroImage, 178, 753–768. [PubMed: 29864520]

Munson MA, & Caruana R (2009). On Feature Selection, Bias-Variance, and Bagging In Buntine W, Grobelnik M, Mladenić D, & Shawe-Taylor J (Eds.), Machine Learning and Knowledge Discovery in Databases (pp. 144–159). Berlin, Heidelberg: Springer Berlin Heidelberg.

Mwangi B, Hasan KM, & Soares JC (2013). Prediction of individual subject's age across the human lifespan using diffusion tensor imaging: A machine learning approach. NeuroImage, 75, 58–67. 10.1016/j.neuroimage.2013.02.055 [PubMed: 23501046]

Mwangi B, Matthews K, & Steele JD (2012). Prediction of illness severity in patients with major depression using structural MR brain scans. Journal of Magnetic Resonance Imaging, 35(1), 64–71. 10.1002/jmri.22806 [PubMed: 21959677]

Mwangi B, Tian TS, & Soares JC (2014). A Review of Feature Reduction Techniques in Neuroimaging. Neuroinformatics, 12(2), 229–244. 10.1007/s12021-013-9204-3 [PubMed: 24013948]

Niehaus KE, Clark IA, Bourne C, Mackay CE, Holmes EA, Smith SM, … Duff EP (2014). MVPA to enhance the study of rare cognitive events: An investigation of experimental PTSD In Pattern Recognition in Neuroimaging, 2014 International Workshop on (pp. 1–4). IEEE Retrieved from http://ieeexplore.ieee.org/abstract/document/6858536/

Nnamoko N, Arshad F, England D, Vora J, & Norman J (2014). Evaluation of filter and wrapper methods for feature selection in supervised machine learning. Age, 21(81), 33–2.

Nooner KB, Colcombe S, Tobe R, Mennes M, Benedict M, Moreno A, … Milham M (2012). The NKI-Rockland Sample: A Model for Accelerating the Pace of Discovery Science in Psychiatry. Frontiers in Neuroscience, 6 10.3389/fnins.2012.00152

Ogutu JO, Schulz-Streeck T, & Piepho H-P (2012). Genomic selection using regularized linear regression models: ridge regression, lasso, elastic net and their extensions. BMC Proceedings, 6(2), S10 10.1186/1753-6561-6-S2-S10 [PubMed: 22640436]

Palczewska A, Palczewski J, Robinson RM, & Neagu D (2014). Interpreting Random Forest Classification Models Using a Feature Contribution Method In Integration of Reusable Systems (pp. 193–218). Springer, Cham 10.1007/978-3-319-04717-1_9

Paulus MP (2015). Pragmatism instead of mechanism: a call for impactful biological psychiatry. JAMA Psychiatry, 72(7), 631–632. [PubMed: 25992540]

Pine DS, & Leibenluft E (2015). Biomarkers with a mechanistic focus. JAMA Psychiatry, 72(7), 633–634. [PubMed: 25992716]

Pinel P, Thirion B, Meriaux S, Jobert A, Serres J, Le Bihan D, … Dehaene S (2007). Fast reproducible identification and large-scale databasing of individual functional cognitive networks. BMC Neuroscience, 8(1), 91 10.1186/1471-2202-8-91 [PubMed: 17973998]

Poldrack RA (2011). Inferring Mental States from Neuroimaging Data: From Reverse Inference to Large-Scale Decoding. Neuron, 72(5), 692–697. 10.1016/j.neuron.2011.11.001 [PubMed: 22153367]

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(3), 2142–2154. 10.1016/j.neuroimage.2011.10.018 [PubMed: 22019881]

Price CJ, Ramsden S, Hope TMH, Friston KJ, & Seghier ML (2013). Predicting IQ change from brain structure: a cross-validation study. Developmental Cognitive Neuroscience, 5, 172–184. 10.1016/j.dcn.2013.03.001 [PubMed: 23567505]

Ramirez J, Gorriz JM, Ortiz A, Martinez-Murcia FJ, Segovia F, Salas-Gonzalez D, … Puntonet CG (2018). Ensemble of random forests One vs. Rest classifiers for MCI and AD prediction using ANOVA cortical and subcortical feature selection and partial least squares. Journal of Neuroscience Methods, 302, 47–57. 10.1016/j.jneumeth.2017.12.005 [PubMed: 29242123]

Ramírez J, Górriz JM, Ortiz A, Padilla P, & Martínez-Murcia FJ (2016). Ensemble Tree Learning Techniques for Magnetic Resonance Image Analysis In Chen Y-W, Torro C, Tanaka S, Howlett RJ, & Jain LC (Eds.), Innovation in Medicine and Healthcare 2015 (pp. 395–404). Cham: Springer International Publishing.

Ramírez J, Górriz JM, Segovia F, Chaves R, Salas-Gonzalez D, López M, … Padilla P (2010). Computer aided diagnosis system for the Alzheimer's disease based on partial least squares and

random forest SPECT image classification. Neuroscience Letters, 472(2), 99–103. 10.1016/j.neulet.2010.01.056 [PubMed: 20117177]

Saeys Y, Inza I, & Larranaga P (2007). A review of feature selection techniques in bioinformatics. Bioinformatics, 23(19), 2507–2517. 10.1093/bioinformatics/btm344 [PubMed: 17720704]

Schnack HG, & Kahn RS (2016). Detecting Neuroimaging Biomarkers for Psychiatric Disorders: Sample Size Matters. Frontiers in Psychiatry, 7, 50 10.3389/fpsyt.2016.00050 [PubMed: 27064972]

Schrouff J, Rosa MJ, Rondina JM, Marquand AF, Chu C, Ashburner J, … Mourão-Miranda J (2013). PRoNTo: Pattern Recognition for Neuroimaging Toolbox. Neuroinformatics, 11(3), 319–337. 10.1007/s12021-013-9178-1 [PubMed: 23417655]

Schrouff J, Monteiro JM, Portugal L, Rosa MJ, Phillips C, & Mourão-Miranda J (2018). Embedding anatomical or functional knowledge in whole-brain multiple kernel learning models. Neuroinformatics, 16(1), 117–143. [PubMed: 29297140]

Schumann G, Loth E, Banaschewski T, Barbot A, Barker G, Büchel C, … Struve M (2010). The IMAGEN study: reinforcement-related behaviour in normal brain function and psychopathology. Molecular Psychiatry, 15(12), 1128–1139. 10.1038/mp.2010.4 [PubMed: 21102431]

Shen K, Fripp J, Mériaudeau F, Chételat G, Salvado O, & Bourgeat P (2012). Detecting global and local hippocampal shape changes in Alzheimer's disease using statistical shape models. NeuroImage, 59(3), 2155–2166. [PubMed: 22037419]

Shen X, Tokoglu F, Papademetris X, & Constable RT (2013). Groupwise whole-brain parcellation from resting-state fMRI data for network node identification. NeuroImage, 82, 403–415. 10.1016/j.neuroimage.2013.05.081 [PubMed: 23747961]

Simmons JP, Nelson LD, & Simonsohn U (2011). False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. Psychological Science, 22(11), 1359–1366. 10.1177/0956797611417632 [PubMed: 22006061]

Stonnington CM, Chu C, Klöppel S, Jack CR, Ashburner J, & Frackowiak RSJ (2010). Predicting clinical scores from magnetic resonance scans in Alzheimer's disease. NeuroImage, 51(4), 1405–1413. 10.1016/j.neuroimage.2010.03.051 [PubMed: 20347044]

Tangaro S, Amoroso N, Brescia M, Cavuoti S, Chincarini A, Errico R, … Bellotti R (2015). Feature Selection Based on Machine Learning in MRIs for Hippocampal Segmentation. Computational and Mathematical Methods in Medicine. 10.1155/2015/814104

Thompson PM, Andreassen OA, Arias-Vasquez A, Bearden CE, Boedhoe PS, Brouwer RM, … Ye J (2017). ENIGMA and the individual: Predicting factors that affect the brain in 35 countries worldwide. NeuroImage, 145, 389–408. 10.1016/j.neuroimage.2015.11.057 [PubMed: 26658930]

Tohka J, Moradi E, Huttunen H, Initiative ADN, & et al. (2016). Comparison of feature selection techniques in machine learning for anatomical brain MRI in dementia. Neuroinformatics, 14(3), 279–296. [PubMed: 26803769]

Tzourio-Mazoyer N, Landeau B, Papathanassiou D, Crivello F, Etard O, Delcroix N, … Joliot M (2002). Automated Anatomical Labeling of Activations in SPM Using a Macroscopic Anatomical Parcellation of the MNI MRI Single-Subject Brain. NeuroImage, 15(1), 273–289. 10.1006/nimg.2001.0978 [PubMed: 11771995]

Van Essen DC, Ugurbil K, Auerbach E, Barch D, Behrens TEJ, Bucholz R, … Yacoub E (2012). The Human Connectome Project: A data acquisition perspective. NeuroImage, 62(4), 2222–2231. 10.1016/j.neuroimage.2012.02.018 [PubMed: 22366334]

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. 10.1016/j.neuroimage.2016.10.038 [PubMed: 27989847]

Wang J, Redmond SJ, Bertoux M, Hodges JR, & Hornberger M (2016). A Comparison of Magnetic Resonance Imaging and Neuropsychological Examination in the Diagnostic Distinction of Alzheimer's Disease and Behavioral Variant Frontotemporal Dementia. Frontiers in Aging Neuroscience, 8 10.3389/fnagi.2016.00119

Wang Y, Goh JO, Resnick SM, & Davatzikos C (2013). Imaging-Based Biomarkers of Cognitive Performance in Older Adults Constructed via High-Dimensional Pattern Regression Applied to MRI and PET. PLoS ONE, 8(12), e85460 10.1371/journal.pone.0085460 [PubMed: 24392010]

Wechsler D (2003). WISC-IV technical and interpretive manual. San Antonio, TX: Psychological Corporation.

Wei D, Zhuang K, Chen Q, Yang W, Liu W, Wang K, … Qiu J (2018). Structural and functional MRI from a cross-sectional Southwest University Adult lifespan Dataset (SALD) | bioRxiv. Retrieved March 26, 2018, from https://www.biorxiv.org/content/early/2018/01/29/177279

Westfall J, & Yarkoni T (2016). Statistically Controlling for Confounding Constructs Is Harder than You Think. PLOS ONE, 11(3), e0152719 10.1371/journal.pone.0152719 [PubMed: 27031707]

Whelan R, Watts R, Orr CA, Althoff RR, Artiges E, Banaschewski T, … Ziesch V (2014). Neuropsychosocial profiles of current and future adolescent alcohol misusers. Nature, 512(7513), 185–189. 10.1038/nature13402 [PubMed: 25043041]

Woo C-W, Chang LJ, Lindquist MA, & Wager TD (2017). Building better biomarkers: brain models in translational neuroimaging. Nature Neuroscience, 20(3), 365–377. 10.1038/nn.4478 [PubMed: 28230847]

Yang D, Pelphrey KA, Sukhodolsky DG, Crowley MJ, Dayan E, Dvornek NC, … & Ventola P (2016). Brain responses to biological motion predict treatment outcome in young children with autism. Translational psychiatry, 6(11), e948. [PubMed: 27845779]

Zahari SM, Ramli NM, & Mokhtar B (2014). Bootstrapped parameter estimation in ridge regression with multicollinearity and multiple outliers. Journal of Applied Environmental and Biological Sciences, 4, 150–156.

Zhou X, Wang S, Xu W, Ji G, Phillips P, Sun P, & Zhang Y (2015). Detection of Pathological Brain in MRI Scanning Based on Wavelet-Entropy and Naive Bayes Classifier In Ortuño F & Rojas I (Eds.), Bioinformatics and Biomedical Engineering (Vol. 9043, pp. 201–209). Cham: Springer International Publishing 10.1007/978-3-319-16483-0_20

Zhu X, Du X, Kerich M, Lohoff FW, & Momenan R (2018). Random forest based classification of alcohol dependence patients and healthy controls using resting state MRI. Neuroscience letters, 676, 27–33. [PubMed: 29626649]

Zou H, & Hastie T (2005). Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67(2), 301–320. 10.1111/j.1467-9868.2005.00503.x

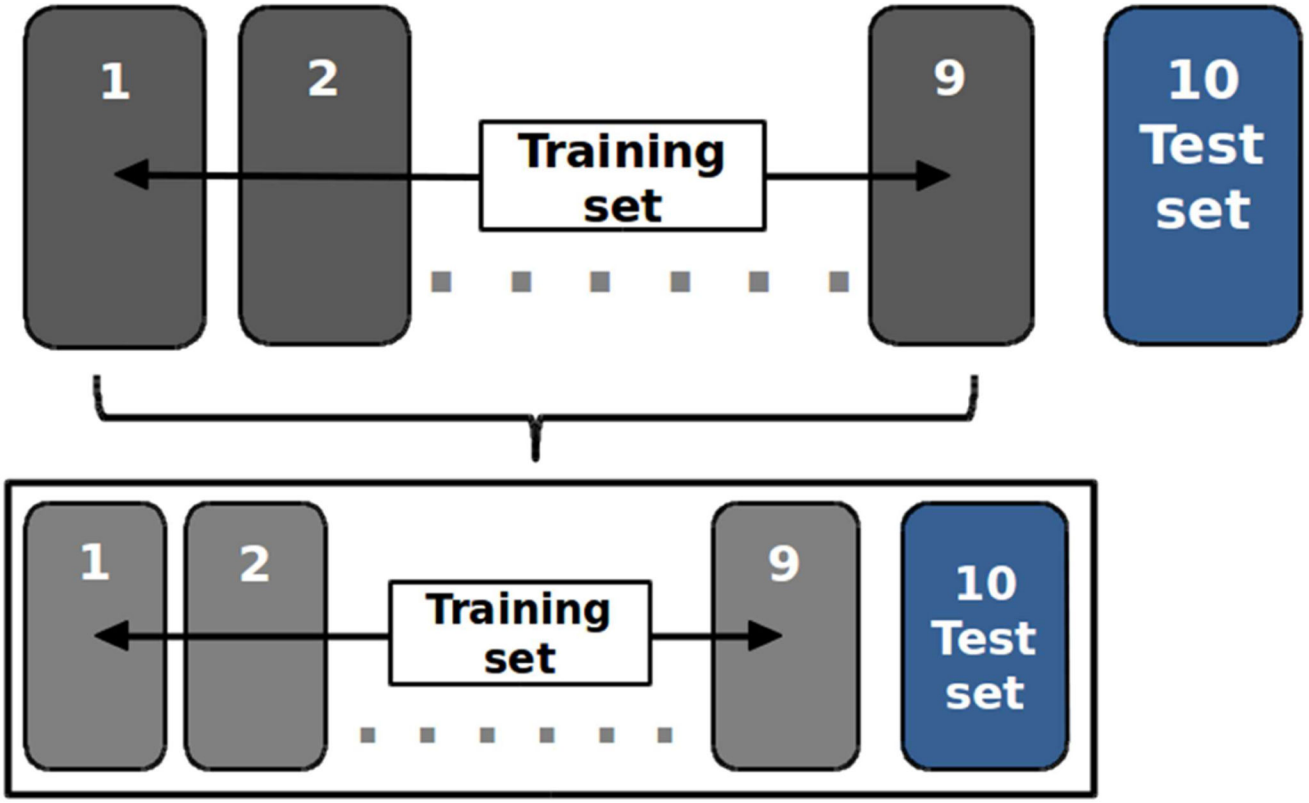

Figure 1:
Representation of the nested cross-validation framework

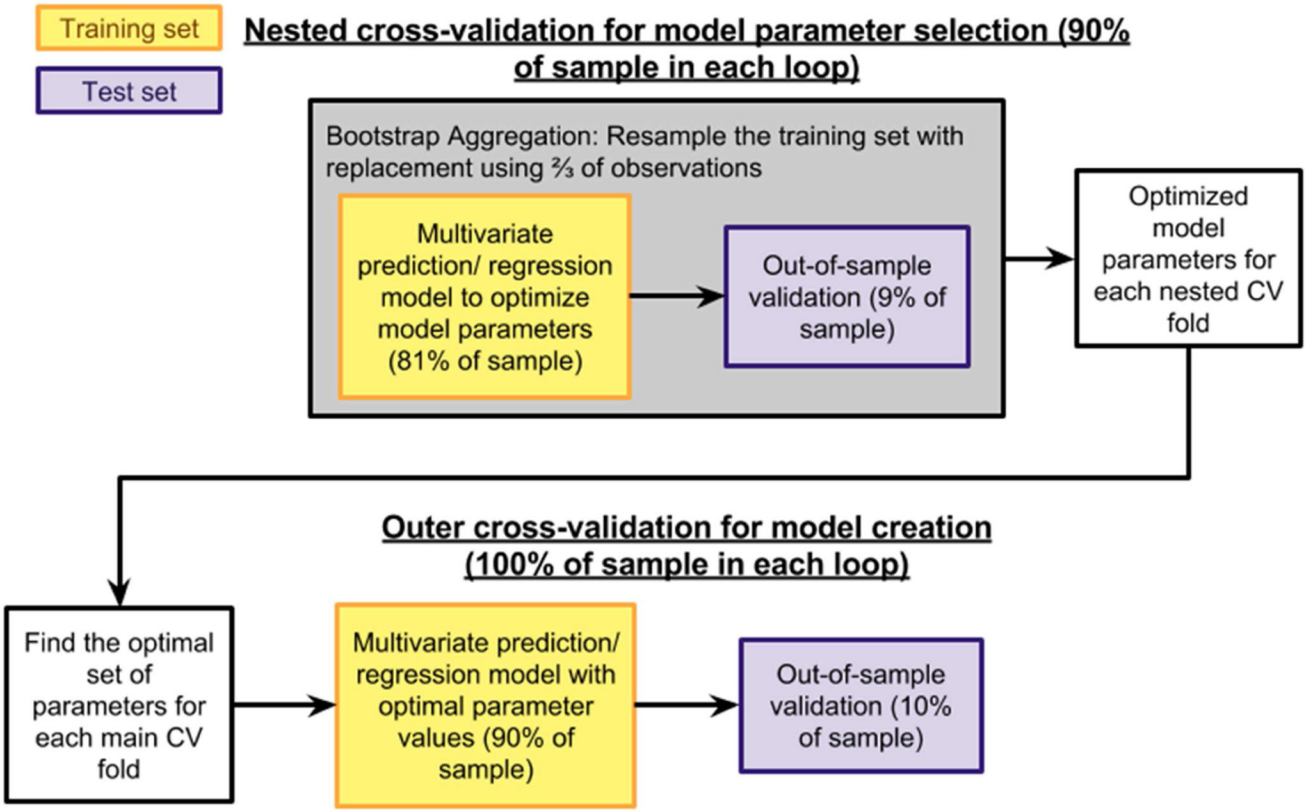

Figure 2:
Analysis framework using bagging and nested cross-validation. CV: Cross-Validation.

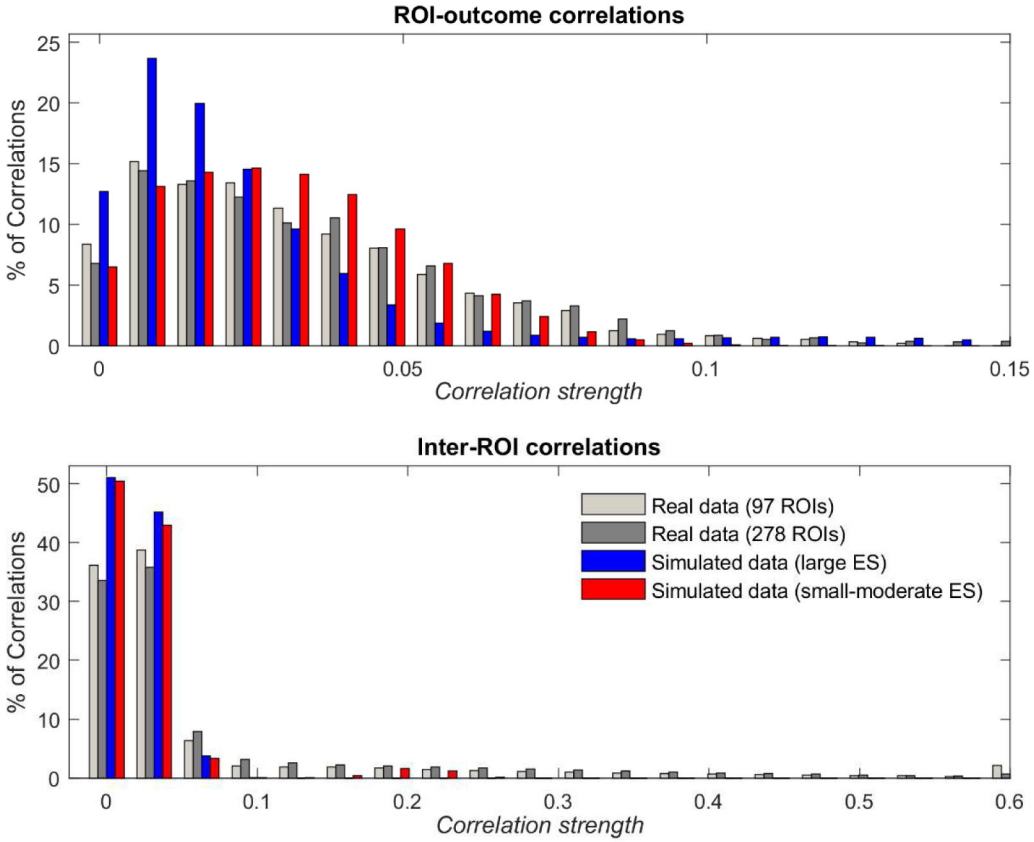

Figure 3.

Correlation strength by percentage of features for correlations between features (Inter-ROI correlations) and between features and the outcome variable (ROI-outcome) for real and simulated datasets. ES: effect size.

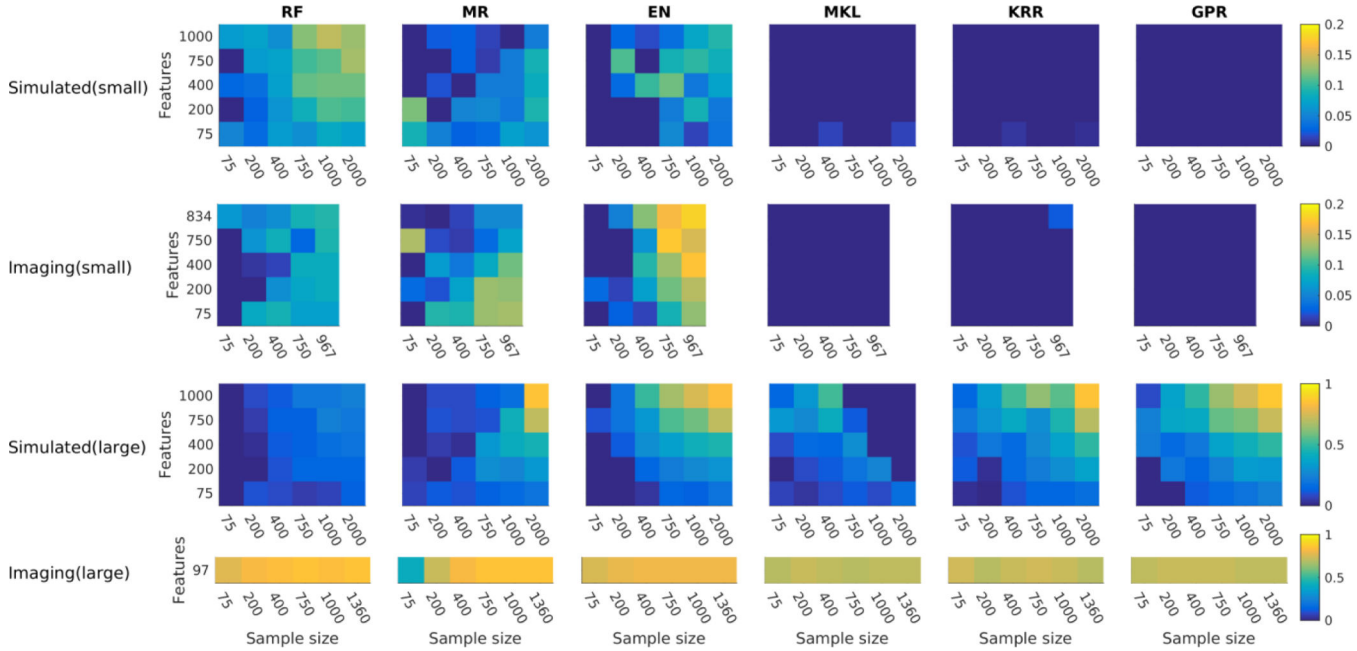

Figure 4.

Median out-of-sample performance by sample size and analysis algorithm ($Simulated_{Large}$, $Simulated_{Small}$, $Imaging_{Large}$ and $Imaging_{Small}$). RF: Random Forest; MR: Multiple Regression; EN: Elastic Net; MKL: Multiple Kernel Learning; KRR: Kernel Ridge Regression; GPR: Gaussian Process Regression. Color bars show the cross-validated Pearson's R value, with higher values (warmer colors) indicating better prediction accuracy. Note that value ranges differ between plots for different data types.

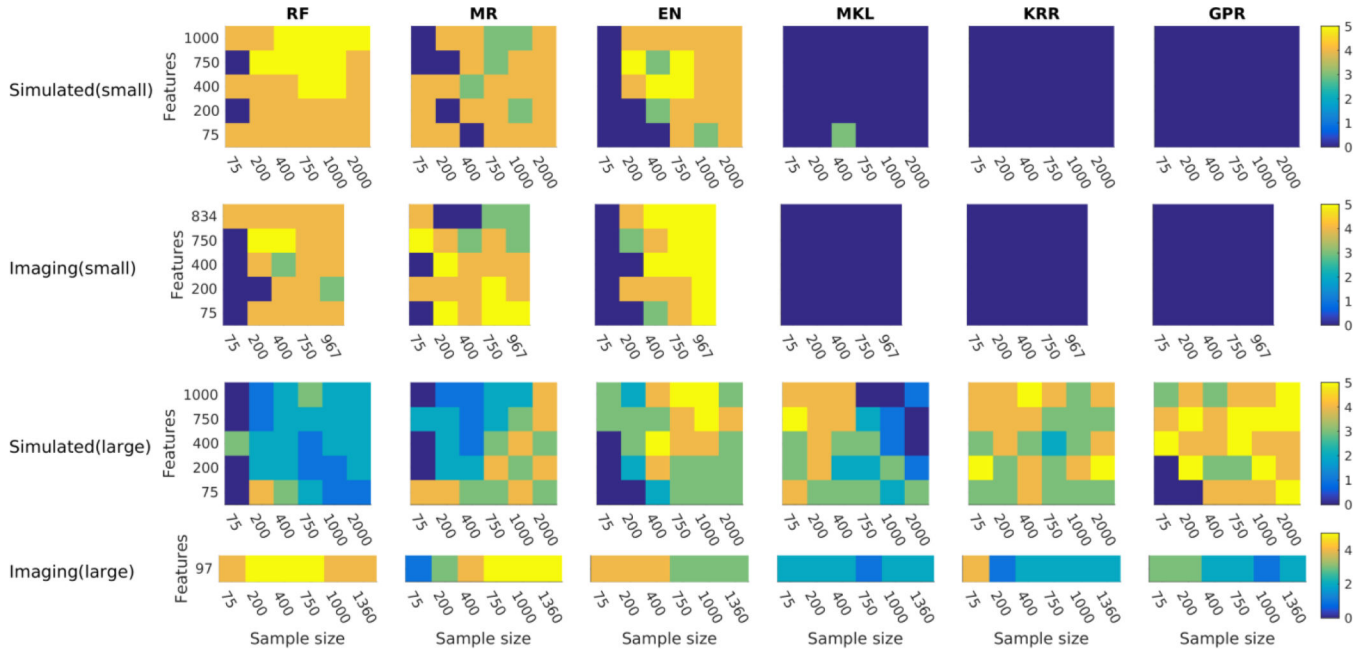

Figure 5.

Quintile rank of prediction accuracy by sample size and analysis algorithm for $Simulated_{Large}$, $Simulated_{Small}$, $Imaging_{Small}$, and $Imaging_{Large}$. Shown ranks are the quintile into which the median prediction accuracy for each method within each data type and cell fell across the distribution of all analysis iteration for each data type and cell. RF: Random Forest; MR: Multiple Regression; EN: Elastic Net; MKL: Multiple Kernel Learning; KRR: Kernel Ridge Regression; GPR: Gaussian Process Regression. Color bars and plot coloring show the rank from zero to five, with higher values (warmer colors) indicating higher rank and therefore better prediction accuracy.

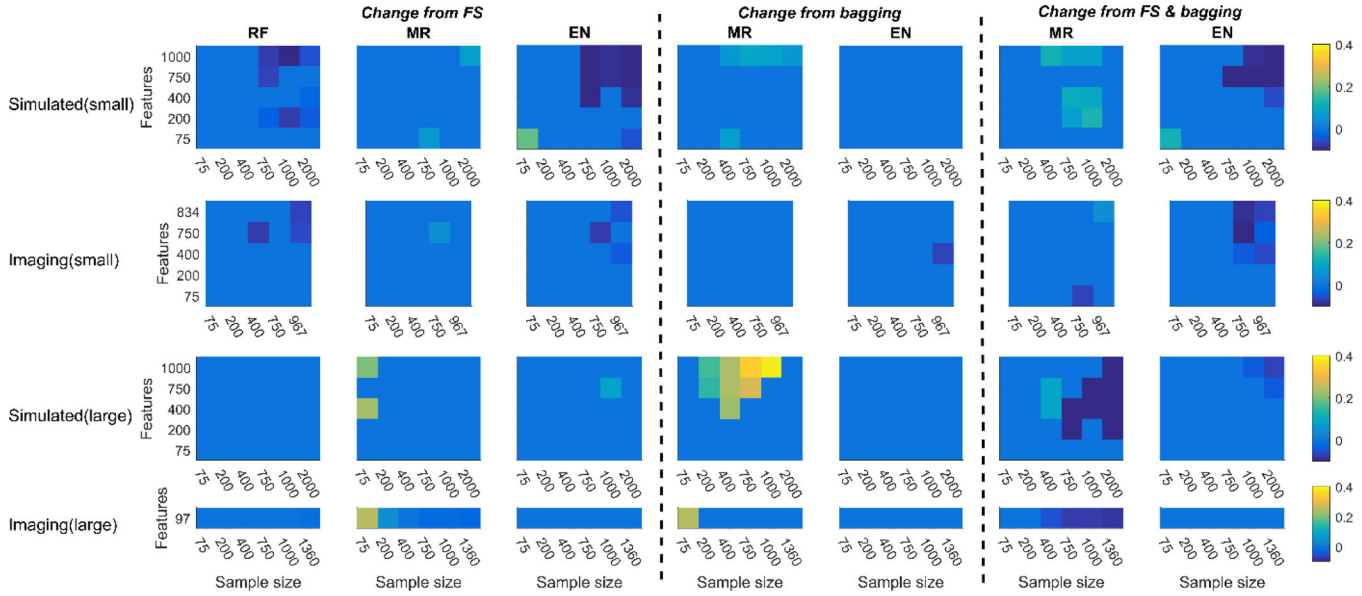

Figure 6.

Significant improvement or decrease in median prediction accuracy ($p<.005$) from adding embedded Feature Selection (FS) and/or bagging to analyses with Random Forest (RF), Multiple Regression (MR), and Elastic Net (EN). Color bars and plot coloring show the difference in median correlation between prediction and truth between standard analyses for each algorithm and analyses with FS and/or bagging. Higher values, indicated by warmer colors, signify an improvement in prediction accuracy, while values below zero indicate a decrease in prediction accuracy.

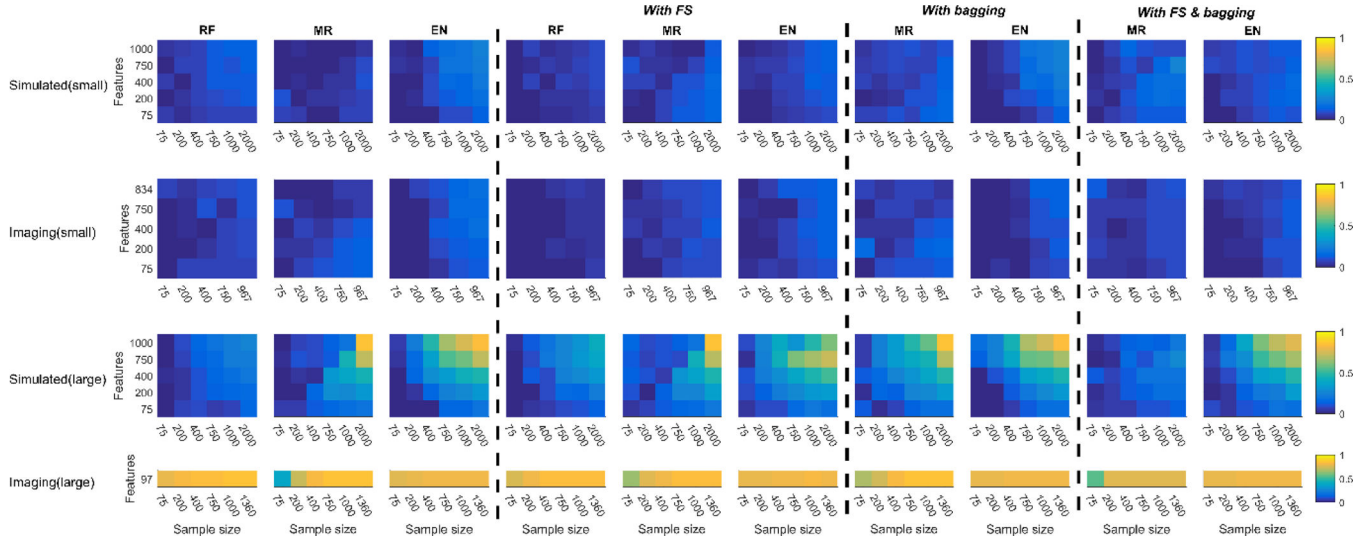

Figure 7.

Mean out-of-sample performance by sample size and analysis algorithm for Random Forest (RF), Multiple regression (MR), and Elastic Net (EN) with and without bagging and embedded feature selection (FS). Color bars show the cross-validated Pearson's R value, with higher values (warmer colors) indicating better prediction accuracy.

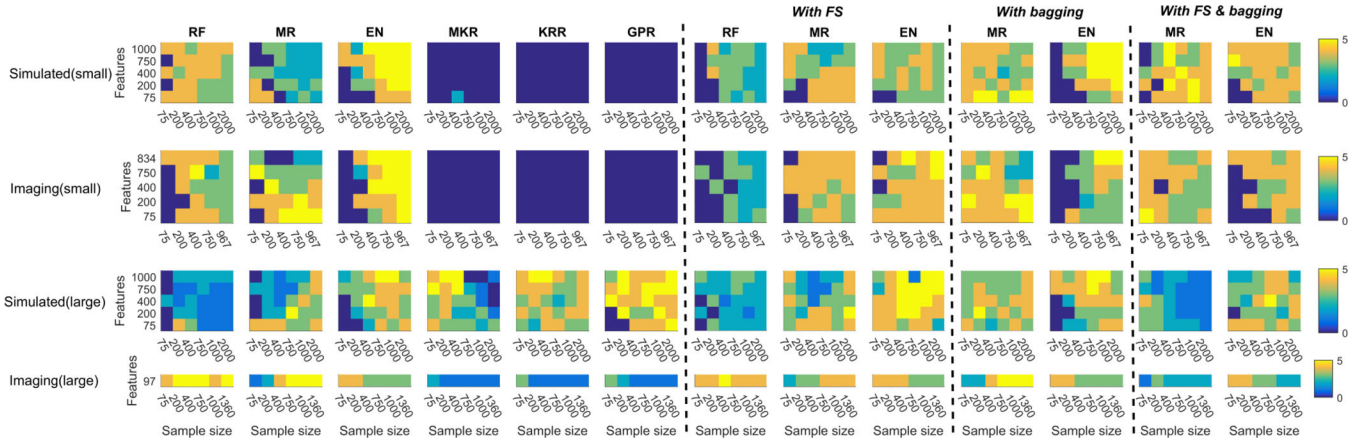

Figure 8.

Quintile rank of prediction accuracy with and without embedded feature selection (FS) and/or bagging by sample size and analysis algorithm for $Simulated_{Small}$, $Imaging_{Small}$, $Simulated_{Large}$, and $Imaging_{Large}$. Shown ranks are the quintile into which the median prediction accuracy for each method within each data type and cell fell across the distribution of all analysis iteration for each data type and cell. RF: Random Forest; MR: Multiple Regression; EN: Elastic Net; MKL: Multiple Kernel Learning; KRR: Kernel Ridge Regression; GPR: Gaussian Process Regression. Color bars and plot coloring show the rank from zero to five, with higher values (warmer colors) indicating higher rank and therefore better prediction accuracy.

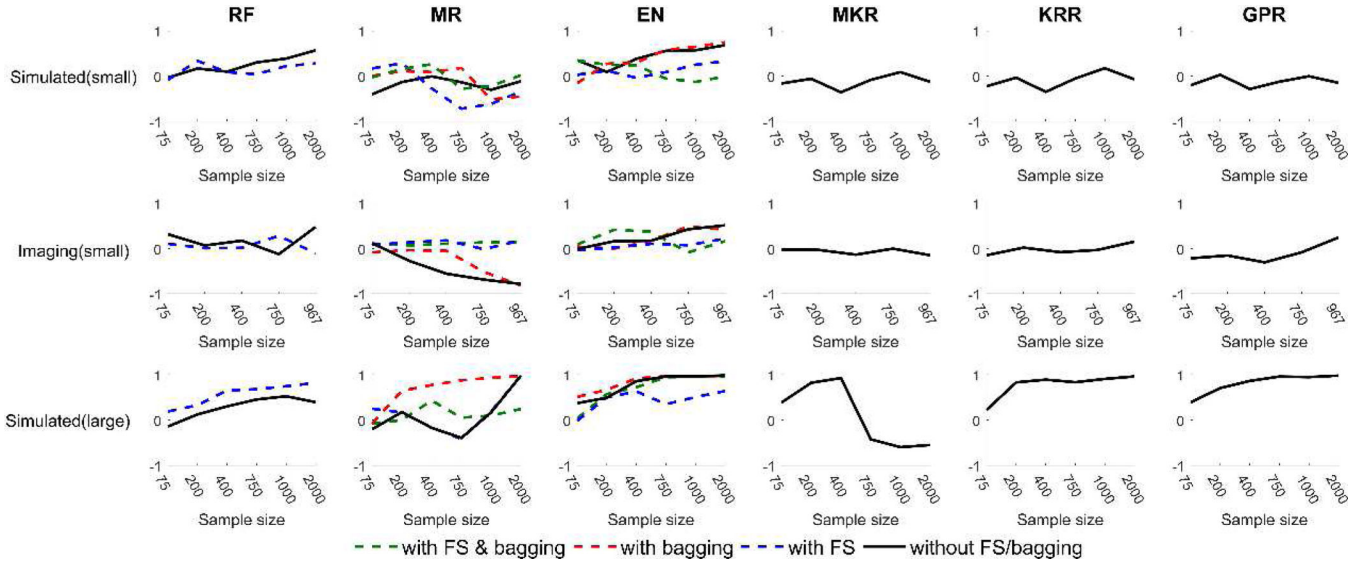

Figure 9.

Correlation between feature set size and prediction accuracy for all analysis approaches and data types. RF: Random Forest; MR: Multiple Regression; EN: Elastic Net; MKL: Multiple Kernel Learning; KRR: Kernel Ridge Regression; GPR: Gaussian Process Regression; FS: Feature Selection.

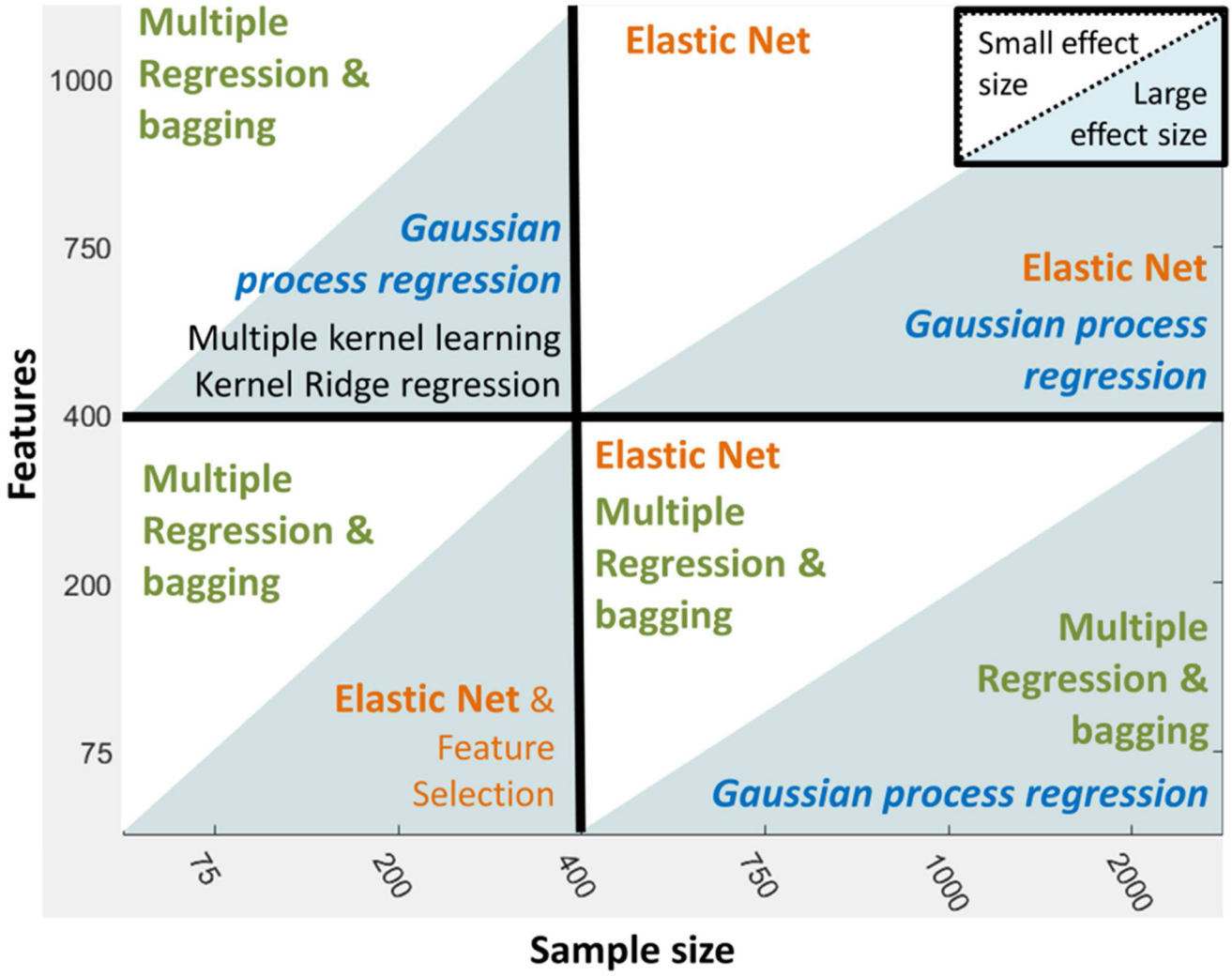

Figure 10.
Graphical representation of best-performing analysis methods by sample size, feature set size, and expected effect size.

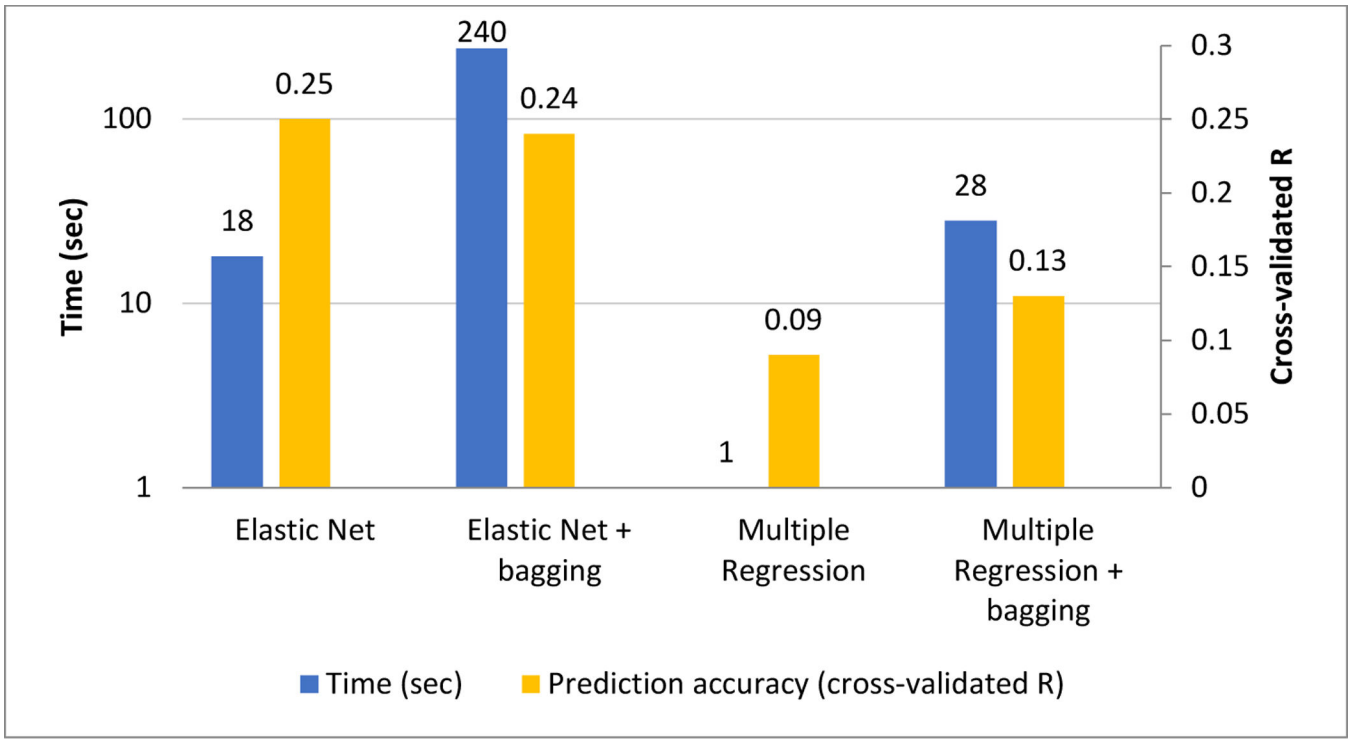

Figure 11.

Example computational time and prediction accuracy for a sample simulated dataset from $\text{Simulated}_{\text{Small}}$ with N=400 and 1000 features.