Published in final edited form as:
Neuroinformatics. 2015 January ; 13(1): 31–46. doi:10.1007/s12021-014-9238-1.

Clinical prediction from structural brain MRI scans: A large-scale empirical study

Mert R. Sabuncu[1,2,*] and **Ender Konukoglu**[1,*] **for the Alzheimer's Disease Neuroimaging Initiative**[**]

[1]Athinoula A. Martinos Center for Biomedical Imaging, Harvard Medical School/Massachusetts General Hospital, Charlestown, MA

[2]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA

Abstract

Multivariate pattern analysis (MVPA) methods have become an important tool in neuroimaging, revealing complex associations and yielding powerful prediction models. Despite methodological developments and novel application domains, there has been little effort to compile benchmark results that researchers can reference and compare against. This study takes a significant step in this direction. We employed three classes of state-of-the-art MVPA algorithms and common types of structural measurements from brain Magnetic Resonance Imaging (MRI) scans to predict an array of clinically relevant variables (diagnosis of Alzheimer's, schizophrenia, autism, and attention deficit and hyperactivity disorder; age, cerebrospinal fluid derived amyloid–β levels and mini-mental state exam score). We analyzed data from over 2,800 subjects, compiled from six publicly available datasets. The employed data and computational tools are freely distributed

Corresponding Author: Mert R. Sabuncu, Athinoula A. Martinos Center for Biomedical Imaging, Massachusetts General Hospital, Building 149, 13th Street, Room 2301, Charlestown, Massachusetts, USA 02129, Phone: 617 643-7460, Fax: 617 726-7422, msabuncu@nmr.mgh.harvard.edu.

*Both authors contributed equally.

**Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database. As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators is available at http://tinyurl.com/ADNI-main.

Author Responsibilities:
Both authors are responsible for manuscript preparation and review and data analysis.

Conflicts of Interest:
No conflicts of interest exist for any of the named authors in this study.

INFORMATION SHARING STATEMENT
The data and computational tools used to generate the cross-validation results presented in this manuscript are made available via: https://www.nmr.mgh.harvard.edu/lab/mripredict.
We note that in compiling these resources, we heavily relied on third-party data collection efforts and software packages. These include the following publicly available datasets, the Alzheimer's Disease Neuroimaging Initiative, or ADNI (www.adni-info.org), the Open-Access Series of Imaging Studies (OASIS, oasis-brains.org), the Autism Brain Imaging Data Exchange (ABIDE, tinyurl.com/fcon1000-abide), the Attention Deficit Hyperactivity Disorder (ADHD) sample from the ADHD-200 Consortium(tinyurl.com/fcon1000-adhd), the Center for Biomedical Research Excellence (COBRE) schizophrenia sample (tinyurl.com/fcon1000-cobre), and the MIND Clinical Imaging Consortium (MCIC) schizophrenia sample (coins.mrn.org). To process the structural MRI scans, we utilized FreeSurfer (https://surfer.nmr.mgh.harvard.edu/). We distribute FreeSurfer-derived morphological measurements in easy-to-read formats. This way, we ensure that researchers with little or no experience in MRI processing can analyze these data. We further employed publicly available implementations of three different classes of Machine Learning algorithms: SVM (csie.ntu.edu.tw/~cjlin/libsvm), RVM (http://people.csail.mit.edu/msabuncu/sw/RVoxM/index.html), and NAF (http://www.nmr.mgh.harvard.edu/~enderk/software.html). We provide all the lists necessary to replicate the 100 random split 5-fold cross-validation sessions we conducted in our analyses. Finally, we distribute a sample script that demonstrates how we compile and evaluate the cross-validation results.

(https://www.nmr.mgh.harvard.edu/lab/mripredict), making this the largest, most comprehensive, reproducible benchmark image-based prediction experiment to date in structural neuroimaging. Finally, we make several observations regarding the factors that influence prediction performance and point to future research directions. Unsurprisingly, our results suggest that the biological footprint (effect size) has a dramatic influence on prediction performance. Though the choice of image measurement and MVPA algorithm can impact the result, there was no universally optimal selection. Intriguingly, the choice of algorithm seemed to be less critical than the choice of measurement type. Finally, our results showed that cross-validation estimates of performance, while generally optimistic, correlate well with generalization accuracy on a new dataset.

Keywords

Image-based prediction; Computer aided diagnosis; machine learning; MRI

INTRODUCTION

Structural Magnetic Resonance Imaging (MRI), a non-invasive and ubiquitous imaging modality, enables the *in vivo* investigation of the morphological features of the human brain macro-anatomy in health and disease, thus offering insights into the underlying neurobiological processes. A growing body of neuroimaging literature (Feinstein et al., 2004; Frisoni et al., 2010; Ho et al., 2003) has demonstrated that markers derived from structural brain MRI scans can aid in clinical decision-making and treatment development, making this imaging technology an invaluable tool for translational science and medical practice.

Multivariate pattern analysis (MVPA), or machine learning, offers a powerful approach in neuroimage analysis, which, until recently, has been dominated by massively univariate (mass-univariate) methods that rely on classical statistical techniques (Ashburner and Friston, 2000). Although MVPA algorithms have been employed for mapping regions of the brain associated with a particular condition of interest (Kriegeskorte *et al.*, 2006), their primary utility is for building image-based predictive models, for example for the purpose of computer-aided diagnosis (Kloppel *et al.*, 2012) or “mind reading” (Friston et al., 2008; Mitchell et al., 2004; Mourao-Miranda et al., 2005). Over the last decade, MVPA has been increasingly applied to structural brain MRI scans, largely for developing models to predict clinical conditions at the individual level (Costafreda et al., 2009; Cuingnet et al., 2011; Davatzikos et al., 2008; Davatzikos et al., 2009; Duchesnay et al., 2007; Duchesne et al., 2009; Ecker et al., 2010; Kawasaki et al., 2007; Kloppel et al., 2009; Kloppel et al., 2008; Koutsouleris et al., 2009; Lao et al., 2004; Lerch et al., 2008; Liu et al., 2012; Mourao-Miranda et al., 2012; Mwangi et al., 2012; Nieuwenhuis et al., 2012; Sabuncu and Van Leemput, 2012; Schnack et al., 2014; Soriano-Mas et al., 2007; Vemuri et al., 2008; Wang et al., 2010; Wilson et al., 2009).

Many prior MVPA studies in neuroimaging have focused on proposing new methods that involve extracting novel types of imaging measurements or using innovative algorithms to improve prediction accuracy or yield more interpretable models (Batmanghelich et al., 2009;

Cho et al., 2012; Davatzikos et al., 2009; Duchesnay et al., 2007; Fan et al., 2007; Nouretdinov et al., 2011; Sabuncu and Van Leemput, 2012; Teipel et al., 2007). However, with notable exceptions (Brown et al., 2012; Cuingnet et al., 2011), there has been little effort to publish benchmark results that researchers can replicate, reference, and objectively compare against. Today, the increasing availability of several widely used, thoroughly validated, and freely distributed

- large-scale clinical neuroimage databases, such as the Alzheimer's Disease Neuroimaging Initiative (ADNI) (Jack *et al.*, 2008), made available through web-based data sharing platforms, such as COINS (Scott *et al.*, 2011) and XNAT (Marcus et al., 2007a),
- neuroimage processing software packages, such as FreeSurfer (Fischl, 2012) and SPM (Friston *et al.*, 1994), and
- implementations of cutting-edge machine learning algorithms, such as LibSVM (Chang and Lin, 2011),

makes such a study possible. This article presents the results of a carefully designed empirical study that employs publicly available computational tools and large-scale multi-site data to report state-of-the-art prediction accuracies and to serve as a *reproducible benchmark* reference for future MVPA studies in structural neuroimaging. In this study, we analyzed data from over 2,800 individuals obtained from six large clinical neuroimaging studies. We used FreeSurfer to extract imaging measurements and publicly available implementations of three different classes of MVPA algorithms to predict clinical diagnoses, for instance of schizophrenia and Alzheimer's disease, and clinically relevant graded variables, such as cognitive performance scores.

The constructed prediction models can directly be useful in clinical practice, e.g., for identifying high-risk subjects, tracking disease progression, or replacing less reliable, more invasive, and/or more expensive diagnostic tests. Furthermore image-based prediction models can also serve basic scientific goals by revealing and quantifying the macro-anatomical footprint of clinical/experimental/behavioral conditions and measuring the information overlap between the image content and non-imaging variables, such as clinical test results.

In addition to reporting experimental results, we also analyze the factors that influence the prediction performance in the domains we considered. We believe that the reported benchmark results, shared data, and presented analyses will catalyze progress and prompt new research in biomedical image analysis, neuroscience, neurology and the intersections between these fields.

MATERIALS AND METHODS

The computational tools and data described in this work have been assembled and made available for download at https://www.nmr.mgh.harvard.edu/lab/mripredict. This website includes instructions and data to reproduce the results presented in this manuscript.

Data

In our experiments, we analyzed data from over 2,800 individuals obtained from six large clinical neuroimaging studies: the Alzheimer's Disease Neuroimaging Initiative, or ADNI (Jack *et al.*, 2008), the Open-Access Series of Imaging Studies (OASIS, oasis-brains.org) (Marcus et al., 2007b), the Autism Brain Imaging Data Exchange (ABIDE, tinyurl.com/fcon1000-abide), the Attention Deficit Hyperactivity Disorder (ADHD) sample from the ADHD-200 Consortium (Milham *et al.*, 2012) (tinyurl.com/fcon1000-adhd), the Center for Biomedical Research Excellence (COBRE) schizophrenia sample (tinyurl.com/fcon1000-cobre), and the MIND Clinical Imaging Consortium (MCIC) schizophrenia sample (Gollub *et al.*, 2013). Table 1 summarizes these data, which are publicly available for download via corresponding websites. We employed the T1-weighted structural brain MRI scans, demographic data (age and gender), site information, and clinical assessments in our analyses. For details of these data, we refer the reader to the associated studies.

We restricted all our analyses to the subjects for which the automatic image processing steps of FreeSurfer (see next sub-section) completed successfully. In the OASIS sample, the AD diagnosis was defined as CDR >= 1 and "AD mild" was defined as CDR > 0, which include subjects suffering from Mild Cognitive Impairment (MCI) (Petersen *et al.*, 1999) and not clinically demented. In the ADHD sample, cases were defined as those with evidence of non-typical development and an ADHD diagnosis, as per the ADHD 200 phenotypic key[1]. Schizophrenia (SCZ) cases in the Center for Biomedical Research Excellence (COBRE) sample were those identified as "Patient" in the COBRE phenotypic key. The ABIDE analyses were restricted to subjects who were at least 10 years old, since we were more confident that the imaging measurements automatically computed from scans in this age group were reliable. In the ABIDE sample, cases were defined as those having a non-zero diagnostic group entry in the phenotype table.

In addition to the binary clinical diagnosis (patient versus control), we analyzed continuous measures derived from non-imaging data (age, mini-mental state exam –MMSE- score, and cerebro-spinal fluid based amyloid-$\beta_{1\text{-}42}$, -CSF-$A\beta_{1\text{-}42}$–). Table 2 provides a list of all (binary and continuous) target variables along with additional information regarding group characteristics. For age, we employed only the control subjects within each dataset. In the ABIDE data, we restricted the age sample to the largest healthy cohort from a single site. The other two continuous variables, MMSE and CSF $A\beta_{1\text{-}42}$ levels, are markers of dementia, and demonstrate meaningful variation across clinical groups, but not necessarily within controls. Hence, for these variables, we combined data across clinical groups.

MRI Processing

We used FreeSurfer (freesurfer.nmr.mgh.harvard.edu) (Fischl, 2012) -version 5.1 – a freely available, widely used and extensively validated brain MRI analysis software package - to process the structural brain MRI scans and compute morphological measurements. The FreeSurfer pipeline is fully automatic and includes steps to compute a representation of the cortical surface between white and gray matter, a representation of the pial surface (Dale et

[1] http://fcon_1000.projects.nitrc.org/indi/adhd200/general/ADHD-200_PhenotypicKey.pdf.

al., 1999; Fischl et al., 1999a), and a segmentation of white matter regions; to perform skull stripping, B1 bias field correction, nonlinear registration of the cortical surface of an individual with a stereotaxic atlas (Fischl et al., 1999b), labeling of regions of the cortical surface (Fischl *et al.*, 2004), and labeling of subcortical brain structures (Fischl *et al.*, 2002). Furthermore, for each MRI scan, FreeSurfer automatically computes subject-specific thickness measurements across the entire cortical mantle and within anatomically defined cortical regions of interest (ROIs), volume estimates of a wide range of sub-cortical structures and estimates of the intra-cranial volume (ICV) and measures of image quality, such as white-matter signal to noise ratio (WM-SNR), which is computed based on the noise level (standard deviation of intensities) within the white matter.

In our analyses, we defined four sets of features to be used by the prediction models.

1. Feature set 1 (*aseg; 45 dimensional vector)*: Volumes of the 45 anatomical structures saved as *stats/aseg.stats* under the FreeSurfer subject directory, which were normalized with each subject's ICV to account for head size variation. The structures we used are: Left and right cerebral white matter, cerebral cortex, lateral ventricle, inferior lateral ventricle, cerebellum white matter, cerebellum cortex, thalamus proper, caudate, putamen, pallidum, hippocampus, and amygdala, plus the 3^{rd} and 4^{th} ventricles.
2. Feature set 2 (*aparc; 68 dimensional vector)*: Average thickness within the following cortical parcellations (saved as *stats/lh.aparc.stats* and *stats/rh.aparc.stats* under the FreeSurfer subject directory. There are 34 measurements per hemisphere). Superior frontal, rostral middle frontal, caudal middle frontal, pars opercularis, pars triangularis, pars orbitalis, lateral orbitofrontal, medial orbitofrontal, precentral, paracentral, frontal pole, superior parietal, inferior parietal, supramarginal, postcentral, precuneus, superior temporal, middle temporal, inferior temporal, banks of the superior temporal sulcus, fusiform, transverse temporal, entorhinal, temporal pole, parahippocampal, lateral occipital, lingual, cuneus, pericalcarine, rostral anterior frontal, caudal anterior frontal, posterior parietal, isthmus parietal, and insula.
3. Feature set 3 (*aparc+aseg; 113 dimensional vector)*: The union of the first two feature sets.
4. Feature set 4 (*thick; 20,484 dimensional vector*): Cortical thickness values sampled onto the *fsaverage5* template (10,242 vertices per hemisphere) and smoothed on the surface with an approximate Gaussian kernel (Han *et al.*, 2006) of a full-width-half-max (FWHM) of 5mm.

Multivariate Pattern Analysis Algorithms

We employed publicly available implementations of three different classes of MVPA algorithms: Support Vector Machines, Neighborhood Approximation Forests, and Relevance Vector Machines. These three algorithms were selected because they have been applied to neuroimage data in prior studies and represent a wide range of methods; each algorithm was derived using a different modeling approach and relying on distinct

assumptions about the data. We emphasize that there is a rich pool of potential algorithms that can be used on these data and we hope that by publicly distributing the data[2] we used in the presented analyses, we will enable other researchers to test, benchmark and publicize other method(s), thus allowing the exploration of a much wider class of machine learning algorithms than we could achieve by our own means. Our primary experiment and associated analyses were constrained to the following three algorithms.

1. The Support Vector Machine (SVM) is one of the most popular generic machine learning methods (Cortes and Vapnik, 1995; Scholkopf and Smola, 2002). In our experiments we used the publicly available implementation LibSVM (csie.ntu.edu.tw/~cjlin/libsvm). We employed the linear kernel, which has been demonstrated to yield good accuracy in prior neuroimaging studies. The hyper-parameters were optimized using a ("nested") cross-validation loop over the training dataset (using the "grid.py" tool available on the LibSVM website). We trained the SVM model for probability estimates. These estimates are directly used for the ROC analysis and thresholded at p=0.5 to compute the correct classification ratio.

2. The Neighborhood Approximation Forest (NAF) (www.nmr.mgh.harvard.edu/~enderk/software.html) (Konukoglu *et al.*, 2013) is a generic variant of random decision forests (Criminisi *et al.*, 2011) that can be applied to regression and classification without any modification of the underlying algorithm. The underlying principle of NAF is to approximate the "closest" training images to a given test image. The proximity between images is defined based on the variable of interest, such as diagnosis. During training, NAF learns to estimate the closest neighbors based on the image-derived measurements, such as ROI volumes or cortical thickness measurements. For a test image, NAF estimates its closest neighbors within the training set along with a weight associated with each neighbor indicating its approximate proximity to the test image. The prediction is then given as the weighted average of the labels of these closest neighbors. To identify the number of closest neighbors used in prediction, we ran a "nested" cross-validation on the training dataset only, similar to our SVM implementation. The remaining hyper-parameters of NAF were set heuristically based on experiments provided in previous publications (Konukoglu *et al.*, 2013). These are: number of trees = 800, maximum tree depth = 12, stopping criteria = 10 samples and number of random samples per node = 20 for feature sets 1–3 and 1000 for feature set 4.

3. The Relevance Voxel Machine (RVoxM, tinyurl.com/rvoxm)(Sabuncu and Van Leemput, 2012), is an adaptation of the Bayesian Relevance Vector Machine (RVM)(Tipping, 2001) customized to handle image data. The RVM model assumes that the target variable is a noisy observation of a linear weighted sum of the feature data. For regression, the noise is an additive Gaussian model. For classification, a logistic link function is used. RVM builds on MacKay's Automatic Relevance Determination (ARD) framework (MacKay, 1992) and employs a Gaussian prior on the weight parameters, which are (approximately) integrated (or

[2] https://www.nmr.mgh.harvard.edu/lab/mripredict

marginalized) out during learning and prediction. RVM's prior encourages sparsity, i.e., a small number of non-zero weights. RVoxM modifies this prior to also encourage spatial smoothness. We note that for Feature set 4 (*thick*), we utilized the neighborhood structure of the *fsaverage5* surface mesh to define the Laplacian matrix that encourages the weights to be spatially smooth. For feature sets 1–3, we used no spatial smoothness, i.e., Laplacian term. Thus for the *aseg* and *aparc* features, the RVoxM model was essentially equivalent to a RVM model on the feature dimensions. We therefore refer to this algorithm as RVM throughout the manuscript.

In total, there were 12 (=3×4) different combinations of algorithm and image feature pairs, or MVPA models, which we applied to the data.

Univariate Prediction Models

Most common image-derived structural biomarkers are univariate descriptions of morphology, such as the volume of a region of interest (ROI). To implement such a biomarker, we used the *aseg* and *aparc* features, which are volume and thickness estimates of anatomical ROIs. These measurements, such as the volume of the hippocampus or size of ventricles, represent most of the classical MRI-derived biomarkers associated with neurological disorders, such as dementia or schizophrenia.

To identify the univariate predictive marker for each variable of interest, we conducted the following unbiased, data-driven analysis. At each cross-validation session, we determined the feature (out of the 113 aparc+aseg measurements) that was most significantly associated with the variable of interest on the training data (based on t-test between two samples for classification; based on Pearson's linear correlation for regression). Next, we computed the affine transformation (scale and shift) that converted the corresponding measurements to best agree with the training labels, which was assessed via the correct classification ratio (the binary prediction was computed by thresholding at zero) or mean squared error. For classification, the scale was restricted to −1/std(measurements) or 1/std(measurements), where the standard deviation was computed on the training sample. The index of the ROI (i.e., identity of the feature), and optimal affine parameters were then saved as the univariate prediction model, to be used on test data. Finally, predictions were computed on the test data by applying the affine transformation to the corresponding measurements. The agreement between these values and ground truth was then computed as in the MVPA case. This whole procedure was repeated across the different cross-validation sessions.

Cross-validation

To quantify the accuracy of an image-based prediction model we utilized 5-fold cross-validation on each sample. For classification, we conducted stratified and balanced cross-validation (Parker *et al.*, 2007), where each partition contained the same number of cases and controls (i.e., was balanced). In each partition, the two groups were also matched based on age, gender and site data, where appropriate. For regression, we partitioned the data into 5 (almost) equally sized groups (if needed, the last partition was allowed to be larger than

the rest to account for all subjects). In each fold, each partition was treated as test data and the remaining subjects constituted training data.

In cross-validation, prediction accuracy was computed by aggregating predictions across the 5 folds, which yielded a single prediction per sample. Binary classification accuracy was then quantified using correct classification rate (CCR), i.e., the empirical ratio of correct predictions across all samples. Regression accuracy was measured with the root mean squared error (RMSE) of the predictions. To normalize RMSE scores, we divided by the range of the target variable in the sample. This allowed a comparison across different variables with different units.

The statistical significance of prediction accuracies for the classification problems were computed using DeLong's method (DeLong *et al.*, 1988) based on the receiver operating characteristic (ROC) analysis. DeLong's test is a non-parametric statistical test for comparing areas-under-the-curve (AUC) for two ROC curves. It is based on estimating an AUC value (which we computed using Matlab's perfcurve function) and an associated variance using the probabilistic predictions for positive and negative samples. A z-score, which has a standard normal distribution, can then be computed for the AUC estimate using the calculated variance and the fact that under the null AUC should equal to 0.5. To compute the p-values we performed a one-sided test on these resulting z-scores. We choose to use the ROC analysis to compute statistical significance because it captures more information than CCR, in particular about how the probabilistic predictions are distributed.

In the regression problems the statistical significance values were computed using Pearson's linear correlation coefficient, r, corresponding t-test.

To assess the uncertainty in the cross-validation based estimates of performance metrics, we repeated the 5-fold cross-validation procedure for the best MVPA models using 100 different 5-fold partitions. The best MVPA models were identified as the ones that yielded the predictions that were most significantly associated with the ground truth variables on the first 5-fold cross-validation (these results are reported in Fig. 1). For each 5-fold partitioning, we computed the cross-validation performance metric, yielding a distribution of 100 values. For the results of Fig. 2 and 4, we computed the mean prediction accuracy as the average of these 100 values and the 95% confidence interval was computed by excluding the highest and lowest two values.

Mass-univariate Analysis of Thickness Maps

We conducted a mass-univariate analysis to map regions where cortical thickness is associated with clinical variables of interest. For this analysis, we used the thickness values sampled onto the highest resolution template, *fsaverage*, which contains over 140k vertices on each hemisphere, and smoothed on the cortical surface with a Gaussian-like filter of a 10 mm FWHM. We then applied a general linear model at each vertex, where the outcome was thickness and the independent variables were age, gender and the clinical variable. The p-value associated with the clinical variables was then saved for each vertex (see Fig. 3). When identifying cortical areas of significant associations, we applied the false discovery rate (Benjamini and Hochberg, 1995) (FDR, q = 0.05) to correct for multiple comparisons.

The total area of significant associations was then computed as the sum of the areas corresponding to the significant vertices in *fsaverage*.

Statistical Analyses of the Influence of Measurement and Algorithm Choice

To gain further insights into the impact of the measurement type and MVPA algorithm on prediction accuracy, we used the 5-fold cross-validation performance estimates presented in Fig. 1. We employed the non-parametric Friedman's test (Wolfe and Hollander, 1973) to assess the difference across measurement types and algorithm classes, adjusting for variation across variables and treating the nuisance factor (e.g., algorithm choice when assessing image feature) as a replicated measurement.

To assess whether the algorithm or image feature design decision had a bigger impact on prediction accuracy, we computed range data as follows. For each variable, we computed the *algorithm range* as the difference between the best and worst performance metrics across the three algorithms (SVM, RVM and NAF), while fixing the feature type. These values were then averaged over feature types. Similarly, for each variable, the *feature range* was defined as the difference between the best and worst performance metrics across the four feature types, while fixing the algorithm type. These values were then average over the algorithms (see Supplementary Fig. S4). We performed the nonparametric Wilcoxon signed rank test (Wolfe and Hollander, 1973) on the paired range values to assess the significance of the difference between the feature and algorithm effects. For the binary variables, the feature range was significantly larger than the algorithm range (P=0.008). For regression, however, the two effects were statistically equivalent (P=0.36).

RESULTS

There was significant variation in the sample sizes across datasets and variables (see Table 2). For example, for the Alzheimer's disease (AD) variable (clinical dementia rating, CDR, greater than or equal to 1), the ADNI sample provided 145 subjects per group, where as the OASIS sample offered only 25. Also, certain datasets were collected at multiple sites (e.g., 20 sites participated in the ABIDE study), whereas others, e.g., COBRE, were acquired at a single location.

Estimating prediction accuracy via cross-validation

To estimate the accuracy of all twelve MVPA models, we utilized a single 5-fold cross-validation on each sample (See Fig. 1. More detailed results are provided in Supplementary Fig. S1). These results revealed that all but two (ADHD diagnosis, and age in the ABIDE sample) of the examined variables exhibited some degree of *predictability* from brain MRI scans, i.e., there was at least one MVPA model that produced a prediction on test data that was statistically significantly associated with the ground truth label (P<1e-3). In practice, there were multiple MVPA models that were significantly associated with each predictable variable, not just one.

Today, most classical image-derived biomarkers are univariate, e.g., the size of a region of interest. To provide a comparison between MVPA models and classical markers, we also quantified the prediction performance of univariate models that use a single measurement,

e.g., volume of an anatomical structure. We applied the univariate models to the same hundred 5-fold cross-validations as the ones used for the MVPA models. Supplementary Table S1 lists the ROIs that were most frequently identified as univariate markers for each variable. Fig. 2 shows the estimated performance metrics for the MVPA and univariate models. For *all* variables, the performance metrics were significantly better for the MVPA model (all $P < 1e\text{-}4$, paired Wilcoxon signed rank test), although the performance boost varied across variables. For example, on the OASIS AD sample, the MVPA model yielded an improvement of more than 10% in Correct Classification Ratio (CCR), while the difference between the prediction accuracies of the MVPA and univariate models was modest for the ADNI:CSF-Aβ phenotype.

From Figures 1 and 2, we observe that there is a dramatic variation in prediction accuracies across datasets, target variables, image features, and algorithms. These results underscore the factors that influence image-based prediction, which include:

1. Biological footprint of the variable, or effect size,
2. Data quality, e.g., the amount of image noise,
3. Sample size,
4. The accuracy and relevance of image-derived measurements,
5. And the prediction algorithm.

In the following, we provide some analyses to gain insights into how these individual factors influence prediction performance.

Dissecting the influence of various factors on prediction performance

Arguably, the most significant determinant of how accurately one can predict a particular variable from a brain MRI scan is the biological footprint. This is observable from Fig. 1, where most of the variation in performance metrics is vertical, i.e., across variables. Fig. 3 illustrates this point further, where MVPA prediction accuracies are shown alongside results from a mass-univariate analysis that reveals the cortical thinning patterns of each disease. In each panel of Fig. 3, we present three variables, where the MVPA and mass-univariate analyses were conducted on samples of roughly the same size (Panel a: ADNI:AD, N=145; ADNI:MCI, N = 135; and ADHD, N = 150. Panel b: ADNI-75:AD, MCIC:SCZ, and ABIDE-75:ASD, each with 75 subjects per group) and commensurate MRI data quality (estimated white matter signal to noise ratio, WM-SNR, mean ± standard deviation. First panel: 16.8±4.2, 17.0±4.1, 16.8±2.3. Second panel: 18.9±3.0, 20.2±3.7, 19.7±3.4.). All MVPA results reported in Fig. 3 were computed with the RVM algorithm, using the cortical thickness maps (i.e., feature type 4). Hence, factors 2–5 have minimal influence on the variation in prediction performance within each panel. This leaves the biological footprint as the only factor that one would expect to largely determine prediction accuracy. The results of Fig. 3 provide compelling support for this hypothesis, since there is a strong agreement between prediction accuracy and the size of the cortical area significantly associated with the disease. AD clearly has the most prominent biological footprint on cortical thickness, which is followed by MCI and schizophrenia. Autism and ADHD seem to have very modest

footprints, which were not detectable using a mass-univariate method in these samples. Intriguingly, the MVPA analysis of the ABIDE:ASD sample demonstrated a significant global association between brain morphology and autism diagnosis (CCR:0.59, with 95% confidence interval [0.57–0.61]), which was not revealed by the mass-univariate analysis.

The influence of sample size on multivariate pattern analysis is twofold. Firstly, increasing training size should in general yield better models and thus improve prediction accuracy. Secondly, increasing test size will typically improve our confidence in the estimates of prediction accuracy, i.e., reduce uncertainty, which will in turn translate into improved statistical power, allowing us to detect more subtle associations. We observed both of these phenomena in our experiments, particularly for predicting age. There was a statistically significant association between sample size and prediction accuracy of age across samples (P=0.0011, Pearson correlation). Furthermore, the statistical significance associated with each sample was correlated with its size (Pearson r=0.88, P=0.02), exposing the strong link between the number of subjects and statistical power.

Finally, we examined the influence of the choice of image-derived measurements and machine learning algorithms. Our primary observation is that among the types of features and algorithms we considered (see Fig. 1 and Supplementary Fig. S1–S3), there was no globally optimal choice that produced the best results overall. However, for the binary phenotypes, feature type 2 (*aparc*) produced significantly worse results than the remaining three types of features (P=0.04), and the performances of the three MVPA algorithms were statistically indistinguishable (P=0.73). For regression, RVM produced inferior results than NAF and SVM (P= 7.4e-6), which were statistically equivalent. Feature types 3 and 4 offered statistically significantly better accuracy than the other two features (P=3.5e-4).

The next question we tackled was whether the algorithm or image feature design decision had a bigger impact on prediction accuracy. The results presented in Supplementary Fig. S3 revealed that for the binary classification cases we analyzed, although the algorithm decision was an important determinant, the choice of image feature had a significantly larger effect on prediction accuracy (P=0.008). For regression, however, both decisions had a statistically indistinguishable (P=0.36), yet large effect. Overall, these results suggest that among the ones we tested, there was no universally optimal choice of imaging measurements or machine learning tool that would produce the best prediction performance, although, these design choices had a substantial impact on accuracy.

Validation on independent datasets

Although, in theory, cross-validation provides an unbiased estimate of performance, validation on independent datasets remains to be the more realistic approach to quantifying generalization accuracy. Here we applied this strategy to four variables, for which we had multiple independent datasets: Alzheimer's disease diagnosis, schizophrenia diagnosis, age and MMSE score. For age, we chose to employ the OASIS and COBRE datasets, which offered a similar range in values.

The results presented in Fig. 4 revealed that all of the eight MVPA models that produced statistically significant predictions on cross-validation, further yielded statistically

significant predictions on independent validation datasets. However, for most models (all but the models of OASIS:AD and COBRE:SCZ), the prediction accuracies on the validation datasets were outside the 95% confidence intervals estimated via cross-validation. On the other hand, there was a strong agreement between the cross-validation and independent validation performances: the rankings of models based on the performance on the independent samples and those based on the estimated cross-validation accuracies were identical within regression and classification. These results suggest that cross-validation can be optimistic in estimating prediction performance, yet provides an informative upper bound.

DISCUSSION

The dramatic variability in the brain's structural anatomy is influenced by genetics, environmental factors, age, disease, and interactions between all these factors. The complexity of these mechanisms makes the problem of predicting diagnosis and clinically relevant variables from structural neuroimaging data very difficult. The problem is further complicated because of our limited understanding of clinical conditions, which introduces heterogeneity and noise into the definitions of the target variables. This phenotype contamination is particularly evident in neurology, where there is an abundance of heterogeneity within and overlap across clinical conditions. Yet, image-based prediction methods can be useful for demonstrating complex and subtle associations, while enabling more accurate individual-level clinical assessments, which in turn can help us refine our clinical definitions.

Multivariate models outperform univariate markers in prediction

Structural brain MRI-derived biomarkers are classically univariate, measuring the volume, size, or thickness of an anatomical ROI, including the whole brain. However, recent studies have demonstrated that many neurological conditions are associated with large-scale networks of distributed regions (Seeley *et al.*, 2009). This suggests that aggregating information across multiple regions within the associated network should improve the sensitivity and specificity of brain biomarkers. Our results generalize prior studies that make similar observations, e.g., (Westman et al., 2011), to a range of target variables. In all our analyses, MVPA models offered a statistically significant boost in prediction performance as assessed via cross-validation. This improvement was reflected as a 5–10% increase in correct classification ratio for binary variables.

An array of variables can be predicted from structural neuroimaging data

Our results demonstrated that MVPA models produce predictions that are statistically significantly associated with the ground truth for a range of variables. However, there is a dramatic variation in the accuracies of these predictions, which determines the utility of these models. On one end of the spectrum, we have autism, which our cross-validation suggests can correctly be discriminated from a healthy state about 59% of the time (95% confidence interval [0.57–0.61]). This, by itself, is unlikely to be useful for making individual-level predictions, especially in the clinical setting, where the problem is particularly more challenging due to sample heterogeneity and lower data quality. However,

it can be used as one line of evidence among an array of other observations. Furthermore, this MVPA result reveals a statistically significant association between brain anatomy and autism, which is so subtle that it cannot be detected via a more traditional mass-univariate analysis. On the other hand of the spectrum, we have Alzheimer's diagnosis and age, which can be predicted very accurately (86% accuracy in discriminating from healthy controls, and root mean squared error less than 9 years, respectively). Thus, these models by themselves might be useful for individualized prognosis in the clinical setting. Age is a particularly interesting variable, which might be informative for detecting deviations from normal aging or healthy development (e.g. when the subject's predicted brain age is substantially different from his/her chronological age).

The results we present in this study, in general, are consistent with prior studies that report structural MRI (sMRI) based clinical predictions. Our AD, MCI, age, and MMSE prediction results are in strong agreement with state-of-the-art structural MRI-based predictions computed on the ADNI data, e.g., as reported in (Cuingnet et al., 2011; Sabuncu and Van Leemput, 2012; Stonnington et al., 2010). For schizophrenia, the classification accuracy we present, which is roughly around 70%, is in line with a previously reported large-scale multi-site MRI-based prediction study (Nieuwenhuis et al., 2012). Finally, the autism prediction accuracy we obtain, which is about 60%, is congruent with the results obtained with resting state functional MRI (rs-fMRI) data on the same ABIDE dataset (Nielsen et al., 2013). This last result suggests that both rs-fMRI and sMRI offer similar prediction accuracy for autism.

Factors that influence prediction accuracy

There are at least five factors that determine prediction accuracy: 1) biological footprint, 2) sample size, 3) data quality, 4) image measurements, and 5) prediction algorithm. We believe that the footprint of the underlying biological process, as captured by the imaging data, is the most important determinant of prediction performance. One way of measuring this footprint is via normalizing the remaining factors, i.e., to compare the footprint of different variables, one could conduct a MVPA prediction analysis, where the last four factors are roughly standardized (same sample size, data quality, imaging measurements and prediction algorithm). We applied this strategy to our data, which provided a clear demonstration of the variable footprint sizes of the different clinical conditions we considered.

Image measurements and prediction algorithms, on the other hand, also have a significant impact on prediction accuracy. Our results further suggest that the former factor has an impact that is at least as important as the latter. Varying these design decisions can lead to radically different conclusions, as our results revealed. However, our analyses also suggest that there is no universally optimal choice for structural neuroimaging. This makes benchmark studies, such as the present, particularly important, since they provide an objective framework for comparing and assessing image processing and analysis methods for different clinical conditions of interest. In this study, we analyzed a small set of possible machine learning algorithms and image measurement types. Future studies will explore

alternative algorithms and image-derived features to identify the optimal design choices for each individual problem.

One particular issue that one needs to pay special attention to is the uncertainty in the performance assessments (Japkowicz and Shah, 2011). We observed a considerable variation between the prediction accuracies estimated using different 5-fold partitions of the data. To quantify this, we employed 100 different partitions of the data, over which performance metric statistics (e.g., average, confidence interval, etc.) were computed. All these lists (i.e., the subject ID's for each fold of each partition) are made publicly available, so that alternative methods can use these data to estimate the prediction accuracy and corresponding uncertainty. We will further distribute the individual predictions computed for each list using each MVPA model. These data will enable a fair and objective comparison across methods.

Validation on independent datasets

Although cross-validation offers a useful strategy for quantifying prediction accuracy, we found that its estimates are often optimistic. We believe this arises due to the variation in (i) the data acquisition protocol, (ii) composition of the populations, and (iii) the application of the diagnostic criteria and/or clinical tests. For example, scan parameters, such as field strength, usually vary and this alters the distributions of the imaging measurements. Furthermore, the precise definitions of the clinical conditions can also change, especially across different clinical centers. These issues can be minimized by standardizing the imaging and clinical protocols. However, in most practical scenarios, inter-site variability will remain a major challenge and impact the clinical application of image-based prediction models. Therefore, we believe using different datasets independently collected at different centers is critical for obtaining a realistic estimate of the generalization accuracy of a prediction model.

Considering and Probing the Underlying Biology

Our experiments suggest that the type of measurements derived from the imaging data have a substantial influence on prediction accuracy. This observation highlights the significance of the utilized image processing tools. Furthermore, it indicates that intelligent feature selection methods might yield improved prediction performance. Feature (variable) selection is an active area of research in machine learning (Guyon and Elisseeff, 2003; Jain and Zongker, 1997; Saeys et al., 2007) and is also being investigated in the context of neuroimaging, e.g. (Nie et al., 2008; Pereira and Botvinick, 2011; Plant et al., 2010; Rondina et al., 2013; Wang et al., 2011; Wang et al., 2006).

While obtaining improved and more efficient prediction is the main motivation of feature selection methods (Chu et al., 2012), by identifying a small, interpretable subset of relevant features, they might also lead to biological insights. From this perspective, feature learning is intimately related to the recent line of research that aims to measure the statistical significance of each variable in a discriminative (predictive) model, e.g., (Gaonkar and Davatzikos, 2013; Lockhart et al., 2012; Meinshausen and Buhlmann, 2010; Rondina et al., 2013). Rather than focusing on statistical significance, which assumes a null hypothesis, an

alternative approach is to quantify the importance of each variable for prediction, e.g., (Sonnenburg et al., 2008; Strobl et al., 2008; Zien et al., 2009). Such methods promise to allow us to probe the prediction models we build and make inferences about the underlying biology.

CONCLUSION

We presented the largest empirical benchmark MVPA study in structural neuroimaging. Our results demonstrate that one can predict a range of clinically relevant variables from structural brain MRI scans with varying degrees of accuracy. MVPA models offer more accurate predictions than univariate markers, such as the volume of a ROI, though the choice of the feature set and machine-learning algorithm has a significant impact on prediction performance. We found no universally optimal MVPA method that would yield the best prediction. Furthermore the biological footprint of the phenotype seems to be the most important determinant of prediction accuracy. Future MVPA studies can compare alternative methods against the published results using the public datasets and distributed cross-validation lists, while properly accounting for the uncertainty in performance estimates.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This research was carried out in whole or in part at the Athinoula A. Martinos Center for Biomedical Imaging at the Massachusetts General Hospital, using resources provided by the Center for Functional Neuroimaging Technologies, P41EB015896, a P41 Biotechnology Resource Grant supported by the National Institute of Biomedical Imaging and Bioengineering (NIBIB), National Institutes of Health. Dr. Sabuncu received support from an AHAF (BrightFocus) Alzheimer's Disease pilot grant (AHAF A2012333), and an NIH K25 grant (NIBIB 1K25EB013649-01).

References

Ashburner J, Friston KJ. Voxel-based morphometry,Äîthe methods. Neuroimage. 2000; 11:805–821. [PubMed: 10860804]

Batmanghelich, N.; Taskar, B.; Davatzikos, C. Information Processing in Medical Imaging. Springer; 2009. A general and unifying framework for feature construction, in image-based pattern classification; p. 423-434.

Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society. Series B (Methodological). 1995:289–300.

Brown MR, Sidhu GS, Greiner R, Asgarian N, Bastani M, Silverstone PH, Greenshaw AJ, Dursun SM. ADHD-200 Global Competition: diagnosing ADHD using personal characteristic data can outperform resting state fMRI measurements. Frontiers in systems neuroscience. 2012:6. [PubMed: 22438838]

Chang CC, Lin CJ. LIBSVM: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology (TIST). 2011; 2:27.

Cho Y, Seong JK, Jeong Y, Shin SY. Individual subject classification for Alzheimer's disease based on incremental learning using a spatial frequency representation of cortical thickness data. Neuroimage. 2012; 59:2217–2230. [PubMed: 22008371]

Chu C, Hsu AL, Chou KH, Bandettini P, Lin C. Does feature selection improve classification accuracy? Impact of sample size and feature selection on classification using anatomical magnetic resonance images. Neuroimage. 2012; 60:59–70. [PubMed: 22166797]

Cortes C, Vapnik V. Support-vector networks. Machine Learning. 1995; 20:273–297.

Costafreda SG, Chu C, Ashburner J, Fu CH. Prognostic and diagnostic potential of the structural neuroanatomy of depression. PLoS One. 2009; 4:e6353. [PubMed: 19633718]

Criminisi, A.; Shotton, J.; Konukoglu, E. Tech. Rep MSRTR-2011-114. Vol. 5. Microsoft Research Cambridge; 2011. Decision forests for classification, regression, density estimation, manifold learning and semi-supervised learning; p. 12

Cuingnet R, Gerardin E, Tessieras J, Auzias G, Lehericy S, Habert MO, Chupin M, Benali H, Colliot O. Automatic classification of patients with Alzheimer's disease from structural MRI: a comparison of ten methods using the ADNI database. Neuroimage. 2011; 56:766–781. [PubMed: 20542124]

Dale AM, Fischl B, Sereno MI. Cortical surface-based analysis: I. Segmentation and surface reconstruction. Neuroimage. 1999; 9:179–194. [PubMed: 9931268]

Davatzikos C, Resnick SM, Wu X, Parmpi P, Clark C. Individual patient diagnosis of AD and FTD via high-dimensional pattern classification of MRI. Neuroimage. 2008; 41:1220–1227. [PubMed: 18474436]

Davatzikos C, Xu F, An Y, Fan Y, Resnick SM. Longitudinal progression of Alzheimer's-like patterns of atrophy in normal older adults: the SPARE-AD index. Brain. 2009; 132:2026–2035. [PubMed: 19416949]

DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics. 1988:837–845. [PubMed: 3203132]

Duchesnay E, Cachia A, Roche A, Rivière D, Cointepas Y, Papadopoulos-Orfanos D, Zilbovicius M, Martinot JL, Régis J, Mangin JF. Classification based on cortical folding patterns. Medical Imaging, IEEE Transactions on. 2007; 26:553–565.

Duchesne S, Rolland Y, Verin M. Automated computer differential classification in Parkinsonian syndromes via pattern analysis on MRI. Academic radiology. 2009; 16:61–70. [PubMed: 19064213]

Ecker C, Rocha-Rego V, Johnston P, Mourao-Miranda J, Marquand A, Daly EM, Brammer MJ, Murphy C, Murphy DG. Investigating the predictive value of whole-brain structural MR scans in autism: a pattern classification approach. Neuroimage. 2010; 49:44–56. [PubMed: 19683584]

Fan Y, Shen D, Gur RC, Gur RE, Davatzikos C. COMPARE: classification of morphological patterns using adaptive regional elements. Medical Imaging, IEEE Transactions on. 2007; 26:93–105.

Feinstein A, Roy P, Lobaugh N, Feinstein K, O,ÄôConnor P, Black S. Structural brain abnormalities in multiple sclerosis patients with major depression. Neurology. 2004; 62:586–590. [PubMed: 14981175]

Fischl B. FreeSurfer. Neuroimage. 2012; 62:774–781. [PubMed: 22248573]

Fischl B, Salat DH, Busa E, Albert M, Dieterich M, Haselgrove C, van der Kouwe A, Killiany R, Kennedy D, Klaveness S. Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. Neuron. 2002; 33:341–355. [PubMed: 11832223]

Fischl B, Sereno MI, Dale AM. Cortical surface-based analysis: II: Inflation, flattening, and a surface-based coordinate system. Neuroimage. 1999a; 9:195–207. [PubMed: 9931269]

Fischl B, Sereno MI, Tootell RB, Dale AM. High-resolution intersubject averaging and a coordinate system for the cortical surface. Human brain mapping. 1999b; 8:272–284. [PubMed: 10619420]

Fischl B, Van Der Kouwe A, Destrieux C, Halgren E, Ségonne F, Salat DH, Busa E, Seidman LJ, Goldstein J, Kennedy D. Automatically parcellating the human cerebral cortex. Cerebral cortex. 2004; 14:11–22. [PubMed: 14654453]

Frisoni GB, Fox NC, Jack CR, Scheltens P, Thompson PM. The clinical use of structural MRI in Alzheimer disease. Nature Reviews Neurology. 2010; 6:67–77.

Friston K, Chu C, Mourao-Miranda J, Hulme O, Rees G, Penny W, Ashburner J. Bayesian decoding of brain images. Neuroimage. 2008; 39:181–205. [PubMed: 17919928]

Friston KJ, Holmes AP, Worsley KJ, Poline JÄ, Frith CD, Frackowiak RS. Statistical parametric maps in functional imaging: a general linear approach. Human brain mapping. 1994; 2:189–210.

Gaonkar B, Davatzikos C. Analytic estimation of statistical significance maps for support vector machine based multi-variate image analysis and classification. Neuroimage. 2013; 78:270–283. [PubMed: 23583748]

Gollub RL, Shoemaker JM, King MD, White T, Ehrlich S, Sponheim SR, Clark VP, Turner JA, Mueller BA, Magnotta V. The MCIC Collection: A Shared Repository of Multi-Modal, Multi-Site Brain Image Data from a Clinical Investigation of Schizophrenia. Neuroinformatics. 2013:1–22. [PubMed: 23224666]

Guyon I, Elisseeff A. An introduction to variable and feature selection. The Journal of Machine Learning Research. 2003; 3:1157–1182.

Han X, Jovicich J, Salat D, van der Kouwe A, Quinn B, Czanner S, Busa E, Pacheco J, Albert M, Killiany R. Reliability of MRI-derived measurements of human cerebral cortical thickness: the effects of field strength, scanner upgrade and manufacturer. Neuroimage. 2006; 32:180–194. [PubMed: 16651008]

Ho BC, Andreasen NC, Nopoulos P, Arndt S, Magnotta V, Flaum M. Progressive structural brain abnormalities and their relationship to clinical outcome: a longitudinal magnetic resonance imaging study early in schizophrenia. Archives of general psychiatry. 2003; 60:585. [PubMed: 12796222]

Jack CR, Bernstein MA, Fox NC, Thompson P, Alexander G, Harvey D, Borowski B, Britson PJ, Whitwell LJ, Ward C. The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods. Journal of Magnetic Resonance Imaging. 2008; 27:685–691. [PubMed: 18302232]

Jain A, Zongker D. Feature selection: Evaluation, application, and small sample performance. Pattern Analysis and Machine Intelligence, IEEE Transactions on. 1997; 19:153–158.

Japkowicz, N.; Shah, M. Evaluating learning algorithms: a classification perspective. Cambridge University Press; 2011.

Kawasaki Y, Suzuki M, Kherif F, Takahashi T, Zhou SY, Nakamura K, Matsui M, Sumiyoshi T, Seto H, Kurachi M. Multivariate voxel-based morphometry successfully differentiates schizophrenia patients from healthy controls. Neuroimage. 2007; 34:235–242. [PubMed: 17045492]

Kloppel S, Abdulkadir A, Jack CR Jr, Koutsouleris N, Mourão-Miranda J, Vemuri P. Diagnostic neuroimaging across diseases. Neuroimage. 2012; 61:457–463. [PubMed: 22094642]

Kloppel S, Chu C, Tan G, Draganski B, Johnson H, Paulsen J, Kienzle W, Tabrizi S, Ashburner J, Frackowiak R. Automatic detection of preclinical neurodegeneration Presymptomatic Huntington disease. Neurology. 2009; 72:426–431. [PubMed: 19188573]

Kloppel S, Stonnington CM, Chu C, Draganski B, Scahill RI, Rohrer JD, Fox NC, Jack CR, Ashburner J, Frackowiak RS. Automatic classification of MR scans in Alzheimer's disease. Brain. 2008; 131:681–689. [PubMed: 18202106]

Konukoglu E, Glocker B, Zikic D, Criminisi A. Neighbourhood Approximation using Randomized Forests. Medical image analysis. 2013

Koutsouleris N, Meisenzahl EM, Davatzikos C, Bottlender R, Frodl T, Scheuerecker J, Schmitt G, Zetzsche T, Decker P, Reiser M. Use of neuroanatomical pattern classification to identify subjects in at-risk mental states of psychosis and predict disease transition. Archives of general psychiatry. 2009; 66:700. [PubMed: 19581561]

Kriegeskorte N, Goebel R, Bandettini P. Information-based functional brain mapping. Proceedings of the National Academy of Sciences of the United States of America. 2006; 103:3863–3868. [PubMed: 16537458]

Lao Z, Shen D, Xue Z, Karacali B, Resnick SM, Davatzikos C. Morphological classification of brains via high-dimensional shape transformations and machine learning methods. Neuroimage. 2004; 21:46–57. [PubMed: 14741641]

Lerch JP, Pruessner J, Zijdenbos AP, Collins DL, Teipel SJ, Hampel H, Evans AC. Automated cortical thickness measurements from MRI can accurately separate Alzheimer's patients from normal elderly controls. Neurobiology of aging. 2008; 29:23–30. [PubMed: 17097767]

Liu F, Guo W, Yu D, Gao Q, Gao K, Xue Z, Du H, Zhang J, Tan C, Liu Z. Classification of different therapeutic responses of major depressive disorder with multivariate pattern analysis method based on structural MR scans. PLoS One. 2012; 7:e40968. [PubMed: 22815880]

Lockhart, R.; Taylor, J.; Tibshirani, RJ.; Tibshirani, R. A significance test for the lasso. 2012.

MacKay DJ. The evidence framework applied to classification networks. Neural computation. 1992; 4:720–736.

Marcus DS, Olsen TR, Ramaratnam M, Buckner RL. The extensible neuroimaging archive toolkit. Neuroinformatics. 2007a; 5:11–33. [PubMed: 17426351]

Marcus DS, Wang TH, Parker J, Csernansky JG, Morris JC, Buckner RL. Open Access Series of Imaging Studies (OASIS): cross-sectional MRI data in young, middle aged, nondemented, and demented older adults. Journal of Cognitive Neuroscience. 2007b; 19:1498–1507. [PubMed: 17714011]

Meinshausen N, Buhlmann P. Stability selection. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2010; 72:417–473.

Milham MP, Fair D, Mennes M, Mostofsky SH. The ADHD-200 consortium: a model to advance the translational potential of neuroimaging in clinical neuroscience. Frontiers in systems neuroscience. 2012; 6:62. [PubMed: 22973200]

Mitchell TM, Hutchinson R, Niculescu RS, Pereira F, Wang X, Just M, Newman S. Learning to decode cognitive states from brain images. Machine Learning. 2004; 57:145–175.

Mourao-Miranda J, Bokde AL, Born C, Hampel H, Stetter M. Classifying brain states and determining the discriminating activation patterns: support vector machine on functional MRI data. Neuroimage. 2005; 28:980–995. [PubMed: 16275139]

Mourao-Miranda J, Reinders A, Rocha-Rego V, Lappin J, Rondina J, Morgan C, Morgan K, Fearon P, Jones P, Doody G. Individualized prediction of illness course at the first psychotic episode: a support vector machine MRI study. Psychological medicine. 2012; 42:1037. [PubMed: 22059690]

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

Nie K, Chen JH, Yu HJ, Chu Y, Nalcioglu O, Su MY. Quantitative analysis of lesion morphology and texture features for diagnostic prediction in breast MRI. Academic radiology. 2008; 15:1513–1525. [PubMed: 19000868]

Nielsen JA, Zielinski BA, Fletcher PT, Alexander AL, Lange N, Bigler ED, Lainhart JE, Anderson JS. Multisite functional connectivity MRI classification of autism: ABIDE results. Frontiers in human neuroscience. 2013:7. [PubMed: 23372547]

Nieuwenhuis M, van Haren NE, Hulshoff Pol HE, Cahn W, Kahn RS, Schnack HG. Classification of schizophrenia patients and healthy controls from structural MRI scans in two large independent samples. Neuroimage. 2012; 61:606–612. [PubMed: 22507227]

Nouretdinov I, Costafreda SG, Gammerman A, Chervonenkis A, Vovk V, Vapnik V, Fu CH. Machine learning classification with confidence: application of transductive conformal predictors to MRI-based diagnostic and prognostic markers in depression. Neuroimage. 2011; 56:809–813. [PubMed: 20483379]

Parker B, Günter S, Bedo J. Stratification bias in low signal microarray studies. BMC bioinformatics. 2007; 8:326. [PubMed: 17764577]

Pereira, F.; Botvinick, M. Classification of functional magnetic resonance imaging data using informative pattern features. Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining; ACM; 2011. p. 940-946.

Petersen RC, Smith GE, Waring SC, Ivnik RJ, Tangalos EG, Kokmen E. Mild cognitive impairment: clinical characterization and outcome. Archives of neurology. 1999; 56:303. [PubMed: 10190820]

Plant C, Teipel SJ, Oswald A, Böhm C, Meindl T, Mourao-Miranda J, Bokde AW, Hampel H, Ewers M. Automated detection of brain atrophy patterns based on MRI for the prediction of Alzheimer's disease. Neuroimage. 2010; 50:162–174. [PubMed: 19961938]

Rondina, J.; Hahn, T.; de Oliveira, L.; Marquand, A.; Dresler, T.; Leitner, T.; Fallgatter, A.; Shawe-Taylor, J.; Mourao-Miranda, J. SCoRS-a method based on stability for feature selection and mapping in neuroimaging. 2013.

Sabuncu M, Van Leemput K. The Relevance Voxel Machine (RVoxM): A Self-tuning Bayesian Model for Informative Image-based Prediction. IEEE Transactions on Medical Imaging. 2012

Saeys Y, Inza Ia, Larrañaga P. A review of feature selection techniques in bioinformatics. Bioinformatics. 2007; 23:2507–2517. [PubMed: 17720704]

Schnack HG, Nieuwenhuis M, van Haren NE, Abramovic L, Scheewe TW, Brouwer RM, Hulshoff Pol HE, Kahn RS. Can structural MRI aid in clinical classification? A machine learning study in two independent samples of patients with schizophrenia, bipolar disorder and healthy subjects. Neuroimage. 2014; 84:299–306. [PubMed: 24004694]

Scholkopf, B.; Smola, AJ. Learning with kernels: Support vector machines, regularization, optimization, and beyond. the MIT Press; 2002.

Scott A, Courtney W, Wood D, De la Garza R, Lane S, King M, Wang R, Roberts J, Turner JA, Calhoun VD. COINS: an innovative informatics and neuroimaging tool suite built for large heterogeneous datasets. Frontiers in neuroinformatics. 2011:5. [PubMed: 21779242]

Seeley WW, Crawford RK, Zhou J, Miller BL, Greicius MD. Neurodegenerative diseases target large-scale human brain networks. Neuron. 2009; 62:42–52. [PubMed: 19376066]

Sonnenburg, Sr; Zien, A.; Philips, P.; Rätsch, G. POIMs: positional oligomer importance matrices‚ÄîunderstandingÄî support vector machine-based signal detectors. Bioinformatics. 2008; 24:i6–i14. [PubMed: 18586746]

Soriano-Mas C, Pujol J, Alonso P, Cardoner N, Menchon JM, Harrison BJ, Deus J, Vallejo J, Gaser C. Identifying patients with obsessive-compulsive disorder using whole-brain anatomy. Neuroimage. 2007; 35:1028–1037. [PubMed: 17321758]

Stonnington CM, Chu C, Klöppel S, Jack CR Jr, Ashburner J, Frackowiak RS. Predicting clinical scores from magnetic resonance scans in Alzheimer's disease. Neuroimage. 2010; 51:1405–1413. [PubMed: 20347044]

Strobl C, Boulesteix AL, Kneib T, Augustin T, Zeileis A. Conditional variable importance for random forests. BMC bioinformatics. 2008; 9:307. [PubMed: 18620558]

Teipel SJ, Born C, Ewers M, Bokde AL, Reiser MF, Möller H Jr, Hampel H. Multivariate deformation-based analysis of brain atrophy to predict Alzheimer's disease in mild cognitive impairment. Neuroimage. 2007; 38:13–24. [PubMed: 17827035]

Tipping ME. Sparse Bayesian learning and the relevance vector machine. The Journal of Machine Learning Research. 2001; 1:211–244.

Vemuri P, Whitwell JL, Kantarci K, Josephs KA, Parisi JE, Shiung MS, Knopman DS, Boeve BF, Petersen RC, Dickson DW. Antemortem MRI based STructural Abnormality iNDex (STAND)-scores correlate with postmortem Braak neurofibrillary tangle stage. Neuroimage. 2008; 42:559–567. [PubMed: 18572417]

Wang, H.; Nie, F.; Huang, H.; Risacher, S.; Ding, C.; Saykin, AJ.; Shen, L. Sparse multi-task regression and feature selection to identify brain imaging predictors for memory performance. Computer Vision (ICCV), 2011 IEEE International Conference on. IEEE; 2011. p. 557-562.

Wang X, Yang J, Jensen R, Liu X. Rough set feature selection and rule induction for prediction of malignancy degree in brain glioma. Computer methods and programs in biomedicine. 2006; 83:147–156. [PubMed: 16893588]

Wang Y, Fan Y, Bhatt P, Davatzikos C. High-dimensional pattern regression using machine learning: From medical images to continuous clinical variables. Neuroimage. 2010; 50:1519–1535. [PubMed: 20056158]

Westman E, Simmons A, Zhang Y, Muehlboeck J, Tunnard C, Liu Y, Collins L, Evans A, Mecocci P, Vellas B. Multivariate analysis of MRI data for Alzheimer's disease, mild cognitive impairment and healthy controls. Neuroimage. 2011; 54:1178–1187. [PubMed: 20800095]

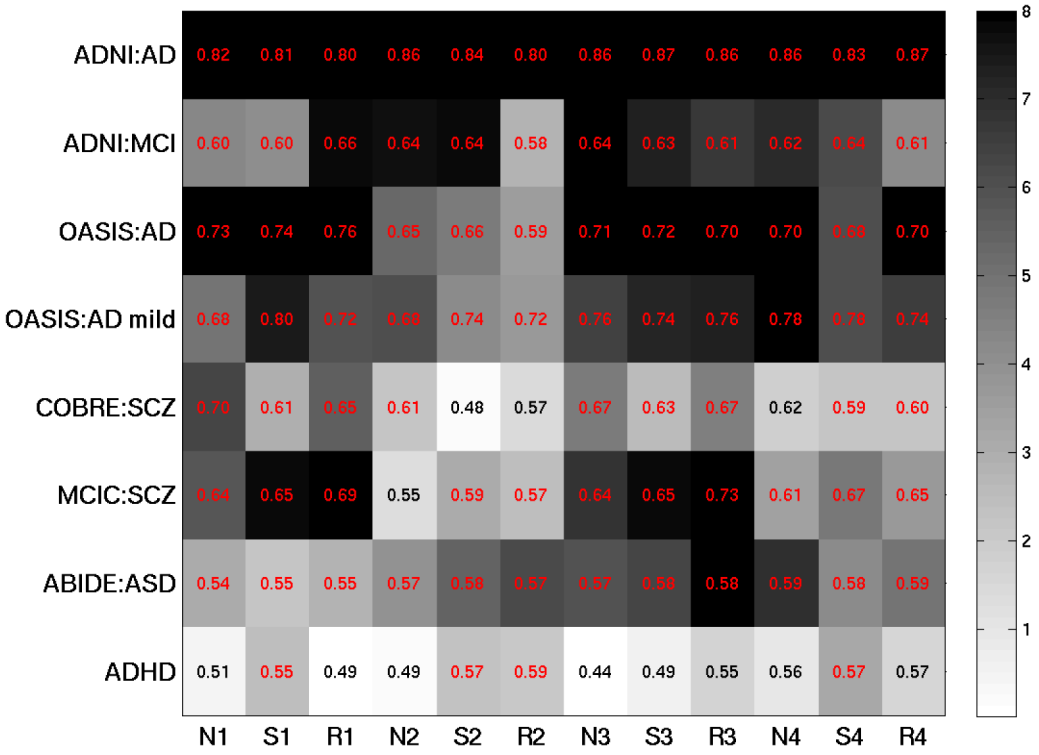
Wilson SM, Ogar JM, Laluz V, Growdon M, Jang J, Glenn S, Miller BL, Weiner MW, Gorno-Tempini ML. Automated MRI-based classification of primary progressive aphasia variants. Neuroimage. 2009; 47:1558–1567. [PubMed: 19501654]

Wolfe DA, Hollander M. Nonparametric statistical methods. Nonparametric statistical methods. 1973

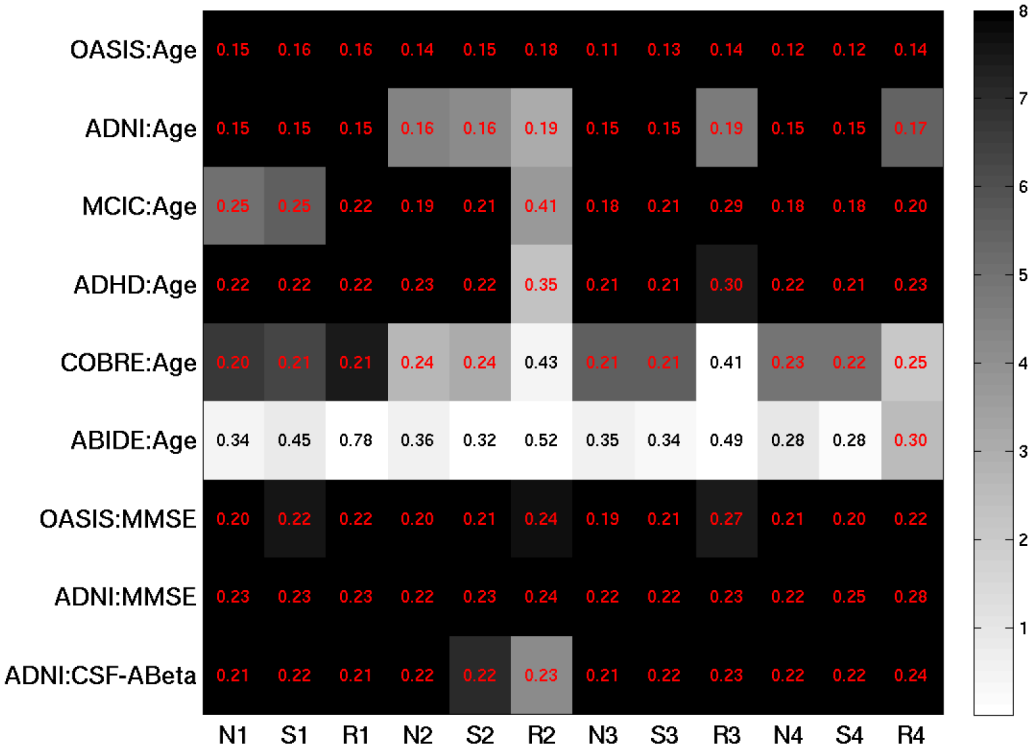
Zien, A.; Krämer, N.; Sonnenburg, Sr; Rätsch, G. Machine Learning and Knowledge Discovery in Databases. Springer; 2009. The feature importance ranking measure; p. 694-709.

Highlights

- A large-scale empirical study (total N>2800) of clinical structural MRI data
- We evaluated the performance of image-based prediction methods
- We considered an array of clinically relevant variables, both binary and continuous
- All tools and data are publicly available to replicate and benchmark results
- We provide a discussion of prediction performance and suggest future research



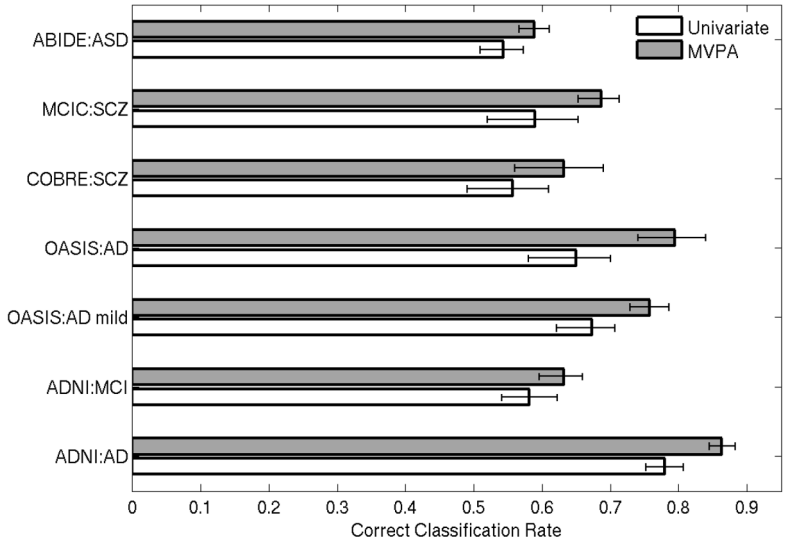
(a)



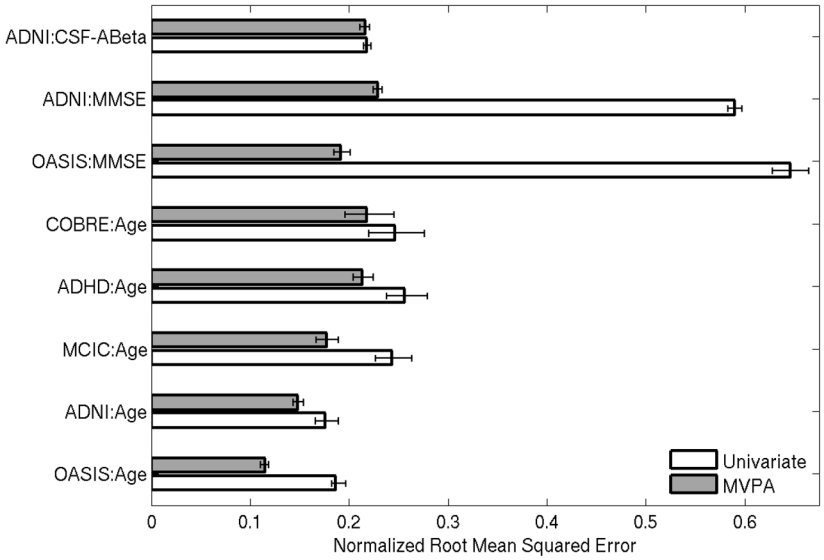
(b)

Figure 1.

Correct classification ratio (CCR) (Panel a) and normalized root mean square (NRMSE) (Panel b) for each variable and MVPA algorithm, estimated via 5-fold cross-validation. The MVPA algorithms are abbreviated as follows: N for neighborhood approximation forest, S for SVMs, and R for RVMs. The number after each letter denotes the feature type (1:*aseg*, 2:*aparc*, 3:*aseg+aparc*, 4:*thick*). The shaded gray color indicates statistical significance (−log10 p-value), where the p-value is computed via DeLong's method(DeLong *et al.*, 1988) for classification (***Panel a***), and Pearson's linear correlation coefficient for regression (***Panel b***). Statistically significant associations with a p-value greater less than 0.01 are shown in red. The RMSE is normalized by dividing by the range of the variable, enabling a comparison between variables with different units.



(a)



(b)

Figure 2.

Average prediction accuracy estimated via repeated 5-fold cross-validation for MVPA and univariate models. The MVPA models were chosen as the ones that yielded the predictions that were most significantly associated with the ground truth variables on the first 5-fold cross-validation (see Fig. 1). ***Panel a***: Binary Classification, ***Panel b:*** Regression. Error bars show the 95% confidence intervals. MVPA models yield better prediction accuracy than univariate models in all variables.

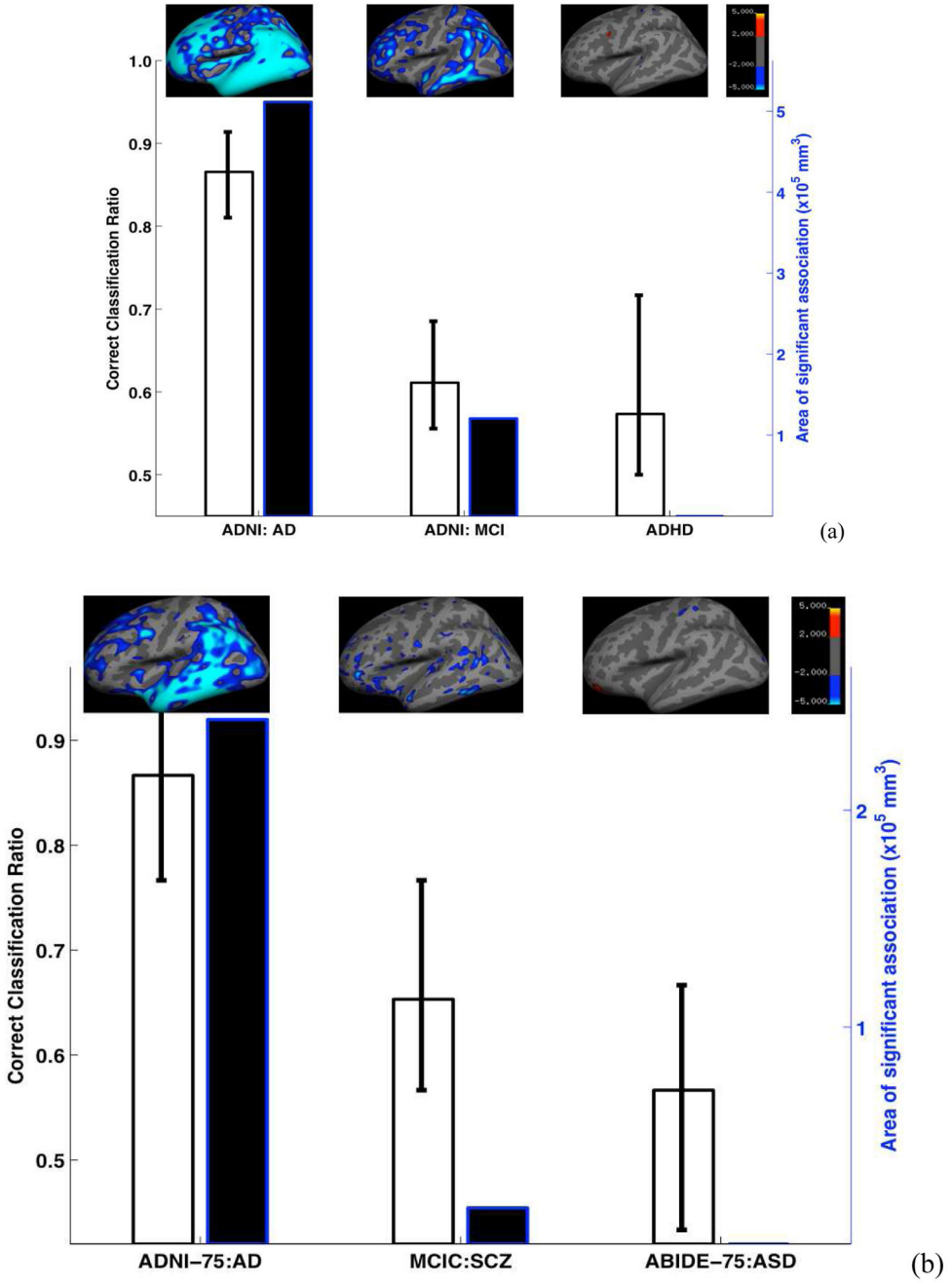
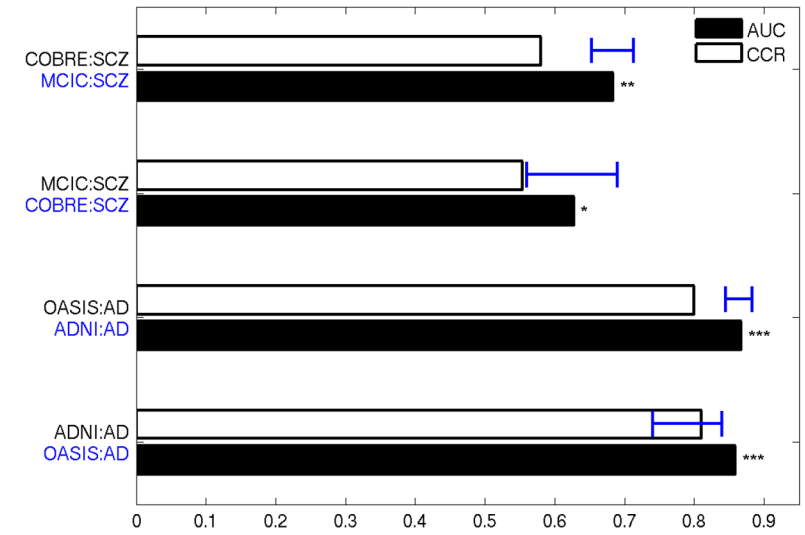


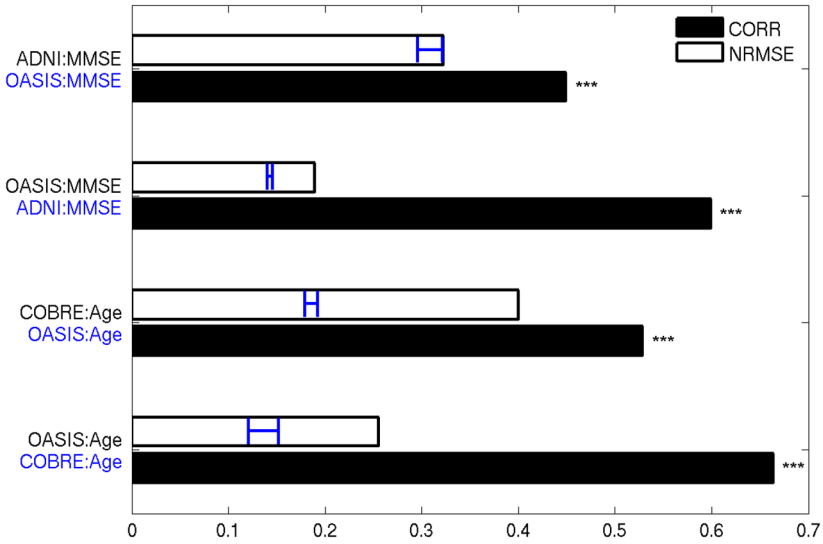
Figure 3.

MVPA prediction accuracy versus biological footprint. There is a strong agreement between the correct classification rate (CCR) and size of cortical area where thickness measurements are statistically significantly associated with the target variable (at False Discovery Rate(Benjamini and Hochberg, 1995), FDR, q=0.05). The errorbars show the full range of CCR values in each fold of cross-validation. Within each panel, the samples contained comparable number of subjects and the scans were of commensurate quality. The Relevance Voxel Machine, a variant of RVM, was applied to cortical thickness maps for the multivariate analysis. The mass-univariate analysis was conducted on the thickness maps

normalized and re-sampled to the standard *fsaverage* template, the left hemisphere of which is visualized with the statistical significance (−log10 p value) of the associations overlaid in color (uncorrected p-value < 0.01). We underscore that these maps are different from the features the MVPA models rely on for making the corresponding predictions.



(a)



(b)

Figure 4.

Eight MVPA models (in blue font) were applied to independent validation datasets (in black font) to assess prediction accuracy. ***Panel a:*** Correct classification ratio, CCR, is shown for each variable (in white). Blue bars show the 95% confidence interval of CCR estimated via cross-validation on original dataset of model. Area under the receiver-operatic charactertic curve (AUC, shown in black) was used to assess statistical significance via DeLong's method (DeLong *et al.*, 1988). ** p-value < 0.001, *** p-value < 0.0001. ***Panel b:*** Normalized root mean squared error (NRMSE) is shown for each variable (in white). Blue bars show the 95% confidence interval of NRMSE, estimated via cross-validation on

original dataset of model. Pearson's correlation (CORR, shown in black) was used to assess statistical significance. *** p-value < 0.0001

Table 1

A summary of the 6 publicly available clinical neuroimaging initiative datasets used in this study.

Dataset	N	Mean Age	Std Age	Min Age	Max Age	% Female	Number of sites
ADNI	810	75.2	6.9	55.0	91.0	42.0	58
OASIS	415	51.4	25.3	18.0	96.0	61.4	1
COBRE	129	35.7	11.9	18.0	65.0	24.8	1
MCIC	194	33.1	11.5	18.0	60.0	28.4	3
ABIDE	935	18.4	8.0	10.0	64.0	14.0	20
ADHD	392	12.9	2.4	8.4	20.9	72.7	6

Table 2

Table 2a						
Dataset	**Variable**	**N per group**	**Age (Mean±Std)**		**Female %**	**Num. of sites**
			Cases	**Controls**		
ADNI	AD	145	76.6±5.7	76.6±5.8	47.6	49
ADNI	MCI	135	75.6±6.7	75.6±6.7	36.3	47
OASIS	AD	25	77.5±6.8	77.5±6.6	72	1
OASIS	AD mild	70	75.9±7.3	76±7.2	68.6	1
COBRE	SCZ	50	34.3±10.6	34.1±10.7	18	1
MCIC	SCZ	75	33.3±11.6	33.4±11.4	26.7	3
ABIDE	ASD	325	17.8±7.4	17.9±7.4	11.4	17
ADHD	ADHD	150	13.2±2.4	13.2±2.3	78.7	6

Table 2b						
Dataset	**Variable**	**Total N**	**Mean±Std**	**Range**	**Female %**	**Num. of sites**
OASIS	age (yrs)	315	43.9±23.8	[18, 94]	62.1	1
ADNI	age (yrs)	213	75.9±5.0	[60, 90]	47.9	55
MCIC	age (yrs)	90	32.3±11.7	[18, 60]	67.8	3
ADHD	age (yrs)	115	11.7±1.7	[8.4, 14.9]	60.9	1
COBRE	age (yrs)	73	35.7±11.6	[18, 65]	31.5	1
ABIDE	age (yrs)	34	23.4±4.2	[17.3, 31.8]	23.5	1
OASIS	MMSE	235	27.1±3.7	[14, 30]	66.4	1
ADNI	MMSE	810	26.8±2.6	[20, 30]	42	58
ADNI	CSF-Aβ (ng/L)	415	169.9±55.7	[50.7, 298.8]	39.8	56

Discrete variables used in the binary classification experiments. AD: Alzheimer's disease, MCI: mild cognitive impairment, SCZ: schizophrenia, ASD: autism spectrum disorder, ADHD: attention-deficit hyperactivity disorder. See Methods for detailed definitions. For each phenotype, we constructed age, sex, site-matched case and control groups. The number of subjects per group, average age, female ratio, and number of unique sites are listed.

Continuous variables used in the regression experiments. Age, MMSE: mini-mental state exam score, and cerebro-spinal fluid (CSF)-derived amyloid burden measured as the concentration of $Ab_{1\text{-}42}$. Number of subjects, variable statistics (mean, standard deviation, minimum/maximum value), female ratio and number of sites are listed.