Brain morphometry in recent decades has increased our understanding of the neural bases of psychiatric disorders by localizing anatomical disturbances to specific nuclei and subnuclei of the brain. At least some of these disturbances precede the overt expression of clinical symptoms and possibly are endophenotypes that could be used to diagnose an individual accurately as having a specific psychiatric disorder. More accurate diagnoses could significantly reduce the emotional and financial burden of disease by aiding clinicians in implementing appropriate treatments earlier and in tailoring treatment to the individual needs. Several methods, especially those based on machine learning, have been proposed that use anatomical brain measures and gold-standard diagnoses of participants to learn decision rules that classify a person automatically as having one disorder rather than another. We review the general principles and procedures for machine learning, particularly as applied to diagnostic classification, and then review the procedures that have thus far attempted to diagnose psychiatric illnesses automatically using anatomical measures of the brain. We discuss the strengths and limitations of extant procedures and note that the sensitivity and specificity of these procedures in their most successful implementations have approximated 90%. Although these methods have not yet been applied within clinical settings, they provide strong evidence that individual patients can be diagnosed accurately using the spatial pattern of disturbances across the brain. **Keywords:** Machine learning, brain morphometry, automated diagnosis, cortical thickness, psychiatric disorders, cross validation, support vector machines.

**Introduction**

Morphometry is the study of forms or shapes. It quantifies anatomical features of interest and then correlates those features with other characteristics of the individual, including clinical, behavioral, and genetic measures (Bookstein, 1997). Many Magnetic Resonance Imaging (MRI) studies have applied the principles of morphometry to understand the association of morphological features with various psychiatric disorders (Figure 1).

MRI generates images of the brain with high signal to noise ratios (SNRs) and exquisite contrast between soft tissues to yield definable boundaries for brain regions that have long been hoped would be useful in discriminating patients who have one clinical diagnosis from patients who have a differing diagnosis. Each brain region, however, consists of several subnuclei that have specific information processing functions and may contribute differentially to the disease process of interest. Soft tissue contrast and SNR, moreover, are generally insufficient to distinguish visual boundaries of the various subnuclei of most brain regions, and therefore initial morphometric studies contented themselves with measuring overall volumes of brain regions. Although these rather simplistic approaches of studying psychiatric illness using overall conventional volumes of brain regions improved our understanding of the anatomical disturbances present in persons who have specific psychiatric disorders, opposing volumetric changes in differing subnuclei of a single brain region likely substantially reduced the ability of the overall volume of that brain region to discriminate diagnostic groups. Recent advances in methods for analyzing MR images, however, have provided substantial information about constituent subnuclei and their roles in the disease process of interest (Bansal, Staib, Wang, & Peterson, 2005; Bansal, Staib, Xu, Zhu, & Peterson, 2007; Csernansky et al., 2004; Davatzikos et al., 2005; Peterson et al., 2009; Plessen et al., 2006; Thompson & Toga, 1996a,b). These advanced imaging techniques permit the detection, comparison, and correlation of anatomical disturbances within subnuclei, thereby facilitating more accurate diagnoses of psychiatric disorders using MR images.

The accurate diagnosis of patients using clinical interviews and family history can be difficult, especially early in the course of illness when the full complex of symptoms have yet to become manifest. Overlapping diagnostic criteria, the presence of comorbid disorders, and the variable expression of psychiatric disorders further complicates accurate diagnosis. Accumulated evidence strongly suggests that psychiatric disorders have unique spatial patterns of anatomical disturbances, and that these
disturbances, together with connectivity measures, metabolite concentrations, and functional disturbances in the brain are detectable even before the overt expression of symptoms. By quantifying brain disturbances in affected patients or in persons at increased risk for developing a psychiatric illness and comparing those measures with persons without an illness, those patients might be accurately diagnosed as having a disorder or not, or as having one disorder rather than another.

Although various computer algorithms have been developed and applied to brain measures to attempt to classify the brains of persons who have psychiatric illnesses, the most common type of algorithm developed for this purpose has been supervised machine learning. A supervised method for machine-based learning and pattern classification learns decision rules and classification boundaries using a labeled set of data. Those data are so-called ‘feature vectors’, features of the image captured in the form of multi-dimensional variables that encode the location and magnitude of the brain measure of interest. Supervised machine learning then associates or links each of these vector variables with a previously known set of numerically coded diagnostic labels or classes. The learning algorithm develops decision rules and boundaries that classify individual brains as having a given diagnostic label based on the imaging feature vectors that are present in each brain. The algorithm that has been ‘trained’ using the previously known labels then should, when tested rigorously, apply the learned rules and boundaries to the classification of new brains that did not participate in the training of the algorithm. The feature vectors can be virtually any brain measure, but in the examples we will provide, they are most commonly measures of cortical thickness or local volumes of gray matter or white matter at each point (or ‘voxel’, a volumetric pixel in three dimensions) in the brain. The diagnostic labels are usually previously established clinical diagnoses. A decision rule or classification boundary can be represented as a function that maps the feature vector to the given diagnostic label. The performance of a classifier is typically assessed by dividing labeled data into two sets, a training set in which the classifier is trained, and a test set in which the trained algorithm is applied to classify a set of brains that were not involved in training the classifier. The correct diagnoses (labels) are known in the test dataset, thereby permitting the accuracy of classification to be discerned and quantified easily when applying the trained algorithm to the test dataset.

If previously established diagnostic labels are unavailable when developing the classifier, other methods can be employed to discover natural groupings in the feature vectors. These methods are termed unsupervised learning algorithms (or, alternatively, data mining or clustering procedures). The validity of the natural groupings that are identified by these unsupervised learning procedures must then be established by applying statistical methods, such as linear discriminant analysis, to assess the statistical significance of differences in clinical and

Figure 1 Spatial patterns of abnormalities in local volumes across cerebral cortex in participants with various neuropsychiatric disorders. Top Row: 44 individuals with velo-cardio-facial syndrome (VCFS); and Second Row: 81 Schizophrenia (SZ) patients. These abnormalities were detected by comparing each group of patients with a group of healthy individuals who were age- and sex-matched with the patient groups. Affected populations have distinct spatial patterns of abnormalities that can be used to train a classifier to diagnose individual people within that population. The p-values of abnormalities are color-encoded and displayed on the template brain. Purple: regions of significant (p-value <0.0001) local volume reductions; Red: regions of significant (p-value <0.0001) local volume increases in the affected populations. VCFS, velo-cardio-facial syndrome; SZ, schizophrenia
Using brain morphometry as a clinical tool

behavioral measures across the naturalistic groupings.

Extant methods for machine-based classification of an individual brain using imaging data (Barnes et al., 2004; Davatzikos, Fan, Wu, Shen, & Resnick, 2008; Duchesnay et al., 2007; Efthymiou et al., 1995; Fan, Shen, & Davatzikos, 2005; Herholz et al., 2002; Jack et al., 2004; Kawasaki et al., 2007; Klöppel et al., 2008a; Lao et al., 2004; Lerch et al., 2006; Liu et al., 2004; Loock, Parsey, Oquendo, & Mann, 2004; Mourao-Miranda, Bokde, Born, Hampel, & Stetter, 2005; Teipel et al., 2007; Wahlund et al., 2005) can generally be characterized as supervised. Imaging data from all participants are first spatially warped, or ‘normalized’, to a template brain to bring all corresponding points into a common spatial register (termed ‘template space’) and to compute the feature vectors of interest for each of the brains. The quantitative features are typically computed from images by applying various technical procedures to extract feature vectors from the imaging dataset, procedures that include volume preserved warping (Xu et al., 2007), mass preserved warping (Haker, Tannenbaum, & Kikinis, 2001), voxel based morphometry (VBM) (Ashburner & Friston, 2000), deformation based morphometry (Paus et al., 2001), surface morphometry (Bansal et al., 2007; Thompson & Toga, 1996b), and spherical harmonics (Chung, Dalton, & Davidson, 2008). Of these various procedures, the most commonly employed has probably been VBM, which generates point-wise (or ‘voxel-wise’) maps of measures at each voxel in the brain. It is an automated method that analyzes the nonlinear deformation field used for the spatial normalization of one particular brain into a common template space and that then generates a voxel-wise map of the local changes in brain volume that are required to normalize the brain to the template brain (i.e., to bring all points in one brain into register with the corresponding points in the template brain). These measures of local changes needed for spatial localization are fine-grained in their spatial resolution (typically on the order of a cubic millimeter), and those measures can then be compared across groups of participants.

We will first review the principles and procedures for machine learning and how they can be applied to the diagnosis of psychiatric illnesses. Each of the various components of these procedures is relatively independent of the others and can be modified to influence significantly the accuracy of classification. We will also review several previous attempts at localizing and associating disturbances in the brain with specific psychiatric disorders and the specific components of machine learning procedures that best exploit these disturbances for more accurately diagnosing illness. Finally, we will discuss the strengths and limitations of extant machine learning algorithms. These methods provide sufficient preliminary evidence that morphometric data adequately localize disturbances in a specific disorder and that machine learning procedures in the foreseeable future will likely yield useful tools for aiding the clinical diagnosis of patients.

Machine learning

Machine learning can be partitioned into components and procedures, including participants, imaging modality, brain regions, brain measures, classifiers, and statistical validations. Each of these components is described briefly in the subsections below.

Selection of participants to which the machine learning procedures are applied

The participants recruited into an imaging study are undoubtedly the most crucial determinant of how generalizable the performance of a machine learning-based classification will be in the larger population of individuals affected with the disorder that is being characterized and classified. Machine learning techniques are usually applied to imaging data that have been acquired in case-control studies (Klöppel et al., 2008b; Macmaster et al., 2008; Patel, Polikar, Davatzikos, & Clark, 2008; Fiven et al., 1995; Plant et al., 2010; Stolley & Schlesselman, 1982), which usually group-match participants on demographic characteristics such as age, sex, handedness, ethnicity, and socioeconomic status in the affected and healthy participants. In addition to matching groups on these variables, these studies also usually control for the effects of these demographic variables on imaging measures by controlling statistically for them and other potential confounders. Controlling for these effects, either by the selection of subjects who enter the study or statistically, is expected to allow the findings to generalize to the entire population of affected individuals. Imaging studies are expensive to conduct, however, and therefore the number of participants who can be included is usually limited to a few dozen participants – indeed, the largest imaging studies have had at most only a few hundred affected and healthy participants combined. The limited number of participants limits the ability to stratify across the key demographic features of the population, and if those demographic influences on brain structure and function overlap those features that are the basis for classification within the machine learning algorithms, or if they influence performance of those algorithms, then the trained machine classifications will likely be unable to diagnose accurately a new person from the larger population of affected individuals. To take an extreme example, if the classification boundary and decision rules were learned using data only from females (Savio, Garcia-Sebastián, Graña, & Villanúa, 2009) and if the features that enter the classifier include female-specific features that differ from
Feature vectors are morphological measures that the algorithm classifies. Inaccurate delineation of brain regions will ultimately undermine the accuracy of diagnostic classification using machine learning algorithms, and in many instances will prove fatal to those efforts. Delineation of a brain region by an expert in brain anatomy remains the gold-standard for definition of brain regions, but manual delineation is tedious and can be prone to drifts in measures over time and to variability across experts. Studies (Peterson, 2000; Peterson, Feineigle, Staib, & Gore, 2001a; Peterson, Thomas, Kane, & Others, 2003; Peterson et al., 1993a,b, 2000, 2001b, 2007, 2009; Plessen et al., 2006) that employ manual delineation should ensure accurate definition by extensively documenting all protocols for region definition, maintaining high reproducibility of measures across experts and across time by reassessing the reliability of the expert at predefined time intervals, and employing quality control checks on the definitions of each expert by an independent rater.

Because of the time, training, and expense required to implement gold-standard manual delineations of brain regions, most studies employ automated tools for region delineations. For example, the automated labeling pathway platform combines a series of publicly available software packages, including Analysis of Functional Neuro-images, Brain Extraction Tool (Smith, 2002), FMRI's Linear Image Registration Tool, and Insight Segmentation and Registration Toolkit (ITK) (Yoo, 2004) to isolate brain from non-brain tissue (Wu, Carmichael, Lopez-Garcia, Carter, & Aizenstein, 2006), to spatially co-register each brain to a template brain using nonlinear deformations [in the ITK (Chen, Zhu, Thulborn, & Ugurbil, 1999) algorithm], to segment brain tissue into gray and white matter, and to label brain regions. Labeling of brain regions can be aided using either the automated anatomical labeling atlas containing 90 manually traced regions (Tzourio-Mazoyer et al., 2002), the Brodmann atlas that has 82 regions (Rorden & Brett., 2000) or locally generated regions defined from fMRI studies (Carter et al., 2000). Automatically delineated regions, however, typically have large errors when compared with gold-standard manual definitions, especially near the outer boundary of the brain region being delineated, and they therefore tend to increase variance in the computed morphometric feature on which the classifier usually operates. These errors in region definition undermine the statistical power to detect brain features that discriminate groups when participant numbers are relatively small, as they usually are in imaging studies.

Methods for delineating brain regions

Because morphometry is the study of shapes and forms, and because machine learning algorithms are applied to measures of regional brain morphology, the accurate delineation of brain regions is an important prerequisite for the accurate measure of feature vectors that the algorithm classifies. Inaccurate delineation of brain regions will ultimately undermine the accuracy of diagnostic classification using machine learning algorithms, and in many instances will prove fatal to those efforts. Delineation of a brain region by an expert in brain anatomy remains the gold-standard for definition of brain regions, but manual delineation is tedious and can be prone to drifts in measures over time and to variability across experts. Studies (Peterson, 2000; Peterson, Feineigle, Staib, & Gore, 2001a; Peterson, Thomas, Kane, & Others, 2003; Peterson et al., 1993a,b, 2000, 2001b, 2007, 2009; Plessen et al., 2006) that employ manual delineation should ensure accurate definition by extensively documenting all protocols for region definition, maintaining high reproducibility of measures across experts and across time by reassessing the reliability of the expert at predefined time intervals, and employing quality control checks on the definitions of each expert by an independent rater.

Because of the time, training, and expense required to implement gold-standard manual delineations of brain regions, most studies employ automated tools for region delineations. For example, the automated labeling pathway platform combines a series of publicly available software packages, including Analysis of Functional Neuro-images, Brain Extraction Tool (Smith, 2002), FMRI's Linear Image Registration Tool, and Insight Segmentation and Registration Toolkit (ITK) (Yoo, 2004) to isolate brain from non-brain tissue (Wu, Carmichael, Lopez-Garcia, Carter, & Aizenstein, 2006), to spatially co-register each brain to a template brain using nonlinear deformations [in the ITK (Chen, Zhu, Thulborn, & Ugurbil, 1999) algorithm], to segment brain tissue into gray and white matter, and to label brain regions. Labeling of brain regions can be aided using either the automated anatomical labeling atlas containing 90 manually traced regions (Tzourio-Mazoyer et al., 2002), the Brodmann atlas that has 82 regions (Rorden & Brett., 2000) or locally generated regions defined from fMRI studies (Carter et al., 2000). Automatically delineated regions, however, typically have large errors when compared with gold-standard manual definitions, especially near the outer boundary of the brain region being delineated, and they therefore tend to increase variance in the computed morphometric feature on which the classifier usually operates. These errors in region definition undermine the statistical power to detect brain features that discriminate groups when participant numbers are relatively small, as they usually are in imaging studies.

Brain measures

Feature vectors are morphological measures that can be compared statistically across groups of participants. The features that typically have been used to study persons with psychiatric disorders include

volumes of entire brain regions, voxel-wise maps of thickness of the cortical mantle, voxel-wise maps of gray matter density (i.e., amount of gray matter per unit volume), local volumes of gray and white matter, and measures of local variation in the deformation fields that spatially normalize a brain to the template brain. Although the dimensionality of a feature vector (i.e., the number of features in the vector) in voxel-wise maps is typically on the order of hundreds of thousands, an imaging study comprises imaging data from at most several hundred participants. Because imaging data are available for only a limited number of participants, the probability distribution of features in a high-dimensional Euclidean space, called the feature space, cannot be discerned reliably. If the dimensionality of the feature space is not reduced using some sort of data reduction procedure, then data from an exorbitantly large number of participants would be required to learn accurate classification boundaries. For example, to estimate a multivariate Gaussian distribution of 10 features, data from more than 800,000 participants are required to ensure that at the mode the estimated distribution is close the true distribution (Silverman, 1986). The dimensionality of the feature space therefore must be reduced significantly to learn the decision rules and classification boundaries that are valid in the affected population. Some of the popular methods for dimensionality reduction include (a) principal component analysis, which generates components that are uncorrelated with one another (Jolliffe, 2002), and (b) independent component analyses, which generates components that are statistically independent (Hyvarinen, Karhunen, & Oja, 2001). Although statistically independent components are uncorrelated, components that are uncorrelated may not necessarily be statistically independent. In the reduced feature space, a machine learning algorithm compares two feature vectors using either their Euclidean distance, which computes the dissimilarity between feature vectors as the length of the line segment connecting them in the feature space, or the Mahalanobis distance, which accounts for the correlation among features and their variances to compute the distance between the feature vectors (Mahalanobis, 1936).

Choice of a classifier

We now briefly review the various classifiers that have been used to diagnose patients as having one of several psychiatric disorders (Mitchell, 1999). The last decade has witnessed an explosion in the use of classifiers in a wide range of disciplines that includes economics, astronomy, molecular biology, and diverse medical specialties such as cardiology and radiology. The use of classifiers in neuroimaging, however, is just beginning. A classifier is a computer algorithm that learns the most concise and accurate decision rules that best discriminate data among classes and applies the learned rules to data that it has not yet seen (Figure 2).

Decision tree classifiers. Decision trees are a popular tool for classifying data (Hunt, Marin, & Stone, 1966). They create a model tree that best predicts the
diagnosis of a patient based on a feature vector in the patient’s brain. In a decision tree, a node represents a feature, an edge between two nodes represents a combination of the features at the two nodes, and a leaf represents a diagnosis. A new patient is diagnosed by first computing its features and then using the computed features to traverse the tree from the root node to a leaf that determines the diagnosis for that individual (Figure 3).

A decision tree is learned by assigning a feature to a node that best partitions the set of brains into two or more subsets, such that the majority of the brains in each subset belong to one or the other diagnostic labels. The feature assigned to the node is removed from the feature vector, and then each subset is recursively partitioned into further subsets using the remaining features in the vector. The recursive partitioning stops when all brains in a subset have the same diagnosis.

Decision tree algorithms have several limitations, including: (a) the learning algorithm may not generalize well to the larger population of affected people, (b) the optimal tree can be learned only by searching over all possible feature assignments to the nodes, and (c) small changes in the feature values of the training set may affect partitioning of the set into subsets, thereby producing large changes in the structure of the learned tree. Therefore, noise in the feature vectors can significantly affect the performance of the classifier. Decision tree classifiers, however, become increasingly stable and more accurate with an increasing number of participants in the training set, thereby producing a more accurate diagnostic classification of new participants.

Naive Bayes’ classifiers. A naive Bayes’ classifier assumes that the features in feature vectors are statistically independent, and it uses this assumption to compute probabilities of a diagnosis from given sets of feature vectors. The estimated probabilities, in turn, are applied to diagnose an individual (Domingos & Pazzani, 1997; Hand & Yu, 2001; Rish, 2001). Bayes’ theorem describes the relationships of the marginal, conditional, and joint probabilities of features and diagnostic labels that are used to learn the classification rule. A naive Bayes’ classifier may outperform other sophisticated classifiers because the assumption that the features are independent reduces the number of parameters of the probability distributions that must be estimated. The parameters in a naive Bayes’ classifier therefore can be estimated accurately using fewer feature vectors, thereby increasing its performance as compared to that of other classifiers.

AdaBoost classifiers. AdaBoost is a meta-algorithm that constructs a diagnostic classifier with high accuracy (i.e., AdaBoost is a strong classifier) by linearly combining several classifiers that have low accuracy (i.e., separately they are weak classifiers) that of other classifiers.

Support vector machine (SVM) classifiers. A support vector machine (SVM) is a linear classifier (Cortes & Vapnik, 1995; Vapnik, 1999) – i.e., it constructs a linear classification boundary in the feature space that best separates feature vectors in the training set. The classification boundary is an optimal hyperplane (a plane in feature spaces of dimension greater than two) that maximizes the distance between the hyperplane and the feature vectors that are closest to the hyperplane. The feature vectors that define the hyperplane are called the ‘support vectors’ (Figure 4).

Support vector machines have been extended to classifying a feature vector among more than two diagnostic classes and for computing nonlinear
separating planes by first mapping the features to a higher dimensional space using nonlinear functions and then estimating the optimal hyperplane that best separates the features in the higher dimensional space (Aizerman, Braverman, & Rozonoer, 1964). Although an SVM may perform well in classifying a new feature vector that was not used to train the SVM, the correct choice of the optimal mapping to the higher dimensional space and the formidable computational complexity remain the greatest limitations of SVM for diagnostic classification (Burges, 1998). In addition, the decision boundary learned by an SVM is defined by the support vectors only, and therefore if the data are sparse, the classification boundary is highly sensitive to the noise in these vectors, such that even small variations in support vectors can lead to a large variation in the decision boundary. Furthermore, feature vectors other than support vectors do not define the decision boundary, and therefore, an SVM classifier ignores a large amount of valuable data, thereby undermining the generalizability of the learned boundary to the larger population of affected individuals.

**Validation procedures**

The performance of a classifier in diagnosing a new patient is evaluated by computing (1) misclassification rates using cross-validation procedures, and (2) prediction error using bootstrap procedures and Akaike’s Information Criteria (AIC). Prediction error measures the accuracy of a classifier in predicting the diagnostic class for an individual from its feature vector. Prediction error is computed as the expected value of the squared difference in the feature vector of the person and its predicted value by the classifier (Efron & Tibshirani, 1993). In classification problems, the prediction error computes the probability of an incorrect classification, which is the probability that the predicted diagnosis differs from the actual diagnosis of a patient. A model with a lower prediction error will generalize better to the population than the model with a higher prediction error. The AIC computes a prediction error by adjusting the error using the complexity c (i.e., the number of diagnostic classes) of the learned model and an estimate of the residual variance of feature vectors in each class.

**Bootstrapping.** A bootstrap procedure estimates prediction error by (a) creating B sets of features with replacement from the original set of feature vectors, (b) generating B models from each of these B sets, and then (c) using the original set to compute the prediction error for each of the B models. The overall prediction error is computed as the average of the B prediction errors, one for each of the B models (Efron & Tibshirani, 1993). However, this simple bootstrap procedure is biased, meaning that it underestimates the prediction error. An improved bootstrap procedure can be used to compute an unbiased estimate of the prediction error by first estimating the bias in the simple bootstrap procedure and then adding the estimated bias to generate the improved estimate of the prediction error (Efron & Tibshirani, 1993).

**Cross-validation.** A cross-validation procedure computes the sensitivity and specificity of a classifier by first dividing the set of feature vectors into two sets: a training set and a test set. Then it learns the classification rules using the feature vectors in the training set and evaluates the learned rules using the feature vectors in the test set. A method for cross-validation therefore divides the data into two parts: one part to generate the model, and the other to assess the model’s accuracy. An algorithm accurately diagnoses or classifies a participant if the person’s true diagnosis as identified by clinical experts matches the diagnosis that the algorithm assigns; otherwise the participant is misclassified. The misclassification rates for a classifier are computed using a three-step algorithm: (a) The set of N feature vectors is divided into n = N/k sets containing roughly equal numbers of feature vectors, where k is a user-specified number between 1 and N/2; (b) one of the n sets is selected as the test set, and the classification rules are learned using the feature vectors in the remaining n-1 sets. The feature vectors in the test set are used to compute the prediction error using the learned classification rules; (c) the second step is repeated using each of the n sets as the test set, yielding n estimates of the prediction errors, which are averaged to estimate the overall prediction error. For example, in a procedure termed ‘Leave-one-out’ (LOO) cross-validation, the validation procedure is repeated N times by treating every one of the N feature vectors as the test feature, training the classifier using the remaining features, and computing the misclassification rates by counting the number of feature vectors that were misclassified. The average misclassification rate is computed for each diagnosis by dividing the number of misclassified individuals by the total number of individuals who have that clinical diagnosis. A procedure termed ‘split-half’ cross-validation, on the other hand, partitions the set into two halves, with an equal number of feature vectors. By dividing the set into several random halves, the split-half procedure can be repeated several times to compute the average and standard deviation of misclassification rates across these several applications of split-half cross-validation procedures.

**Examples of classifiers for diagnosing psychiatric illnesses using MRI**

MRI has enabled us to localize brain disturbances to specific brain regions, and these disturbances have been analyzed using machine learning algorithms to diagnose an individual as having a specific disorder.
of these disorders.

Alzheimer’s disease

Alzheimer’s disease (AD) is a neurodegenerative disease characterized by progressive impairment of memory and other cognitive functions. Brain MRI measures have been demonstrated to be more sensitive indicators of AD than cognitive assessment alone (Scheltens & Korf, 2000). Several MRI studies have identified severe localized volume loss, presumably representing synaptic and neuronal degeneration, in hippocampus (Ball et al., 1985; Bobinski et al., 1996; Braak & Braak, 1991a,b; Brion, Octave, & Couck, 1994; Hyman, Van Horsen, Damasio, & Barnes, 1984; Mesulam, 2000), entorhinal cortex (Ball et al., 1985; Bobinski et al., 1996; Hyman et al., 1984; Janke et al., 2001), amygdala (Braak & Braak, 1991a,b; Brion et al., 1994; Mesulam, 2000), parahippocampal formation (Braak & Braak, 1991a,b; Brion et al., 1994; Mesulam, 2000), and tempoparietal association cortices (Decarli, 2000). Volume loss in the anterior hippocampus and amygdala occur even before the onset of clinical symptoms and in individuals who are at genetic risk of developing AD (Fox et al., 1996; Lehtovirta, Laakso, Frisoni, & Soininen, 2000). Furthermore, in the earlier stages of AD, hippocampal volume has been shown to be 1.75 standard deviations smaller than those in age-matched healthy participants (Jack et al., 2000). MRI findings are possibly a sensitive measure of dementia (Scheltens & Korf, 2000), and early atrophy of hippocampus and the ventromedial temporal lobe is consistent with the presence of early memory impairment (Braak, Braak, & Bohl, 1993; Erkinjuntti et al., 1993; Scheltens & Fox, 2002). Hippocampal atrophy, however, is not unique to AD, as it has been associated with Parkinson’s dementia (Laakso et al., 1996) and otherwise healthy individuals who have impairments in recent memory (De Leon, Convit, De Santis, & Bobinski, 1999).

A number of procedures for machine learning using anatomical MRI data have been developed to diagnose individuals as having AD or not. Two studies (Hinrichs, Singh, Xu, & Johnson, 2009; Zhang, W., Zhou, Yuan, & Shen, 2011) used data from both anatomical MRI and PET for classification. All of these procedures applied an SVM to learn the decision boundary in a feature space, and most validated performance of the classifier using an LOO cross-validation. Two studies applied 10-fold cross-validation (Hinrichs et al., 2009; Zhang et al., 2011), and another (Cuinnette et al., 2010) applied split-half validation by dividing the participants into a training set or a test set. These procedures classified individuals with high sensitivity (SN ± SD = 83% ± 14%, range = 57%–100%) and specificity (SP ± SD = 88% ± 9%, range = 66%–97%) (Table 1).

Major depressive disorder

Major depressive disorder (MDD) is defined by the presence of depressed mood and associated disturbances in sleep, appetite, energy, pleasure, self-esteem, and concentration that persist for more than 2 weeks. Using MRI, atrophy in cortex (Morris & Rapoport, 1990) (Ballmaier et al., 2004a; Pantel et al., 1997; Rabins, Pearlson, Aylward, Kumar, & Dowell, 1991), frontal lobe (Kumar, Bilker, Jin, & Udupa, 2000), orbitofrontal cortex (Ballmaier et al., 2004b; Lai, Payne, Byrum, Steffens, & K.R.R., 2000; Lee et al., 2003), gyrus rectus and anterior cingulate (Ballmaier et al., 2004b), hippocampus (Krishnan et al., 1992; Shah, Ebmeier, Glabus, & Goodwin, 1998; Sheline, Sanghavi, Mintun, & Gado, 1999; Steffens & Krishnan, 1998; Steffens et al., 2000, 2002), amygdala (Sheline et al., 1999; Steffens et al., 2000), and lesions in the basal ganglia (Rabins et al., 1991; Steffens & Krishnan, 1998; Tupler et al., 2002), especially in the caudate (Krishnan et al., 1992) and putamen (Steffens & Krishnan, 1998; Tupler et al., 2002), have been associated with MDD. MRI studies have also identified differences between early versus late-onset MDD (Ballmaier et al., 2004a,b), with late-onset illness showing less frontal and more temporal and parietal atrophy. In addition, several studies using semi-quantitative ratings (Butters et al., 2004; Greenwald et al., 1998) and semi-automated measures (Taylor et al., 2003) have reported an increased prevalence of white matter hyperintensities in periventricular and subcortical regions, especially in the elderly (Salloway et al., 1996). DTI has shown decreased anisotropy in white matter diffusion, and therefore possible loss of integrity in neural pathways across white matter, of the prefrontal cortex (Alexopoulos, Kiosses, Choi, Murphy, & Lim, 2002; Hickie et al., 1995; Taylor et al., 2004). Patients with subsyndromal depression had similar patterns of atrophy in the prefrontal cortex as those with MDD (Kumar, Jin, Bilker, Udupa, & Gottlieb, 1998). Although several studies have associated disturbances with MDD using anatomical MRI, only one fMRI study in 19 patients with MDD and 19 healthy controls has attempted to diagnose an individual as having MDD using an affect task of sad faces (Fu et al., 2007). This latter study applied SVM to the voxel-wise maps of activation across the brain and reported high sensitivity (84%) and specificity (89%) in LOO cross-validation analyses. However, because the study employed fMRI data from only 38 participants to learn the classification rules, and then computed sensitivity and specificity using LOO cross-validation, the method may not generalize well to classify an individual in clinical populations.
Autism spectrum disorders (ASDs)

Autism spectrum disorders are characterized by abnormalities in reciprocal social interaction and communication, as well as the presence of restricted interests and behaviors. These disorders usually manifest before the age of 3 years. Imaging studies have demonstrated disturbances in the cortex, corpus callosum (Piven, Bailey, Ransons, & Arndt, 1997), basal ganglia (Sears et al., 1999), and cerebellum.
(Piven et al., 1992), and they have reported correlations among gray matter volume, Intelligence Quo-
tient (IQ), and measures of symptom severity (Rojas et al., 2006). Previous attempts to diagnose an individual as having ASD have applied either SVM (Ecker et al., 2010a,b) or LPBoost (Singh, Mukherjee, & Chung, 2008) to voxel-wise maps of (a) localized volume change in GM, WM, and CSF (Ecker et al., 2010b), (b) cortical thickness and average convexity, mean curvature, local area, and distortion of the cortical surface (Ecker et al., 2010b), or (c) cortical thickness (Singh et al., 2008). These methods evaluated classification accuracy, which was greater than 85%, using methods for either LOO (Ecker et al., 2010b), leave-two-out (Ecker et al., 2010a), or ninefold (Singh et al., 2008) cross-validation. The classification rules, however, were learned using imaging data from fewer than 40 participants, whereas the dimensionality of the feature space was very large, on the order of thousands of voxels across the entire brain. In addition, the several brain regions that contributed to classifying an individual were small, typically fewer than 100 voxels, and they were randomly scattered across the entire brain. Although the small, randomly scattered regions may have classified the small number of participants with high accuracy, disturbances in the brains of individuals with ASDs are unlikely to be randomly distributed in this way in small regions across the brain. Thus, because these methods used brain regions that may not be representative of the disturbances in ASDs, and because they used a small number of participants that may not be representative of the larger population of people with ASDs, the learned classification rules cannot yet be used within clinical settings to diagnose a patient as having an ASD (Table 2).

### Discussion

Many sophisticated imaging and statistical methods have been developed to diagnose an individual automatically as having a specific neuropsychiatric disorder. These methods learned classification rules and decision boundaries by applying SVMs to selected image features, such as a voxel-wise map of gray matter intensities, gray matter volumes, white matter volumes, or volumes and shape features of various subcortical regions of the brain. Classification accuracy was evaluated using either leave-one-out (LOO) or 10-fold cross-validation, with reported sensitivities and specificities frequently being greater than 90%. Most of these machine learning algorithms were applied to diagnose an individual as having AD. However, these methods may not generalize well in the affected populations because the classification rules were learned from small groups of participants and from image features that may not be clinically relevant, and the classification accuracies were not assessed in clinical settings. Furthermore, although LOO cross-validation is a widely accepted method for evaluating the performance of classification algorithms, it may have overestimated the sensitivity and specificity of the classifiers, especially because data were available for only a small number of participants. Therefore, more rigorous validation and statistical procedures, in particular split-half validation, should be used to evaluate accurately the performance of these classifiers. Thus, although the classification methods used thus far in imaging studies have shown that brain morphometry has the potential to diagnosis individuals accurately, evaluation of the performance of the classifiers using more representative features and more rigorous validation procedures are essential to ensure accurate and valid performance estimates and to improve the generalizability of the learned classification rules within the affected population.

All attempts thus far to use imaging-based automated classifiers to diagnose psychiatric disorders have assessed performance of the classifiers against the ‘gold standard’ clinical diagnoses, though exactly how golden those standards are, is questionable. For

**Table 2.** Recent procedures for diagnosing an individual as having autism spectrum disorders (ASDs) or not using anatomical MRI data. The classification accuracy was computed using methods for leave-\(k\)-out (LOO) cross-validation.

<table>
<thead>
<tr>
<th>Study</th>
<th>Participants</th>
<th>Features</th>
<th>Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ecker et al., 2010a</td>
<td>20 with ASD and 20 healthy controls</td>
<td>Five morphological measures brain surfaces: (1) convexity and concavity, (2) mean curvature, (3) metric distortion, (4) cortical thickness, and (5) pial area</td>
<td>SN &gt;85%, SP &gt;85% of a SVM classifier computed using leave-two-out cross validation</td>
</tr>
<tr>
<td>Ecker et al., 2010b</td>
<td>22 with ASD and 22 healthy controls</td>
<td>Locally averaged voxels from GM, WM, and CSF</td>
<td>SN = 88%, SP = 86% of a SVM classifier computed using LOO cross validation</td>
</tr>
<tr>
<td>Singh et al., 2008</td>
<td>16 with ASD and 11 healthy controls</td>
<td>Cortical thickness</td>
<td>Accuracy (number of correctly classified) ~90% of a LPBoost classifier computed using ninefold cross validation</td>
</tr>
</tbody>
</table>

CSF, cerebrospinal fluid; GM, gray matter; LOO, leave-one-out; SN, sensitivity; SP, specificity; SVM, support vector machine; WM, white matter.
example, even when those clinical diagnoses are made rigorously using structured or semi-structured diagnostic interviews, those diagnoses may not be perfect. Moreover, and possibly more problematically, the unquestionable etiological and neurobiological heterogeneity of psychiatric diagnoses has for some time brought the categorical boundaries for clinical diagnoses under increasing suspicion and criticism as being invalid. The extent to which those diagnoses are invalid is the extent to which the assessment of performance of imaging-based classifiers will be undermined. Certainly the performance of a classifier can be no better than the quality of the gold standard against which it is evaluated, because even if the brain-based classifier were to diagnose patients more accurately than do symptom-based clinical diagnoses, we could never know it, given that we have nothing better or more golden than the clinical diagnosis by which to judge the brain-based classifier. Some kind of independent gold standard would be needed to evaluate and validate both the brain-based classification and clinical diagnosis, such as their ability to predict future clinical course, treatment response, behavioral or cognitive profiles, or comorbid illnesses. Nevertheless, the difficulties that the etiological and neurobiological heterogeneity pose for the validity of symptom-based clinical diagnoses does not entirely doom or preclude their utility as a benchmark for evaluating imaging-based classifiers, given that whatever degree of validity they have provides the opportunity to demonstrate the diagnostic potential of a putative classifier and can provide proof-of-concept support for the entire enterprise of imaging-based classification. Also certainly, it is impossible for classifiers consistently perform at a level that is statistically significantly better than chance if symptom-based clinical diagnoses are entirely invalid. Furthermore, the presence of neurobiological heterogeneity for any single clinical diagnoses does not mean that those neurobiological subtypes have nothing in common with one another. Indeed, the morphological signatures in the brain of those neurobiological subtypes may as a whole have more in common with one another than they have with the morphological signatures of other symptom-based clinical diagnoses. Those subtypes, in fact, may very well share certain features that constitute in essence a ‘final common pathway’ that is responsible for those features producing overlapping symptom complexes and that have caused those symptom complexes to be classified as a single clinical disorder in the first place. The fact that the best performing imaging-based classifiers consistently report diagnostic accuracies in the 80–90% range suggests that the ‘gold standard’ clinical diagnoses must be at least that good themselves in terms of having brain features that are more alike than they are like the brain features of the comparator group.

Methods for diagnostic classification typically have used SVMs to generate classification boundaries and decision rules using imaging data from participants who have gold-standard clinical diagnoses already available, and they have used those classifiers to diagnose individuals who were not included in the datasets used to generate the diagnostic algorithms. Reported classification accuracies generally have exceeded 85% and typically have been estimated using either LOO cross-validation or 10-fold cross-validation. LOO cross-validation is sensitive to noise in the imaging data, especially if the feature space consists of voxelwise maps of brain measures and if the imaging data are available for only a small number of participants. Although 10-fold cross-validation in principle should be more robust and more accurate in assessing the accuracy of diagnostic classification than LOO cross-validation, the limited number of participants in these studies (typically fewer than 100 combined across diagnostic groups) likely rendered performance of these two methods comparable. The limitations of cross-validation techniques undermine the accuracy estimates of these diagnostic methods and argue for use of more rigorous validation procedures, such as split-half cross-validation, to assess diagnostic accuracy of the classifiers.

The accuracy of diagnostic classifications using imaging data alone is limited by the features of the image that are used to generate the classification boundaries and decision rules. Although the computational and statistical complexities of classifiers have increased rapidly in the last two decades, their classification accuracy has typically not surpassed 85–90%. Further improvements in diagnostic accuracy will require more accurate localization and measurement of anatomical disturbances that are both robust and specific to particular psychiatric illnesses. Classifiers have used increasingly fine-grained features from imaging data across the entire brain in an effort to capture anatomical disturbances that are more circuit-based and distributed across multiple brain regions (Peterson, 2010). Using fine-grained, voxel-wise maps of local volume disturbances prohibitively increases noise in the imaging features and increases the dimensionality of the feature space, thereby requiring data from many more participants to generate models for diagnostic classification that are accurate and generalizable. Therefore, using features that are computed at an appropriate level of spatial resolution might be essential to help reduce the dimensionality of the feature space and make more tractable the study of a sufficient sample of participants to generate classifiers that generalize to the larger population of affected individuals.

The sampling frames used to identify patients and healthy control participants in the attempts of research studies thus far to use classifiers to diagnose patients on the basis of features in brain images...
also likely will prove to limit the generalizability and practical utility of these approaches to diagnosis. Patients in these studies have been identified from samples of convenience, usually from psychiatric clinics, and therefore they are not truly representative of all affected people in the general population (Berkson, 1946). Similarly, the healthy participants, although usually recruited to match the patient group demographically, have been screened so that they do not have any psychiatric disorder or neurological disease, and therefore also are not representative of the general community. The generalizability of a classifier therefore is limited by the demographic, clinical, and behavioral characteristics of the participants used to generate the classification boundaries. For example, excluding participants who fear the tight spaces of the MRI scanner may yield a non-representative sample of research participants in a study of generalized anxiety disorders. Similarly, excluding patients with DSM-IV Axis 1 comorbidities would create a sample that is not representative of the population affected with Bipolar Disorder (BPD). Thus, for a classifier to diagnose accurately a patient within clinical settings, and for performance estimates of the classifier generated in research settings to estimate likely performance within clinical settings, the classifier must be generated using imaging data from participants that are as representative as possible of affected people in the general population.

The greatest challenge in generating classifiers in research studies that can be applied to the affected population in clinical settings is the availability of imaging data from only a small number of participants. The limitation of sample size can perhaps be overcome by the growing trend to establish national repositories of imaging data from federally funded imaging studies. Pooling data into one repository should significantly increase the imaging data available for the training and testing of diagnostic classifiers. Large samples of pooled data also afford the possibility of being more representative of both the affected and healthy populations across the country. An example of such a public database is AD Neuroimaging Initiative (ADNI), which has acquired imaging data in 895 elderly persons who are healthy, with AD, or with mild cognitive impairment. Merging imaging data across multiple laboratories has its own challenges, however, including differing ascertainment procedures for recruiting participants, differing manufacturers and performance of the MRI scanners, differing SNR and contrast in the images, differing patterns of inhomogeneity in image intensity across the imaging volume, and differing patterns of geometric distortion in the images. In addition, differing methods for processing the images and for defining brain regions have large effects on the imaging measures that enter classifiers. Each of these confounding factors reduces the accuracy and statistical power of the diagnostic classifiers and therefore must be carefully addressed to generate valid decision rules and classification boundaries in datasets combined across sites in multi-site imaging studies.

The presence of differing anatomical subtypes of a single, phenotypically defined illness presumably will reduce the accuracy of classifiers that are based upon identifying common anatomical features that discriminate between two or more phenotypically identified disorders. A classification procedure that is trained using imaging data from one anatomical subtype perhaps cannot accurately classify a participant from a differing anatomical subtype unless those two subtypes share more imaging features with one another than they do with the conditions against which they are being compared and classified. Because anatomical subtypes of neuropsychiatric disorders have not yet been identified, increasing the number of participants seems to be the only available recourse at present to reduce the adverse effects of biological heterogeneity on the performance of a diagnostic classifier.

The performance of a classifier can be improved significantly by reducing errors when measuring the feature vectors that are used to train and test the classifier. Errors in delineating brain regions as well as errors in localizing and quantifying disturbances across multiple regions across the brain will increase measurement error in the feature vectors and thereby will undermine the accuracy of any given classifier. Automated methods for delineating brain regions, for example, or failing to account adequately for artifactual variations in image intensity across the brain, typically introduce large measurement errors near the surfaces of brain regions, particularly for small regions located deep in the brain. Yet current methods for classification typically employ automated methods for delineating brain regions because manual delineation is labor-intensive and expensive. However, manual delineation of brain regions is at present the only way to minimize errors sufficiently in the feature vectors to improve substantially the accuracy of the diagnostic algorithms.

Although current methods for making computer-based diagnoses using brain images alone cannot yet be applied within clinical settings, a growing number of studies thus far provides strong proof-of-concept evidence that these techniques can and will be used to diagnose individuals with high accuracy in the near future. Although the accuracy of diagnosis using these algorithms thus far has generally not exceeded 90%, imaging datasets from much larger and more representative populations, when subjected to advanced image processing and statistical techniques that provide precise and fine-grained techniques that provide precise and fine-grained features across the entire brain, will be able to generalize better to the larger population of affected individuals and therefore will make the automated diagnosis of psychiatric illness using...
brain images alone clinically realistic and logistically feasible.

Acknowledgements
This work was supported in part by NIMH grants MH036197, MH068318, MH16434, MH089582, 1PS0MH090966, and K02-74677, NIDA grants DA017820 and DA027100, and the Brain & Behavior Research Foundation (formerly NARSAD).

This review article was invited by the journal, for which the principal author has been offered a small honorarium payment towards personal expenses. This manuscript is based on material funded by the NIH. The authors have declared that they have no competing or potential conflicts of interest.

Correspondence to
Ravi Bansal, 1051 Riverside Drive, Unit 74, New York NY, 10032, USA; Tel: +212 543 6145; Fax: +212 543 6660; Email: bansalr@nyspi.columbia.edu

Glossary
Classification Boundaries are hypersurfaces that partition the feature spaces into regions of features such that brains within each region have the same diagnosis.

Cross-Validation Procedures compute the sensitivity and specificity of a classifier first by generating the classifier using a set of features called the training set, and then by evaluating the performance of the classifier using an independent set of features called the test set.

Key points
- Numerous recent studies using advanced imaging technologies provide strong evidence that individuals can be correctly diagnosed using appropriate brain imaging measures and automated, machine-based methods for classification.
- Nevertheless, the automated diagnosis of an individual as having a specific psychiatric disorder or not using only MRI data in routine clinical settings still faces considerable challenges.
- The generalizability of learned classification rules to affected populations is limited by the availability of imaging data from a relatively small numbers of participants and the high dimensionality of imaging measures.
- Imaging measures that encode the nature and locations of disturbances in the brain for specific disorders will be necessary for the accurate clinical diagnosis of individual patients in clinical settings.
- Rigorous, split-half cross-validations procedures should be used alongside other less stringent, but more common, validation procedures to assess performance of classifiers within real-world, clinical settings.

References


Accepted for publication: 25 January 2012
Published online: 7 March 2012