Elsevier Editorial System(tm) for Schizophrenia Research

Manuscript Draft

Manuscript Number: SCHRES-D-17-00102R1

Title: Can we accurately classify schizophrenia patients from healthy controls using magnetic resonance imaging and machine learning? A multi-method and multi-dataset study.

Article Type: Full Length Article

Keywords: structural magnetic resonance imaging, machine learning, classification, schizophrenia, voxel-based morphometry, cortical thickness

Corresponding Author: Ms. Julie L Winterburn,

Corresponding Author's Institution: Kimel Family Translational Imaging Genetics Research Laboratory, Research Imaging Centre, Centre for Addiction and Mental Health

First Author: Julie L Winterburn

Order of Authors: Julie L Winterburn; Aristotle N Voineskos; Gabriel A Devenyi; Eric Plitman; Camilo de la Fuente-Sandoval; Nikhil Bhagwat; Ariel Graff; Jo Knight; M. Mallar Chakravarty

Abstract: Machine learning is a powerful tool that has previously been used to classify schizophrenia (SZ) patients from healthy controls (HC) using magnetic resonance images. Each study, however, uses different datasets, classification algorithms, and validation techniques. Here, we perform a critical appraisal of the accuracy of machine learning methodologies used in SZ/HC classifications studies by comparing three machine learning algorithms (logistic regression [LR], support vector machines [SVMs], and linear discriminant analysis [LDA]) on three independent datasets (435 subjects total) using two tissue density estimates and cortical thickness (CT). Performance is assessed using 10-fold cross-validation, as well as a held-out validation set. Classification using CT outperformed tissue densities, but there was no clear effect of dataset. LR, SVMs, and LDA each yielded the highest accuracies for a different feature set and validation paradigm, but most accuracies were between 55-70%, well below previously reported values. The highest accuracy achieved was 73.5% using CT data and an SVM. Taken together, these results illustrate some of the obstacles to constructing effective disease classifiers, and suggest that tissue densities and CT may not be sufficiently sensitive for SZ/HC classification given current available methodologies and sample sizes.

Can we accurately classify schizophrenia patients from healthy controls using magnetic resonance imaging and machine learning? A multi-method and multi-dataset study.

Running title: Comparison of classification methods for schizophrenia in MRI

Julie L. Winterburn[a,b,c], Aristotle N. Voineskos[c,d,e,f], Gabriel A. Devenyi[a], Eric Plitman[6,7], Camilo de la Fuente-Sandoval[h,i], Nikhil Bhagwat[a,b,c], Ariel Graff[d,e,f,g], Jo Knight[e,f,j], M. Mallar Chakravarty[a,b,k,l]

[a]Computational Brain Anatomy Laboratory, Douglas Mental Health Institute, McGill University, Montreal, Quebec, Canada
[b]Institute of Biomaterials and Biomedical Engineering, University of Toronto, Toronto, Ontario, Canada
[c]Kimel Family Translational Imaging-Genetics Research Lab, Research Imaging Centre, Campbell Family Mental Health Research Institute, Centre for Addiction and Mental Health, Toronto, Ontario, Canada
[d]Geriatric Mental Health Division, Centre for Addiction and Mental Health, Toronto, Ontario, Canada
[e]Department of Psychiatry, University of Toronto, Toronto, Ontario, Canada
[f]Institute of Medical Science, University of Toronto, Toronto, Ontario, Canada
[g]Multimodal Imaging Group, Research Imaging Centre, Centre for Addiction and Mental Health, Toronto, Ontario, Canada
[h]Laboratory of Experimental Psychiatry, Instituto Nacional de Neurología y Neurocirugía, Mexico City, Mexico
[h]Neuropsychiatry Department, Instituto Nacional de Neurología y Neurocirugía, Mexico City, Mexico
[j]Biostatistics Division, Dalla Lana School of Public Health, University of Toronto, Toronto, Ontario, Canada
[k]Departments of Psychiatry and Biomedical Engineering, McGill University, Montreal, Quebec, Canada
[l]Biological and Biomedical Engineering, McGill University, Montreal, Quebec, Canada

Correspondence should be addressed to:
Julie Winterburn: winterburn.julie@gmail.com, +1 (647) 466-4802
OR Mallar Chakravarty: mallar@cobralab.ca
Computational Brain Anatomy Laboratory
Brain Imaging Centre, Douglas Mental Health University Institute
6875 Boulevard LaSalle
Montreal, Quebec, Canada
H4H 1R3

Abstract

Machine learning is a powerful tool that has previously been used to classify schizophrenia (SZ) patients from healthy controls (HC) using magnetic resonance images. Each study, however, uses different datasets, classification algorithms, and validation techniques. Here, we perform a critical appraisal of the accuracy of machine learning methodologies used in SZ/HC classifications studies by comparing three machine learning algorithms (logistic regression [LR], support vector machines [SVMs], and linear discriminant analysis [LDA]) on three independent datasets (435 subjects total) using two tissue density estimates and cortical thickness (CT). Performance is assessed using 10-fold cross-validation, as well as a held-out validation set. Classification using CT outperformed tissue densities, but there was no clear effect of dataset. LR, SVMs, and LDA each yielded the highest accuracies for a different feature set and validation paradigm, but most accuracies were between 55-70%, well below previously reported values. The highest accuracy achieved was 73.5% using CT data and an SVM. Taken together, these results illustrate some of the obstacles to constructing effective disease classifiers, and suggest that tissue densities and CT may not be sufficiently sensitive for SZ/HC classification given current available methodologies and sample sizes.

Keywords

- Structural magnetic resonance imaging
- Machine learning
- Classification
- Schizophrenia
- Voxel-based morphometry
- Cortical thickness

1 Introduction

Schizophrenia (SZ) and related psychoses are typically diagnosed using criteria defined in the 5th edition of *The Diagnostic and Statistical Manual of Mental Disorders* (American Psychiatric Association, 2013). However, the etiology of SZ is poorly understood (American Psychiatric Association, 2013; Demirci and Calhoun, 2009; Kim et al., 2015). While differences among brain regions have been consistently observed in univariate analyses at the group level (Csernansky et al., 2002; Gaser et al., 2004; Narr et al., 2005), these differences cannot be used for automated, patient-by-patient, biologically-defined diagnosis. To this end, studies have recently explored the diagnostic value of magnetic resonance (MR) imaging data (Davatzikos et al., 2005) in combination with machine learning, a field of computer science that uses pattern recognition for classification and predictive tasks (Kambeitz et al., 2015; Zarogianni et al., 2013).

These machine learning studies in SZ can be difficult to compare to one another due to differences in study populations, data processing steps, classification algorithms, and validation techniques. This heterogeneity was recently highlighted in a study that surveyed 38 publications that applied machine learning to SZ classification (Kambeitz et al., 2015). The 18 studies from this review that used structural MRI data are highlighted in Table 1 (along with sample characteristics, choice of inputs, methodology, and validation techniques).

We have performed a comprehensive comparison of machine learning techniques and methods used in SZ classification. Our evaluation included the most frequently-used

classifiers (linear discriminant analysis and linear and non-linear support vector machines), one less common one (logistic regression), and a popular publicly-available method (COMPARE (Fan et al., 2007)) across three independent datasets (435 subjects acquired at 1.5T and 3T) of both first-episode and chronic patients. We also used three MR-derived metrics including voxel based morphometry (VBM) (Ashburner and Friston, 2000), RAVENS maps (Davatzikos et al., 2001), and cortical thickness measures (Lerch and Evans, 2005) yielding a total of 42 comparisons (summarized in Table 2). Finally, we assessed the impact of two different validation methods on generalizability: 10-fold cross-validation, and a held-out dataset. To the best of our knowledge, this study is the most expansive and systematic on this topic to date.

2 Materials and Methods

2.1 Datasets Evaluated

Classification performance was evaluated on three independently-collected datasets, all of which have been published previously. For descriptions and demographics, see Table 3 and Supplementary Materials and Methods Section 1.

1. *Centre for Addiction and Mental Health (CAMH)* (Voineskos et al., 2011; Wheeler et al., 2013).
2. *Northwestern University Schizophrenia Data and Software Tool (NUSDAST)* (Wang et al., 2013).
3. *National Institute of Neurology and Neurosurgery (INNN)* (de la Fuente-Sandoval et al., 2011, 2013; Plitman et al., 2015).

2.2 Image Processing

Three neuroanatomical metrics were selected based on their prevalence in previous studies (Table 1). Further information can be found in Supplementary Materials and Methods Section 2.

1) Modulated GM VBM (Ashburner and Friston, 2000; Karageorgiou et al., 2011; Nieuwenhuis et al., 2012; Schnack et al., 2013). Briefly, this is a voxel-wise estimate of the local density of GM in a given voxel region.

2) GM RAVENS maps (Davatzikos et al., 2001; Fan et al., 2007; Zanetti et al., 2013). A RAVENS map is a metric similar to modulated VBM, but is derived using a different method for computing local GM density. RAVENS maps are calculated using an open-source software package (Davatzikos et al., 2001).

3) Cortical thickness (Lerch and Evans, 2005; Takayanagi et al., 2011; Yoon et al., 2007). This metric measures the thickness of the GM cortical mantle at ~80,000 vertices across the brain.

2.3 Machine Learning Analyses

All analyses were performed using R (www.R-project.org) (R Core Team, 2013). Following the work of previous studies (Borgwardt et al., 2010; Karageorgiou et al., 2011; Kasparek et al., 2011; Santos et al., 2010), a Principal Component Analysis (PCA) was applied to reduce dimensionality; only those PCs explaining >1% of the input feature

variance were retained (Liu et al., 2012) **Table S1 contains a breakdown of the PCs. Classification methods and parameter optimization are described in greater detail in Supplementary Materials and Methods Section 3.** The following algorithms were explored:

- **Logistic Regression (LR)** uses continuous independent variables to describe a single dependent categorical variable. To avoid over-fitting, elastic net regularization (a combination of LASSO and ridge regression) were added to the models.
- **Support Vector Machines (SVMs)** distinguish two distinct classes by constructing a hyperplane between the classes using the most ambiguous datapoints. The performances of both linear and radial basis function (RBF; ie: non-linear) kernels were assessed.
- **Linear Discriminant Analysis (LDA)** performs classification based on continuous variables, and maps input features to a lower dimensional space using a linear transformation. In this space, between-class variance is maximized, and within-class variance is minimized.
- **COMPARE** is a software tool that combines feature reduction techniques with an SVM-based classifier (RBF kernel) to perform classifications based on tissue densities (Davatzikos et al., 2005; Fan et al., 2007; Zanetti et al., 2013).

Algorithm Validation

Two validation methods were compared: 10-fold cross-validation and classification on a previously "unseen" and held-out subset of each sample (Figure 1). **The data were first split randomly (SZ:HC ratio retained) into two subsets (2:1 ratio), which we will refer to henceforth as the 'training' and 'validation' sets. PCA-based dimensionality reduction was applied to the full training set, and then the validation dataset was projected onto this PC space to ensure consistency in the features between datasets. To assist with parameter tuning and to allow for comparison with multiple validation methods, we performed 10-fold cross-validation in the training set and report the performance. We then applied our trained models to the validation set (the 1/3rd of the data that had been held out) to get an estimation of performance on an unseen dataset.** Algorithm accuracy (percentage of correctly-classified subjects) averaged across folds is reported for the training set, while accuracy, sensitivity, and specificity are reported for the validation set. **Given that our data had only one of two possible classes (HC or SZ), the significance of each accuracy result was assessed using a test statistic for the binomial distribution.**

3 Results

All training set results (using 10-fold cross-validation) are summarized in Table 4. Results from the held-out subset are summarized in Table 5. Further details are provided in the Supplementary Materials and Methods.

3.1 Modulated VBM demonstrates poor accuracy

Results using modulated VBM demonstrate poor accuracy through 10-fold cross-validation (<65% throughout) (Table 4). Non-linear SVMs offer the highest performance (mean accuracy of 63.2% across the three datasets versus 59.8% mean accuracy for LR, 55.1% for linear SVM, and 57.7% for LDA). The pairing of non-linear SVMs and modulated VBM consistently performed the best within each dataset (accuracies 61.7-64.2%; best performance in NUSDAST dataset). In all cases, performance is better than chance based on a binomial probability statistic, with the exception of the linear SVM applied to the INNN and NUSDAST datasets.

The validation results (Table 5) are on par with the training results for most algorithms. In the cases of LR and linear SVM, validation set accuracy exceeds 10-fold training accuracy in almost all cases, with the biggest gain in the NUSDAST data set (e.g.: LR training: 57.4%; LR validation: 65.2%). However, when using a non-linear SVM, the accuracy observed for the validation set is notably lower than the training set for both NUSDAST (training: 64.2%; validation: 56.5%) and INNN (training: 63.8%; validation:

59.4%). Similarly, NUSDAST accuracy improves with the use of LDA in the validation set over the training set (training: 56.4%; validation: 63.0%). Nonetheless, based on the performance in the validation phase, no single method outperformed the others. There is no fixed trend for sensitivity and specificity performance; however, a number of the models selectively show very poor sensitivity and specificity (<50% for most trials in the CAMH dataset). Conversely, other models demonstrate exceedingly high sensitivity at the expense of very low specificity (NUSDAST dataset). Only those methods with >60% accuracy in the CAMH and NUSDAST datasets performed better than chance for validation.

3.2 Performance with RAVENS is not better than modulated VBM

As with modulated VBM, the use of RAVENS maps demonstrates poor accuracy across all datasets and algorithm types. Within the training data (Table 4), LR performed the best for the NUSDAST and INNN datasets (66.0% and 70.0%; respectively), while non-linear SVM performs the best for the CAMH dataset (63.3%). Across methods, LR was also the best performer (mean accuracy of 65.6% versus 60.1% for linear SVM, 62.0% for non-linear SVM, and 57.6% for LDA). Regardless, there was no discernable impact of datasets on performance. Almost all of the training algorithms performed better than chance, except for linear SVM and LDA in the CAMH and INNN datasets, respectively.

Within the validation dataset (Table 5), the RAVENS accuracies do not seriously over- or under-perform compared to the training data. Overall, the accuracies were slightly higher than the modulated VBM results (61.4% versus 59.9%). LDA in the CAMH dataset (69.5%) and linear SVM in the NUSDAST dataset (69.6%) outperformed all other

methods. In general, LDA outperformed all other methods with 65.7% accuracy averaged across the three datasets (compared to 61.7% for LR, 58.5% for linear SVM, and 59.8% for non-linear SVM). None of the datasets notably outperformed the others. Only seven of the twelve methods performed better than chance.

Sensitivity and specificity results were very similar to the observations made using modulated VBM. Some models show very poor sensitivity (<50% for most trials in the CAMH dataset as with modulated VBM) and specificity, while other models demonstrated exceedingly high sensitivity at the expense of very low specificity (NUSDAST dataset, as with modulated VBM).

3.3 Cortical thickness offers improved accuracy, specificity, and sensitivity

Compared to both modulated VBM and RAVENS, the use of cortical thickness as an input improves many of the classification results within the training set (Table 4). In this particular case we see that many of the accuracies trend towards or exceed 70% (nonlinear SVM with CAMH: 68.0%; LR, linear SVM, and nonlinear SVM with NUSDAST: 68.8%, 71.9%, and 68.8% respectively). However, the group with the smallest sample size, INNN, did not achieve these accuracies. Non-linear SVMs outperformed all other methods in the training data (67.0% mean accuracy versus 59.1% for LR, 64.2% for linear SVM, and 59.6% for LDA). In this particular case there is a clear effect of the dataset used, where NUSDAST dataset yields observably better results overall, with a mean accuracy of 68.8% across the four algorithms (compared with 58.6% for CAMH and 60.0% for INNN). All training sets performed better than chance except for LR in CAMH and INNN datasets.

In the validation phase, many improvements in classification accuracy are observed relative to the two tissue density metrics (Table 5). Similar to the training phase, many of the methods trend towards or exceed 70% classification accuracy. The best performance was observed using nonlinear SVM in the INNN dataset, with an accuracy of 73.5% (CAMH with LDA: 68.3%; NUSDAST with LR, linear SVM, and nonlinear SVM: 70.8, 68.1, and 70.8%, respectively; LR for INNN: 69.7%). LR outperformed all other methods in the validation set, with 66.9% mean accuracy across the three datasets (compared with 62.9% for linear SVM, 63.7% for non-linear SVM, and 63.9% for LDA). The NUSDAST dataset performed better overall across the four algorithms, with 68.1% mean accuracy (compared with 63.5% for CAMH and 61.5% for INNN). Similar to the modulated VBM data, a number of the models show very poor (<50%) sensitivity, mostly with the CAMH dataset. However, this trend is not nearly as prevalent as observed with modulated VBM and RAVENS maps. All models performed better than chance except linear SVM and LDA using the INNN dataset.

3.4 Using COMPARE improves classification accuracy using modulated VBM

The input, training, and validation datasets used with the COMPARE algorithm were identical to those used for the modulated VBM analyses. On the training sets, COMPARE performed with an overall accuracy of 63.3%, 71.3%, and 71.2% on the CAMH, NUSDAST, and INNN datasets, respectively (Table 4). These accuracies were higher than any of the other methods using modulated VBM in training. On the validation set, the performance was 55.9%, 61.0%, and 67.7% for the CAMH, NUSDAST, and INNN datasets, respectively (Table 5). These results are similar to the non-COMPARE

validation results. However, the accuracy for INNN is better than what was observed in any of the validation results for modulated VBM (maximum accuracy 65.2% with LR). All results were statistically greater than chance except for the validation subset of the CAMH data.

The subjects selected for inclusion, training, and testing in the RAVENS COMPARE analysis were the same as those used in the modulated VBM analysis. With RAVENS inputs, COMPARE achieved surprisingly low overall accuracies of 55.8%, 50.0%, and 50.0% in the training set, and 51.2%, 51.0%, and 50.6% in the validation set for the CAMH, NUSDAST, and INNN datasets, respectively. Only the CAMH training data yielded an accuracy statistically greater than chance.

4 Discussion

The effectiveness of machine learning algorithms commonly used in the literature for classifying patients with SZ from HC were compared on three independent datasets using cortical thickness and two estimates of tissue density, and validated using three different techniques. The performance of all algorithms on all datasets was poor relative to previously reported results; however algorithms constructed using cortical thickness generally outperformed the others. Non-linear SVMs marginally outperformed the other methods using 10-fold cross-validation with modulated VBM and cortical thickness. LR was slightly superior to the other methods using a left-out validation dataset with both modulated VBM and cortical thickness, with a maximum accuracy of 70.8% for cortical thickness in the NUSDAST dataset. The best algorithms using RAVENS maps were LR in the training set and LDA in the validation set. The best result overall was 73.5% in the INNN validation subset using a non-linear SVM and cortical thickness.

No single dataset consistently over- or under-performed relative to the other two, although the NUSDAST dataset performed the best of the three datasets when using cortical thickness data, which turned out to be the most effective discriminatory metric used in this study. This result is not trivial to explain, as the datasets differ in multiple ways, including: primary diagnosis of the patient group, illness duration and severity, sample size, and the MR field strength used to acquire the images (Table 3). Additionally, the NUSDAST dataset is neither the smallest nor the largest of the datasets used in this study, and its patient population is not well-characterized

in its original publication (only an general diagnosis of schizophrenia is provided). It could be that this was the most homogeneous of the three datasets (in terms of illness duration, illness severity, medication status), and thus classification into two distinct groups was the most straight-forward.

In general, the diagnoses present across the three datasets are heterogeneous. The CAMH dataset contains subjects with both schizophrenia and schizoaffective disorder; as mentioned, the subjects in the NUSDAST dataset are not characterized diagnostically beyond schizophrenia/control; and the INNN subjects all have non-affective psychosis (either brief psychotic disorder, schizophreniform disorder, or schizophrenia). To explore if the heterogeneity of the samples was causing our low accuracies, we reanalyzed the CAMH VBM dataset with the schizoaffective subjects (n=26) excluded. Accuracies in this non-affective subset were even lower (<54% for all algorithms). Therefore it is unclear if the clinical heterogeneity is affecting the accuracies. Given the enforced homogeneity of the INNN dataset (early in disease course, unmedicated), as well as the high symptom severity (relative to the CAMH dataset, as measured using the Positive and Negative Syndrome Scale (Kay et al., 1987)), we expected the clearest differentiation between patient and control groups compared with the CAMH and NUSDAST datasets. This, however, was not what we observed. It may be that this population was composed of patients so early in their disease course that they were not yet correctly diagnosed. More likely, the well-documented effects of medication on neuroanatomy actually aid classification, as they further differentiate patient and control groups beyond the subtle natural

differences in their neuroanatomy (Ho et al., 2011). The effect of medication on classification performance merits significant further study.

We acknowledge that many of our results are difficult to explain and may seem a bit atypical. We believe our results illustrate that, contrary to existing results in the literature, reliable and versatile SZ/HC classifiers are difficult to construct, and existing classifiers in the literature may be over-estimating performance.

To the best of our ability we attempted to replicate the methodologies used in previous studies (Table 1). The accuracies we observed are considerably lower than what others report. Specifically, studies using SVMs with an RBF kernel trained on tissue densities report accuracies between 81.1% and 91.8%, compared with 55.4% - 66.1% achieved here (Davatzikos et al., 2005; Fan et al., 2007; Zanetti et al., 2013). Likewise, accuracies of 71.4% have been reported using linear SVMs (Nieuwenhuis et al., 2012), compared with 51.9% - 65.6% in this study; and 86.1% using LR (Sun et al., 2009), compared with 56.3% - 67.8%. The cortical thickness results were also lower than what has been previously reported (Yoon et al., 2007). Many of these studies, however, have small sample sizes (<36 subjects/group) (Fan et al., 2007; Karageorgiou et al., 2011; Sun et al., 2009), which may limit their results. Additionally, many of the studies use LOOCV, and do not validate the model on a held-out dataset. The results from the COMPARE analysis were not as high as those initially recorded by Fan and colleagues (Fan et al., 2007). A similar result was reported by Zanetti and colleagues (Zanetti et al., 2013), and they postulated that it was likely due to the forced homogeneity of the dataset used in the Fan study, which does not reflect the 'real-word' heterogeneity of a patient population. It has been proposed that an SVM classifier with an RBF kernel is not applicable for very high-

dimensional feature sets, and under certain conditions can lead to severe underfitting (subjects are all assigned to the majority class) or overfitting (Keerthi and Lin, 2006). Underfitting may have been an issue in our datasets, as the sensitivity and specificity results suggest that classifications were sometimes driven by overclassifying a specific group (ie: the most stable results may have come from classifying a disproportionate number of individuals as SZ or HC). It is possible that this effect could be corrected using a weighting term, and this should be explored in future studies. **Classification studies have also been performed using other imaging modalities. A recent study reviewed methods for multivariate analysis in functional MRI (Pereira et al., 2009). Much like our study, they were not able to conclude that one particular combination of data processing and algorithm was optimal; however hopefully this type of work will set a precedent for future systematic studies across imaging modalities.**

It is interesting that we had difficulty replicating the findings of previous manuscripts given that our samples are on par with or substantially larger than samples previously used in the literature (Fan et al., 2007; Sun et al., 2009; Yoon et al., 2007; Zanetti et al., 2013). **In support of this, there is literature that suggests that larger samples may in fact be problematic (Schnack and Kahn, 2016). Smaller studies typically have more tightly-controlled exclusion criteria, and therefore are often more homogeneous. As sample sizes grow, so does the heterogeneity of the sample. This may have the somewhat counter-intuitive and unexpected effect of decreasing performance as sample size increases.**

Leave-one-out cross-validation (LOOCV) is used throughout the SZ classification literature, although it has been criticized due to its tendency to over-fit data and

provide an inconsistent estimate of the true model (Shao, 1993). LOOCV was performed in the training set of only the best performing dataset and metric combination to illustrate the possibility of overfitting. Given this tendency of LOOCV to over-estimate generalizability, we expected to see higher estimates of accuracy from the LOOCV experiments, although this was only observed in the non-linear SVM case (69.8% LOOCV accuracy versus 68.8% using 10-fold cross-validation). For all algorithms, the LOOCV-derived models performed worse on the left-out dataset than the models derived from 10-fold cross-validation, which supports the hypothesis that LOOCV is prone to over-fitting in the training dataset, and does not always create the best model for unseen data.

Given the above factors, it is difficult to conclude that there is an optimal combination of image preprocessing, training, testing, and machine learning methods. However, we propose that cortical thickness inputs, when paired with LR or SVM, appear to provide the most robust estimates across datasets and methods. Further, we propose the use of 10-fold cross validation and testing on an "unseen" dataset may be critical to buffer against over-fitting models and to provide robust estimates of classification accuracy, sensitivity, and specificity. Larger datasets may also help to understand the behaviour of all permutations within the methodologies tested. **It would have been prohibitive to assess all possible variables that contribute to algorithm performance and conduct a fully controlled study where all variables are independent. We endeavored in this study, however, to shed some light on possible major contributing factors. Significant future work is needed to tease out in more detail the specific main effects and interactions.**

Although the sample sizes of our three datasets compare favourably with previous studies, a limitation in our study (and in most machine learning studies in neuroimaging) is the number of subjects we had available to us. This meant that the number of subjects we included was much lower than the number of variables we used. Increasing sample size may improve the reliability of our results, or perhaps better expose the limitations of dimensionality reduction, training, and machine learning methods. **Some of our results showed high sensitivity at the expense of low specificity. With a larger sample size, this could be mitigated by preserving the patient:control ratio within all folds of the 10-fold cross-validation.** Additionally, the number of features available from an MR image is enormous, and overfitting can be an issue with such large feature spaces. In order to manage this amount of data, some sort of dimensionality reduction must be performed, whether it be selecting a limited number of regions-of-interest (Greenstein, Malley, Weisinger, Clasen, & Gogtay, 2012; Nakamura et al., 2004; Ota et al., 2012; Pettersson-Yeo et al., 2013; Takayanagi et al., 2010, 2011; Yushkevich et al., 2005), image downsampling (Davatzikos et al., 2005; Nieuwenhuis et al., 2012), a PCA (Kasparek et al., 2011; Santos et al., 2010), an entirely novel method, (Fan et al., 2007; Zanetti et al., 2013), or some combination of thereof (Borgwardt et al., 2010; Karageorgiou et al., 2011; Nieuwenhuis et al., 2012). With improved computational power, it may be possible to retain all variables, and capture more of the subtle variability that exists between the brains of SZ patients and HC.

We used a binomial probability statistic, which indicates how likely a given result is assuming the data follows a binomial distribution, to assess how meaningful our accuracy results were. Another method for assessing this is permutation testing,

which although more computationally expensive, is more generalizable as it does not assume an inherent probability distribution. Additionally, it may be more suitable to classification studies that use cross-validation (Noirhomme et al., 2014). In our case, since our data is non-random and can only be assigned to one of two classes, the binomial probability test was sufficient; however, the possible benefits of permutation testing should be explored in future studies.

In conclusion, this study illustrates some of the limitations of applying machine learning to neuroimaging data, and suggests that perhaps cortical thickness and tissue densities are not reliable features for distinguishing between SZ and HC groups on a patient-by-patient basis using these methods. It is always tempting to adjust datasets and algorithms to boost accuracy after seeing the final results, but in this study, we endeavored to estimate the true discriminative ability of the algorithms, and limit data pre-processing and tailoring. Along with future, more extensive studies, the results from this study can be used to construct guidelines (such as most discriminative feature type, data preprocessing steps, dimensionality reduction, model selection, and training/validation paradigm) for performing classifications on novel datasets, which holds the promise of a clinical application supplementing symptom-based diagnoses.

Disclosure/Conflict of Interest

CF has received support from Janssen (Johnson & Johnson), and has served as consultant and/or speaker for AstraZeneca, Eli Lilly, and Janssen.

Acknowledgements

MMC is funded by the Fonds de Recherches Santé Québec, Canadian Institutes of Health Research (CIHR), Natural Sciences and Engineering Research Council of Canada, the Weston Brain Institute, the Alzheimer's Society of Canada, and Michael J. Fox Foundation for Parkinson's Research (MMC). ANV is supported by CHIR, the Ontario Mental Health Foundation, NARSAD, and the National Institute of Mental Health (R01MH099167) (ANV). ANV also acknowledges support from the CAMH Foundation, Michael and Sonja Koerner, the Kimel Family, and the Paul E. Garfinkel New Investigator Catalyst Award. This work was also supported by Consejo Nacional de Cienca y Technologia (CONACyT; 182279 to CF/AG), CONACyT project 261895 (CF), and CONACyT's Sistema Nactional de Investigadores (CF/AG). CF has also received support from the United States National Institute of Health, and the Instituto de Cienca y Technologia del DF.

References

American Psychiatric Association (2013). Diagnostic and Statistical Manual of Mental Disorders (DSM-5). Washington, D.C.: American Psychiatric Association.

Ashburner, J., Friston, K.J., 2000. Voxel-based morphometry–the methods. NeuroImage 11, 805–21. https://doi.org/10.1006/nimg.2000.0582

Bansal, R., Staib, L.H., Laine, A.F., Hao, X., Xu, D., Liu, J., Weissman, M., Peterson, B.S., 2012. Anatomical Brain Images Alone Can Accurately Diagnose Chronic Neuropsychiatric Illnesses. PLoS ONE 7, e50698. https://doi.org/10.1371/journal.pone.0050698

Borgwardt, S.J., Picchioni, M.M., Ettinger, U., Toulopoulou, T., Murray, R., McGuire, P.K., 2010. Regional Gray Matter Volume in Monozygotic Twins Concordant and Discordant for Schizophrenia. Biol. Psychiatry 67, 956–964. https://doi.org/10.1016/j.biopsych.2009.10.026

Csernansky, J.G., Wang, L., Jones, D., Rastogi-Cruz, D., Posener, J. a, Heydebrand, G., Miller, J.P., Miller, M.I., 2002. Hippocampal deformities in schizophrenia characterized by high dimensional brain mapping. Am. J. Psychiatry 159, 2000–6.

Davatzikos, C., Genc, a, Xu, D., Resnick, S.M., 2001. Voxel-based morphometry using the RAVENS maps: methods and validation using simulated longitudinal atrophy. NeuroImage 14, 1361–9. https://doi.org/10.1006/nimg.2001.0937

Davatzikos, C., Shen, D., Gur, R.C., Wu, X., Liu, D., Fan, Y., Hughett, P., Turetsky, B.I., Gur, R.E., 2005. Whole-brain morphometric study of schizophrenia revealing a spatially complex set of focal abnormalities. Arch. Gen. Psychiatry 62, 1218–27. https://doi.org/10.1001/archpsyc.62.11.1218

de la Fuente-Sandoval, C., León-Ortiz, P., Azcárraga, M., Stephano, S., Favila, R., Díaz-Galvis, L., Alvarado-Alanis, P., Ramírez-Bermúdez, J., Graff-Guerrero, A., 2013. Glutamate levels in the associative striatum before and after 4 weeks of antipsychotic treatment in first-episode psychosis: a longitudinal proton magnetic resonance spectroscopy study. JAMA Psychiatry 70, 1057–66. https://doi.org/10.1001/jamapsychiatry.2013.289

de la Fuente-Sandoval, C., León-Ortiz, P., Favila, R., Stephano, S., Mamo, D., Ramírez-Bermúdez, J., Graff-Guerrero, A., 2011. Higher levels of glutamate in the associative-striatum of subjects with prodromal symptoms of schizophrenia and patients with first-episode psychosis. Neuropsychopharmacol. Off. Publ. Am. Coll. Neuropsychopharmacol. 36, 1781–91. https://doi.org/10.1038/npp.2011.65

Demirci, O., Calhoun, V.D., 2009. Functional magnetic resonance imaging–implications for detection of schizophrenia. Eur. Neurol. Rev. 4, 103.

Fan, Y., Shen, D., Gur, R., 2007. COMPARE: classification of morphological patterns using adaptive regional elements. Med. Imaging IEEE … 26, 93–105.

Gaser, C., Nenadic, I., Buchsbaum, B.R., Hazlett, E. a., Buchsbaum, M.S., 2004. Ventricular Enlargement in Schizophrenia Related to Volume Reduction of the Thalamus, Striatum, and Superior Temporal Cortex. Am. J. Psychiatry 161, 154–156. https://doi.org/10.1176/appi.ajp.161.1.154

Greenstein, D., Malley, J.D., Weisinger, B., Clasen, L., Gogtay, N., 2012. Using multivariate machine learning methods and structural MRI to classify childhood onset schizophrenia and healthy controls. Front. Psychiatry 3, 53. https://doi.org/10.3389/fpsyt.2012.00053

Ho, B.-C., Andreasen, N.C., Ziebell, S., Pierson, R., Magnotta, V., 2011. Long-term antipsychotic treatment and brain volumes: a longitudinal study of first-episode schizophrenia. Arch. Gen. Psychiatry 68, 128–137. https://doi.org/10.1001/archgenpsychiatry.2010.199

Kambeitz, J., Kambeitz-Ilankovic, L., Leucht, S., Wood, S., Davatzikos, C., Malchow, B., Falkai, P., Koutsouleris, N., 2015. Detecting Neuroimaging Biomarkers for Schizophrenia: A Meta-Analysis of Multivariate Pattern Recognition Studies. Neuropsychopharmacology 40, 1742–1751. https://doi.org/10.1038/npp.2015.22

Karageorgiou, E., Schulz, S.C., Gollub, R.L., Andreasen, N.C., Ho, B.C., Lauriello, J., Calhoun, V.D., Bockholt, H.J., Sponheim, S.R., Georgopoulos, A.P., 2011. Neuropsychological testing and structural magnetic resonance imaging as diagnostic biomarkers early in the course of schizophrenia and related psychoses. Neuroinformatics 9, 321–333. https://doi.org/10.1007/s12021-010-9094-6

Kasparek, T., Thomaz, C.E., Sato, J.R., Schwarz, D., Janousova, E., Marecek, R., Prikryl, R., Vanicek, J., Fujita, A., Ceskova, E., 2011. Maximum-uncertainty linear discrimination analysis of first-episode schizophrenia subjects. Psychiatry Res. - Neuroimaging 191, 174–181. https://doi.org/10.1016/j.pscychresns.2010.09.016

Kawasaki, Y., Suzuki, M., Kherif, F., Takahashi, T., Zhou, S.Y., Nakamura, K., Matsui, M., Sumiyoshi, T., Seto, H., Kurachi, M., 2007. Multivariate voxel-based morphometry successfully differentiates schizophrenia patients from healthy controls. NeuroImage 34, 235–242. https://doi.org/10.1016/j.neuroimage.2006.08.018

Kay, S.R., Flszbein, A., Opfer, L.A., 1987. The positive and negative syndrome scale (PANSS) for schizophrenia. Schizophr. Bull. 13, 261.

Keerthi, S.S., Lin, C.-J., 2006. Asymptotic Behaviors of Support Vector Machines with Gaussian Kernel. Neural Comput. 15, 1667–1689.

Kim, D., Kim, J., Koo, T., Yun, H., Won, S., 2015. Shared and Distinct Neurocognitive Endophenotypes of Schizophrenia and Psychotic Bipolar Disorder 13, 94–102.

Lerch, J.P., Evans, A.C., 2005. Cortical thickness analysis examined through power analysis and a population simulation. NeuroImage 24, 163–173. https://doi.org/10.1016/j.neuroimage.2004.07.045

Liu, F., van der Lijn, F., Schurmann, C., Zhu, G., Chakravarty, M.M., Hysi, P.G., Wollstein, A., Lao, O., de Bruijne, M., Ikram, M.A., van der Lugt, A., Rivadeneira, F., Uitterlinden, A.G., Hofman, A., Niessen, W.J., Homuth, G., de Zubicaray, G., McMahon, K.L., Thompson, P.M., Daboul, A., Puls, R., Hegenscheid, K., Bevan, L., Pausova, Z., Medland, S.E., Montgomery, G.W., Wright, M.J., Wicking, C., Boehringer, S., Spector, T.D., Paus, T., Martin, N.G., Biffar, R., Kayser, M., 2012. A Genome-Wide Association Study Identifies Five Loci Influencing Facial Morphology in Europeans. PLoS Genet. 8, e1002932. https://doi.org/10.1371/journal.pgen.1002932

Nakamura, K., Kawasaki, Y., Suzuki, M., Hagino, H., Kurokawa, K., Takahashi, T., Niu, L., Matsui, M., Seto, H., Kurachi, M., 2004. Multiple structural brain measures obtained by three-dimensional magnetic resonance imaging to distinguish between schizophrenia patients and normal subjects. Schizophr. Bull. 30, 393.

Narr, K.L., Bilder, R.M., Toga, A.W., Woods, R.P., Rex, D.E., Szeszko, P.R., Robinson, D., Sevy, S., Gunduz-Bruce, H., Wang, Y.-P., DeLuca, H., Thompson, P.M., 2005. Mapping cortical thickness and gray matter concentration in first episode schizophrenia. Cereb. Cortex N. Y. N 1991 15, 708–19. https://doi.org/10.1093/cercor/bhh172

Nieuwenhuis, M., van Haren, N.E.M., Hulshoff Pol, H.E., Cahn, W., Kahn, R.S., Schnack, H.G., 2012. Classification of schizophrenia patients and healthy controls from structural MRI scans in two large independent samples. NeuroImage. https://doi.org/10.1016/j.neuroimage.2012.03.079

Noirhomme, Q., Lesenfants, D., Gomez, F., Soddu, A., Schrouff, J., Garraux, G., Luxen, A., Phillips, C., Laureys, S., 2014. Biased binomial assessment of cross-validated estimation of classification accuracies illustrated in diagnosis predictions. NeuroImage Clin. 4, 687–694.

Ota, M., Sato, N., Ishikawa, M., Hori, H., Sasayama, D., Hattori, K., Teraishi, T., Obu, S., Nakata, Y., Nemoto, K., Moriguchi, Y., Hashimoto, R., Kunugi, H., 2012. Discrimination of female schizophrenia patients from healthy women using multiple structural brain measures obtained with voxel-based morphometry: Screening for

schizophrenia. Psychiatry Clin. Neurosci. 66, 611–617. https://doi.org/10.1111/j.1440-1819.2012.02397.x

Pereira, F., Mitchell, T., Botvinick, M., 2009. Machine learning classifiers and fMRI: a tutorial overview. Neuroimage 45, S199–S209.

Pettersson-Yeo, W., Benetti, S., Marquand, A.F., Dell'Acqua, F., Williams, S.C.R., Allen, P., Prata, D., McGuire, P., Mechelli, A., 2013. Using genetic, cognitive and multi-modal neuroimaging data to identify ultra-high-risk and first-episode psychosis at the individual level. Psychol. Med. 43, 2547–2562. https://doi.org/10.1017/S003329171300024X

Plitman, E., de la Fuente-Sandoval, C., Reyes-Madrigal, F., Chavez, S., Gómez-Cruz, G., León-Ortiz, P., Graff-Guerrero, A., 2015. Elevated Myo-Inositol, Choline, and Glutamate Levels in the Associative Striatum of Antipsychotic-Naive Patients With First-Episode Psychosis: A Proton Magnetic Resonance Spectroscopy Study With Implications for Glial Dysfunction. Schizophr. Bull. 1–10. https://doi.org/10.1093/schbul/sbv118

R Core Team (2013). R Foundation for Statistical Computing, Vienna, Austria; R: A language and environment for statistical computing.

Santos, P.E., Thomaz, C.E., Santos, D. dos, Freire, R., Sato, J.R., Louza, M., Sallet, P., Busatto, G., Gattaz, W.F., 2010. Exploring the knowledge contained in neuroimages : Statistical discriminant analysis and automatic segmentation of the most significant changes 49, 105–115. https://doi.org/10.1016/j.artmed.2010.03.003

Schnack, H.G., Kahn, R.S., 2016. Detecting Neuroimaging Biomarkers for Psychiatric Disorders: Sample Size Matters. Front. Psychiatry 7. https://doi.org/10.3389/fpsyt.2016.00050

Schnack, H.G., Nieuwenhuis, M., van Haren, N.E.M., Abramovic, L., Scheewe, T.W., Brouwer, R.M., Hulshoff Pol, H.E., Kahn, R.S., 2013. 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. https://doi.org/10.1016/j.neuroimage.2013.08.053

Shao, J., 1993. Linear model selection by cross-validation. J. Am. Stat. Assoc. 88, 486–494.

Sun, D., van Erp, T.G.M., Thompson, P.M., Bearden, C.E., Daley, M., Kushan, L., Hardt, M.E., Nuechterlein, K.H., Toga, A.W., Cannon, T.D., 2009. Elucidating a Magnetic Resonance Imaging-Based Neuroanatomic Biomarker for Psychosis: Classification Analysis Using Probabilistic Brain Atlas and Machine Learning Algorithms. Biol. Psychiatry 66, 1055–1060. https://doi.org/10.1016/j.biopsych.2009.07.019

Takayanagi, Y., Kawasaki, Y., Nakamura, K., Takahashi, T., Orikabe, L., Toyoda, E., Mozue, Y., Sato, Y., Itokawa, M., Yamasue, H., Kasai, K., Kurachi, M., Okazaki, Y., Matsushita, M., Suzuki, M., 2010. Differentiation of first-episode schizophrenia patients from healthy controls using ROI-based multiple structural brain variables. Prog. Neuropsychopharmacol. Biol. Psychiatry 34, 10–17. https://doi.org/10.1016/j.pnpbp.2009.09.004

Takayanagi, Y., Takahashi, T., Orikabe, L., Mozue, Y., Kawasaki, Y., Nakamura, K., Sato, Y., Itokawa, M., Yamasue, H., Kasai, K., Okazaki, Y., Suzuki, M., 2011. Classification of First-Episode Schizophrenia Patients and Healthy Subjects by Automated MRI Measures of Regional Brain Volume and Cortical Thickness 6, 1–10. https://doi.org/10.1371/journal.pone.0021047

Voineskos, A.N., Lerch, J.P., Felsky, D., Shaikh, S., Rajji, T.K., Miranda, D., Lobaugh, N.J., Mulsant, B.H., Pollock, B.G., Kennedy, J.L., 2011. The brain-derived neurotrophic factor Val66Met polymorphism and prediction of neural risk for Alzheimer disease. Arch. Gen. Psychiatry 68, 198–206. https://doi.org/10.1001/archgenpsychiatry.2010.194

Wang, L., Kogan, A., Cobia, D., Alpert, K., Kolasny, A., Miller, M.I., Marcus, D., 2013. Northwestern University Schizophrenia Data and Software Tool (NUSDAST). Front. Neuroinformatics 7, 25. https://doi.org/10.3389/fninf.2013.00025

Wheeler, A.L., Chakravarty, M.M., Lerch, J.P., Pipitone, J., Daskalakis, Z.J., Rajji, T.K., Mulsant, B.H., Voineskos, A.N., 2013. Disrupted Prefrontal Interhemispheric Structural Coupling in Schizophrenia Related to Working Memory Performance. Schizophr. Bull. 1–11. https://doi.org/10.1093/schbul/sbt100

Yoon, U., Lee, J.M., Im, K., Shin, Y.W., Cho, B.H., Kim, I.Y., Kwon, J.S., Kim, S.I., 2007. Pattern classification using principal components of cortical thickness and its discriminative pattern in schizophrenia. NeuroImage 34, 1405–1415. https://doi.org/10.1016/j.neuroimage.2006.11.021

Yushkevich, P., Dubb, A., Xie, Z., Gur, R., Gur, R., Gee, J., 2005. Regional structural characterization of the brain of schizophrenia patients. Acad. Radiol. 12, 1250–1261. https://doi.org/10.1016/j.acra.2005.06.014

Zanetti, M. V, Schaufelberger, M.S., Doshi, J., Ou, Y., Ferreira, L.K., Menezes, P.R., Scazufca, M., Davatzikos, C., Busatto, G.F., 2013. Neuroanatomical pattern classification in a population-based sample of first-episode schizophrenia. Prog. Neuropsychopharmacol. Biol. Psychiatry 43, 116–25. https://doi.org/10.1016/j.pnpbp.2012.12.005

Zarogianni, E., Moorhead, T.W.J., Lawrie, S.M., 2013. Towards the identification of imaging biomarkers in schizophrenia, using multivariate pattern classification at a single-subject level. NeuroImage Clin. 3, 279–89. https://doi.org/10.1016/j.nicl.2013.09.003

Figure(s)

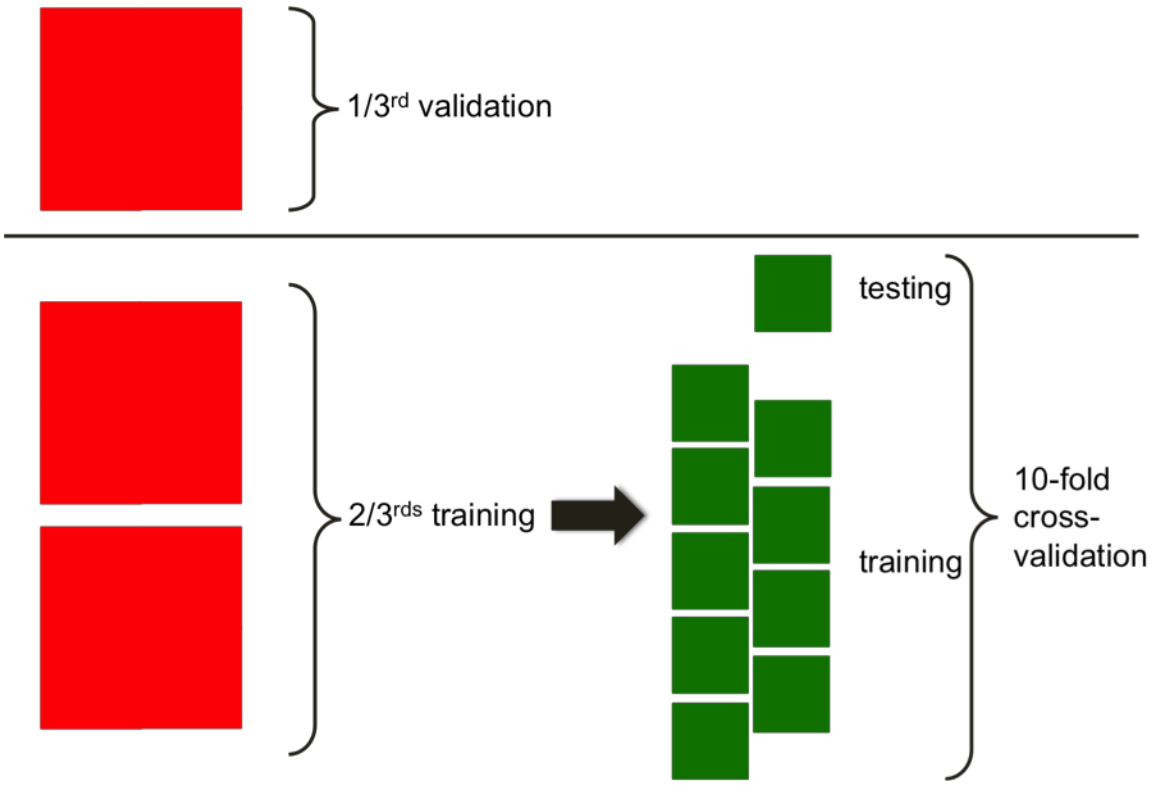

Figure 1: Validation scheme for training algorithms. Data was split into training and validation subsets (ratio 2:1, SZ:HC ratio retained). The validation data was set aside, and the training subset was further divided into training and testing groups to tune the model parameters using 10-fold cross-validation. The tuned models were then applied to the validation set. Performance is reported for both training and validation groups.

Table(s)

Table 1: Studies using structural MR imaging to classify schizophrenia patients and healthy controls included in Kambeitz et al., 2015

Study	Sample	Method	Validation	Metric	Feature Reduction	Accuracy
Bansal et al., 2012	65 HC, 40 SZ	Hierarchical Clustering	10 rounds of split-half & LOOCV	Surface morphology of cortical & subcortical ROIs	Spherical wavelet representation	93.3%
Borgwardt et al., 2012	22 HC, 23 FEP	SVM (non-linear)	Nested cross-validation	RAVENS (GM)	Multivariate filter method & PCA	86.7%
Davatzikos et al., 2005	79 HC, 69 SZ	SVM (non-linear)	LOOCV	RAVENS (GM, WM, CSF)	Image downsampling & COMPARE	81.1%
Fan et al., 2007	Female Sample: 38 HC, 23 SZ Male Sample: 41 HC, 46 SZ	SVM (non-linear)	LOOCV	RAVENS (GM, WM, CSF)	COMPARE	91.8% 90.8%
Greenstein et al., 2012	99 HC, 98 SZ	Random Forests	Out-of-bag (33% left out at each tree)	74 ROIs (cortical and subcortical volumes)	None	73.6%
Karageorgiou et al., 2011	47 HC, 28 ROS	LDA	LOOCV	95 ROIs & 75 neuropsychological variables	PCA	64.3% sensitivity, 76.6% specificity
Kasparek et al., 2011	39 HC, 39 FEP	LDA	Jackknife (LOOCV)	Local intensity features	PCA	71.8%
Kawasaki et al., 2007	Training Sample: 30 HC, 30 SZ Held-out Group: 16 HC, 16 SZ	LDA	LOOCV Held-out group	VBM (GM)	Eigenimage decomposition	76.7% 84.4%
Nakamura et al., 2004	Female Sample: 22 HC, 27 SZ Male Sample: 25 HC, 30 SZ	LDA	Unknown	8 ROIs on 3 coronal slices	Stepwise variable addition	81.6% 80.0%

Nieuwenhuis et al., 2012	Training Sample: 111 HC, 128 SZ Validation Sample: 122 HC, 155 SZ	SVM (linear)	LOOCV Independent sample	VBM (GM)	Image downsampling, selection of top 10% ranked discriminatory features & some ROI selection	71.4% 70.4%
Ota et al., 2012	Female Sample: 128 HC, 61 SZ	LDA	Held-out group (23 HC, 23 SZ)	VBM (GM & CSF)	ROIs & stepwise variable addition	71.7%
Petterson-Yeo et al., 2013	19 HC, 19 FEP	SVM (linear)	LOOCV	GM	None	63.2%
Sun et al., 2009	36 HC, 36 ROS	Sparse Multinomial Logistic Regression	LOOCV	Surface-based GM densities	None	86.1%
Santos et al., 2010	25 HC, 43 SZ	LDA	LOOCV	Voxel intensities	PCA	66.2%
Takayanagi et al., 2010	Male Sample: 24 HC, 17 FEP Female Sample: 24 HC, 17 FEP	LDA	LOOCV	Select ROIs	Stepwise variable addition	75.6% 82.9%
Takayanagi et al., 2011	Male Sample: 22 HC, 29 FEP Female Sample: 18 HC, 23 FEP	LDA	Held-out group (~30%) Held-out group (~30%)	Select ROIs & cortical thickness	Stepwise variable addition	86.7% 81.2%
Yushkevich et al., 2005	46 HC, 46 SZ	SVM	LOOCV	Select ROIs	LOOCV-based feature selection	70.7%
Zanetti et al., 2013	62 HC, 62 FE	SVM (non-linear)	LOOCV	RAVENS (GM, WM, CSF)	COMPARE	73.4%

Sample: FEP: First-Episode Psychosis; HC: Healthy Control; ROS: Recent-Onset Schizophrenia; SZ: Schizophrenia
Method: LDA: Linear Discriminant Analysis; SVM: Support Vector Machine
Validation: LOOCV: Leave-One-Out Cross-Validation
Metric: CSF: Cerebrospinal Fluid GM: Grey Matter; RAVENS: Regional Analysis of Volumes Examined in Normalized Space; ROI: Region of Interest; VBM: Voxel-Based Morphometry; WM: White Matter
Feature Reduction: COMPARE: Classification of Morphological Patterns using Adaptive Regional Elements; PCA: Principal Component Analysis

Table(s)

Table 2: Summary of study design. 42 combinations were studied in total across two validation schemes. VBM = voxel-based morphometry; GM = grey matter, RAVENS = Regional Analysis of Volumes in Normalized Space (Davatzikos et al., 2001); LR = logistic regression; LDA = linear discriminant analysis; SVM = support vector machine; COMPARE = Classification of Morphological Patterns using Recursive Feature Elimination (Fan et al., 2007).

Dataset	Feature Set	Classification Method	Validation Scheme
CAMH	VBM (GM)	LR	10-fold cross-val
NUSDAST	RAVENS (GM)	LDA	Held-out subset
INNN	Cortical thickness	Linear SVM	
		Non-linear SVM	
		COMPARE	

Table(s)

Table 3: Demographic characteristics of all datasets

	CAMH (1.5T)				NUSDAST (1.5T)				INNN (3T)			
Demographic	Schizophrenia Patients		Healthy Controls		Schizophrenia Patients		Healthy Controls		FEP Patients		Healthy Controls	
	(n=88)		(n=103)		(n=91)		(n=67)		(n=50)		(n=50)	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Age	36.6	12.4	35.2	12.4	32.1	12.2	25.8	9.80	25.8	7.35	23.8	4.79
Education (years)	13.3[a]	2.33	15.5[a]	1.91	12.0[a]	2.00	13.9[a]	2.56	11.9[a]	3.26	15.6[a]	2.57
Parental Education (years)	12.7[a]	7.00	17.0[a]	4.37	13.8	3.12	14.3	2.66	9.58[a]	5.19	13.9[a]	3.01
WTAR (IQ)	109[a]	15.7	117[a]	8.27	--	--	--	--	--	--	--	--
MMSE	28.9[a]	1.75	29.4[a]	0.833	--	--	--	--	--	--	--	--
CIRS	1.55[a]	0.811	0.824[a]	0.639	--	--	--	--	--	--	--	--
Age of onset	24.0	6.78	NA	NA	--	--	NA	NA	24.7	7.51	NA	NA
Illness Duration (weeks)	658	641	NA	NA	--	--	NA	NA	35.1	56.0	NA	NA
PANSS												
Positive	14.18	5.98	NA	NA	--	--	NA	NA	23.9	4.97	NA	NA
Negative	14.27	6.11	NA	NA	--	--	NA	NA	24.3	5.83	NA	NA
General	25.33	6.99	NA	NA	--	--	NA	NA	49.4	8.61	NA	NA
SAPS	--	--	NA	NA	22.4	16.8	NA	NA	--	--	NA	NA
SANS	--	--	NA	NA	30.5	17.3	NA	NA	--	--	NA	NA
	N		N		N		N		N		N	
Diagnosis	60 SZ	27 SA	NA		--	--	NA		11 BPD/ 18 SFD/ 21 SZ		NA	
Sex	58 M	30 F	56 M	46 F	61 M	30 F	39 M	28 F	31 M	19 F	32 M	18 F
Handedness	80 R	6 L	89 R	7 L	80 R	9 L	58 R	9 L	50 R		50 R	

[a] significant at $p<0.05$ between diagnostic groups

Abbreviations: SD = Standard Deviation; WTAR = Weschler Test for Adult Reading; MMSE = Mini-Mental State Exam; CIRS = Cumulative Illness Rating Scale; PANSS = Positive And Negative Syndrome Scale; SAPS = Scale for Assessment of Positive Symptoms; SANS = Scale for Assessment of Negative Symptoms; SZ = Schizophrenia; SA= Schizoaffective Disorder; BPD = Bipolar Disorder; SFD = Schizophreniform Disorder

Table(s)

Table 4: All training set results (using 10-fold cross-validation). Results are reported as % accuracy (averaged across all folds).

	Modulated VBM			RAVENS Maps			Cortical Thickness		
	CAMH	NUSDAST	INNN	CAMH	NUSDAST	INNN	CAMH	NUSDAST	INNN
LR	60.0[b]	57.4[a]	62.1[a]	60.8[c]	66.0[c]	70.0[c]	50.8	68.8[c]	57.6
SVM (linear)	55.0[a]	54.3	56.1	51.7	55.3[a]	66.7[c]	60.2[b]	71.9[c]	60.6[a]
SVM (RBF)	61.7[c]	64.2[b]	63.8[b]	63.3[c]	60.9[b]	69.8[c]	68.0[c]	68.8[c]	64.1[b]
LDA	59.2[b]	56.4[a]	57.6[a]	59.2[b]	57.5[a]	56.1	55.5[a]	65.6[c]	57.6[a]
COMPARE	63.3[c]	71.3[c]	71.2[c]	55.8[a]	50.0	50.0	--	--	--

Significance in binomial probability test (accuracies are significantly > than chance):
[a]$p<0.05$
[b]$p<0.01$
[c]$p<0.005$

Table(s)

Table 5: All validation set results. Results are reported as % accuracy (sensitivity/specificity).

	Modulated VBM			RAVENS Maps			Cortical Thickness		
	CAMH	NUSDAST	INNN	CAMH	NUSDAST	INNN	CAMH	NUSDAST	INNN
LR	62.7[a] (23.1/93.9)	65.2[a] (100/18.8)	62.5 (68.8/56.33)	67.8[c] (30.1/97.0)	60.9[a] (92.6/15.8)	56.3 (68.8/43.8)	60.3[a] (24.1/91.1)	70.8[c] (78.6/60.0)	69.7[a] (59.0/76.5)
SVM (linear)	55.9 (38.5/69.7)	63.0[a] (63.0/63.2)	62.5 (62.5/62.5)	52.5 (11.5/84.9)	69.6[b] (77.8/57.9)	65.6[a] (68.8/62.5)	61.9[a] (65.5/58.8)	68.1[c] (75.0/60.0)	58.8 (52.9/64.7)
SVM (RBF)	62.7[a] (38.5/81.8)	56.5 (92.6/5.26)	59.4 (56.3/62.5)	66.1[b] (46.2/81.8)	58.7 (100/0)	53.1 (87.5/18.8)	63.5[a] (44.8/79.4)	70.8[c] (78.6/60.0)	73.5[c] (58.8/88.2)
LDA	57.6 (34.6/75.8)	63.0[a] (77.8/42.1)	56.3 (56.3/56.3)	69.5[c] (38.5/94.0)	65.2[a] (81.5/42.1)	62.5 (68.8/56.3)	68.3[c] (65.5/70.6)	64.6[a] (67.9/60.0)	58.8 (41.2/76.5)
COMPARE	55.9 (65.4/48.5)	61.0[a] (81.5/31.6)	67.7[a] (56.3/80.0)	51.2 (52.0/45.1)	51.0 (53.1/50.4)	50.6 (43.2/52.4)	--	--	--

Significance in binomial probability test (accuracies are significantly > than chance):
[a]$p<0.05$
[b]$p<0.01$
[c]$p<0.005$

Acknowledgements

The authors have no acknowledgements to state.

Funding Sources

MMC is funded by the Fonds de Recherches Santé Québec, Canadian Institutes of Health Research (CIHR), Natural Sciences and Engineering Research Council of Canada, the Weston Brain Institute, the Alzheimer's Society of Canada, and Michael J. Fox Foundation for Parkinson's Research (MMC). ANV is supported by CHIR, the Ontario Mental Health Foundation, NARSAD, and the National Institute of Mental Health (R01MH099167) (ANV). ANV also acknowledges support from the CAMH Foundation, Michael and Sonja Koerner, the Kimel Family, and the Paul E. Garfinkel New Investigator Catalyst Award. This work was also supported by Consejo Nacional de Cienca y Technologia (CONACyT; 182279 to CF/AG), CONACyT project 261895 (CF), and CONACyT's Sistema Nactional de Investigadores (CF/AG). CF has also received support from the United States National Institute of Health, and the Instituto de Cienca y Technologia del DF.

***Contributors**

Author Winterburn designed the study, performed the data analysis, and wrote the manuscript. Authors Voineskos and Knight assisted with study design and general project supervision. Authors Devenyi and Bhagwat assisted with image processing, data analysis, and algorithm implementation. Authors de la Fuente-Sandoval, Plitman, and Graff provided insight and guidance in using the INNN dataset. Author Chakravarty assisted with study design, overall supervision, and manuscript writing. All authors contributed to and have approved the final manuscript.

Conflict of Interest

CF has received support from Janssen (Johnson & Johnson), and has served as consultant and/or speaker for AstraZeneca, Eli Lilly, and Janssen.

Supplementary Material for online publication only

Click here to download Supplementary Material for online publication only: Supplementary_Materials_and_Methods.doc