Learning discriminative functional network features of schizophrenia

Gheiratmand, Mina, Rish, Irina, Cecchi, Guillermo, Brown, Matthew, Greiner, Russell, et al.

Mina Gheiratmand, Irina Rish, Guillermo Cecchi, Matthew Brown, Russell Greiner, Pouya Bashivan, Pablo Polosecki, Serdar Dursun, "Learning discriminative functional network features of schizophrenia," Proc. SPIE 10137, Medical Imaging 2017: Biomedical Applications in Molecular, Structural, and Functional Imaging, 101371A (13 March 2017); doi: 10.1117/12.2264102

SPIE.

Event: SPIE Medical Imaging, 2017, Orlando, Florida, United States

Learning Discriminative Functional Network Features of Schizophrenia

Mina Gheiratmand[a,b,c], Irina Rish[d*], Guillermo Cecchi[d], Matthew Brown[a,c], Russell Greiner[a,b], Pouya Bashivan[e], Pablo Polosecki[d], Serdar Dursun[c]

[a]Dept. Computing Science, University of Alberta, 2-21 Athabasca Hall, Edmonton, AB T6G 2E8, Canada; [b]Alberta Machine Intelligence Institute, 2-47 Computing Science Centre, Edmonton, AB T6G 2E8, Canada; [c]Dept. Psychiatry, University of Alberta, 1E1 Walter Mackenzie Health Sciences Centre, 8440 112 St NW, Edmonton, AB T6G 2B7, Canada; [d]IBM T. J. Watson Research Center, 1101 Kitchawan Rd, Yorktown Heights, NY 10598, USA; [e]Dept. Electrical and Computer Engineering, University of Memphis, 3720 Alumni Ave, Memphis, TN 38152, USA

ABSTRACT

Associating schizophrenia with disrupted functional connectivity is a central idea in schizophrenia research. However, identifying neuroimaging-based features that can serve as reliable "statistical biomarkers" of the disease remains a challenging open problem. We argue that generalization accuracy and stability of candidate features ("biomarkers") must be used as additional criteria on top of standard significance tests in order to discover more robust biomarkers. Generalization accuracy refers to the utility of biomarkers for making predictions about individuals, for example discriminating between patients and controls, in novel datasets. Feature stability refers to the reproducibility of the candidate features across different datasets. Here, we extracted functional connectivity network features from fMRI data at both high-resolution (voxel-level) and a spatially down-sampled lower-resolution ("supervoxel" level). At the supervoxel level, we used whole-brain network links, while at the voxel level, due to the intractably large number of features, we sampled a subset of them. We compared statistical significance, stability and discriminative utility of both feature types in a multi-site fMRI dataset, composed of schizophrenia patients and healthy controls. For both feature types, a considerable fraction of features showed significant differences between the two groups. Also, both feature types were similarly stable across multiple data subsets. However, the whole-brain supervoxel functional connectivity features showed a higher cross-validation classification accuracy of 78.7% vs. 72.4% for the voxel-level features. Cross-site variability and heterogeneity in the patient samples in the multi-site FBIRN dataset made the task more challenging compared to single-site studies. The use of the above methodology in combination with the fully data-driven approach using the whole brain information have the potential to shed light on "biomarker discovery" in schizophrenia.

Keywords: schizophrenia, functional magnetic resonance imaging (fMRI), functional networks, multivariate predictive modeling, classification, predictive features

1. INTRODUCTION

Associating schizophrenia with disrupted functional connectivity is a central idea in modern schizophrenia research ("dysconnection" hypothesis[1]), which can be also traced back to the original work by Wernicke[2] and Bleuler[3]. However, identifying neuroimaging-based features that can serve as reliable "statistical biomarkers" of the disease remains a challenging open problem. Most prior work on biomarker discovery involves mass-univariate hypothesis testing to identify individual features that have significantly different empirical distributions across two populations, for example patients with schizophrenia vs. healthy controls. Mass-univariate analysis may reveal cross-population differences, but it is incapable of predicting the status (e.g. patient vs. healthy) of previously unseen individuals, given their feature sets.

* rish@us.ibm.com; phone 1 914 945 1896.

Medical Imaging 2017: Biomedical Applications in Molecular, Structural, and Functional Imaging
edited by Andrzej Krol, Barjor Gimi, Proc. of SPIE Vol. 10137, 101371A · © 2017 SPIE
CCC code: 1605-7422/17/$18 · doi: 10.1117/12.2264102

Moreover, statistical significance of individual features is neither a necessary nor a sufficient criterion for high performance in a predictive modeling setup[4], where combinations of features are considered and generalization to novel subjects is a key goal.

Predictive modeling has potential clinical applications in psychiatry – for example, early diagnosis of schizophrenia using neuroimaging data, predicting onset of psychosis in youth at high-risk using free speech[5] or identifying effective treatment for an individual. To date, no objective biomarker has been discovered in psychiatry that can be used to aid precise diagnosis or prognosis of treatment response. In recent years, predictive modeling in combination with neuroimaging data, particularly functional magnetic resonance imaging (fMRI) data, has shown promising results in discriminating patients diagnosed with schizophrenia from healthy controls, in datasets with small sample sizes[6] (see also Wolfers, et al.[7] for a review). For example, using functional connectivity network features extracted from fMRI data, Rish, et al.[6] achieved 93% accuracy in classification of schizophrenia patients from healthy control subjects in a dataset that included 11 patients and 11 healthy controls. Our objective here is to investigate the extent to which such findings generalize to different groups of patients and experimental paradigms in larger datasets. The move towards larger datasets that include heterogeneous patient samples is necessary for deployment of such predictive technologies in practical settings.

We evaluated the discriminative utility of functional network features derived from fMRI data using a wide range of classification models on a multi-site dataset (FBIRN)[8,9], in which cross-site variability introduced additional challenges. We argue that generalization accuracy and stability of candidate feature sets or "biomarkers" must be used as additional criteria on top of standard significance tests in order to discover more robust biomarkers. Here, we define biomarkers as multivariate patterns, i.e. specific combinations of features, rather than individual features. Generalization accuracy refers to the utility of features for making predictions about individuals, for example discriminating between patients and controls, in novel datasets. Feature stability refers to the reproducibility of the candidate features across different datasets. Functional connectivity network features (i.e. pairwise Pearson correlation coefficients) were extracted from fMRI blood oxygen-level dependent (BOLD) signals at two different spatial resolutions (scales): high (voxel-level) and low (supervoxel-level), with the aim of identifying the most discriminative set of features. Reducing the dimensionality of the fMRI data by spatial down-sampling permitted the use of all pairwise correlations across the whole brain. In contrast, for the original higher-resolution data, random subsampling from the feature space (>360M features) was required to reduce the computational expense. We compared statistical significance, stability and discriminative utility of both feature types in the multi-site fBIRN dataset[8], composed of fMRI and structural MRI data acquired from schizophrenia patients and healthy controls recruited at five different scanning sites.

For both feature types, whole-brain supervoxel-level pairwise correlations and a random subset of voxel-level pairwise correlations, a considerable fraction of features showed significant differences between the two groups. The most-significant features were similarly stable across multiple data subsets for both feature types. The whole-brain supervoxel functional connectivity features, however, showed a higher classification accuracy of 78.7% vs. 72.4% for the voxel-level features. Prediction tasks in multi-site studies are more challenging compared to single site studies because of the increased variability in the data due to both cross-site differences in image acquisition equipment as well as heterogeneity among patient samples compared to studies with a homogeneous patient group, such as Rish, et al. [6]. The use of the proposed tripartite framework, including assessment of prediction accuracy in addition to stability and statistical significance of candidate features, may lead to identification of more robust neuroimaging-based objective biomarkers. Finally, for each feature type, we display the stable subset of features that contributed to accurately discriminating between patients and healthy controls. Such analysis can also provide insights into abnormalities in brain functional connectivity networks in schizophrenia.

2. MATERIALS AND METHODS

2.1 Data

Our study used the FBIRN phase II multi-site fMRI dataset[8,9] (downloaded from fbirnbdr.nbirn.net:8080/BDR), focusing on the subset of the data acquired in response to an Auditory Oddball task[9]. We performed a routine series of per-subject preprocessing steps using the FSL software package[10]. These included removal of the first three fMRI volumes, motion correction, tCompCorr denoising[11], spatial smoothing (5 mm FWHH), high pass temporal filtering (cutoff period 100 s) and linear registration to the MNI template through subjects T1 images[12]. A brain mask was created by taking intersection of all subjects' brains. The total number of brain voxels remaining after applying the mask was 26,949. Voxel size was 3.4375 × 3.4375 × 5 mm. We excluded any subjects with translational motion, in any of the x, y, or z

directions, larger than the voxel size or with rotation larger than 0.06 radians. After the data preprocessing and quality control, 95 subjects remained, including 46 patients (35 male, 11 female; mean age: 38.98 ± 12.49 s.d.) with a diagnosis of schizophrenia or schizoaffective disorder, according to DSM-IV criteria, and 49 age- and sex-matched healthy controls (37 male, 12 female; mean age: 36.57 ± 12.96 s.d.). The data in this study came from five different scanning sites. Numbers of subjects from each site in the final set of subjects were 23, 23, 22, 21 and 6, respectively. Subjects participated in two sessions of fMRI data collection. For this study, data from the second scanning session was used. There were four runs per session, for each subject, resulting in a total of 380 (95 × 4) fMRI runs across all subjects. Subjects with missing runs were excluded.

We also downsampled[1] the preprocessed fMRI data at a rate of 4 × 4 × 3 relative to the original-resolution data (64 × 53 × 37) to reduce dimensionality. This allowed the use of whole-brain information in our analysis as described below in Section 2.2. The intersection of downsampled fMRI data from all subjects was computed to generate a universal brain mask at the lower resolution. All 137 brain volumes per subject were considered in computing the intersection.

2.2 Feature Extraction

For each subject, and each run, we extracted the functional network features at two different spatial resolutions (scales): *voxel*-level and *supervoxel*-level. In each case, the correlation matrix was computed as the pairwise Pearson correlation coefficients among all pairs of time-series $v_i(t)$, $v_j(t)$, where $v_k(t)$ corresponds to the BOLD signal of the *k*-th voxel or supervoxel. Time-series length was 137 volumes, sampled every 2 s (volume time = 2s).

1- ***voxel-level edge weights***: this feature set included a randomly-selected subset of 200,000 pairwise correlations out of 26,949 × 26,949 entries of the correlation matrix. The locations of pairs were randomly selected once, and then the same locations were used to derive features for all subjects. The rationale behind random sampling from the correlation matrix was to reduce the computational complexity of working with the full set of correlations, which would exceed 360 million features. Here we generated and analyzed only one random subset from the pool of the voxel-level correlations; Other random subsets may provide results that might vary in terms of feature subset stability, prediction accuracy, and/or the set of stable most-predictive features (described later in Section 3). A more comprehensive exploration of this feature space may identify better performing feature subsets ("biomarkers").

2- Supervoxel-level edge weights*:* since the total number of pair-wise correlations among all voxels in the brain was over 360 million, we spatially downsampled the fMRI data 4 × 4 × 3 times, which resulted in a total of 569 *supervoxel*s (size: 13.75 × 13.75 × 15 mm), in the subjects' universal mask. There were 161,596 pair-wise correlations among supervoxels. This number was low enough to enable a whole brain approach, using all within-brain supervoxel edge weights as features for classification.

Within-site feature standardization: The data in this study came from five different sites. To account for between-site variability in the features, we also used the within-site z-transformed version of the above features, where each feature was z-transformed within each site before combining the samples from different sites.

2.3 Analysis

Our methodology included (1) standard mass-univariate hypothesis testing of candidate features, complemented with (2) feature stability evaluation across multiple subsets of the data and (3) testing generalization accuracy of multivariate discriminative models built on top of those features. The procedure was applied to each feature type separately, either supervoxel-level edge weights or voxel-level edge weights.

2.4 Mass-univariate hypothesis testing

For each feature, we ran a t-test, with the null-hypothesis that the feature's values in the patient vs. control group came from distributions with equal means (assuming that the distributions were Gaussian with equal variances). The resulting p-values were corrected for multiple comparisons using Bonferroni and False Discovery Rate (FDR) correction with a corrected threshold of $\alpha = 0.05$.

We also used a p-value-based feature ranking as a simple feature selection (feature space dimension reduction) step during predictive modeling to reduce the risk of overfitting (see section 2.6).

[1] Downsampling was implemented using FSL FLIRT[12] using a reference image of dimensions 13 × 16 × 12 and voxel size 13.75 × 13.75 × 15mm, and nearest neighbors interpolation.

2.5 Feature stability evaluation

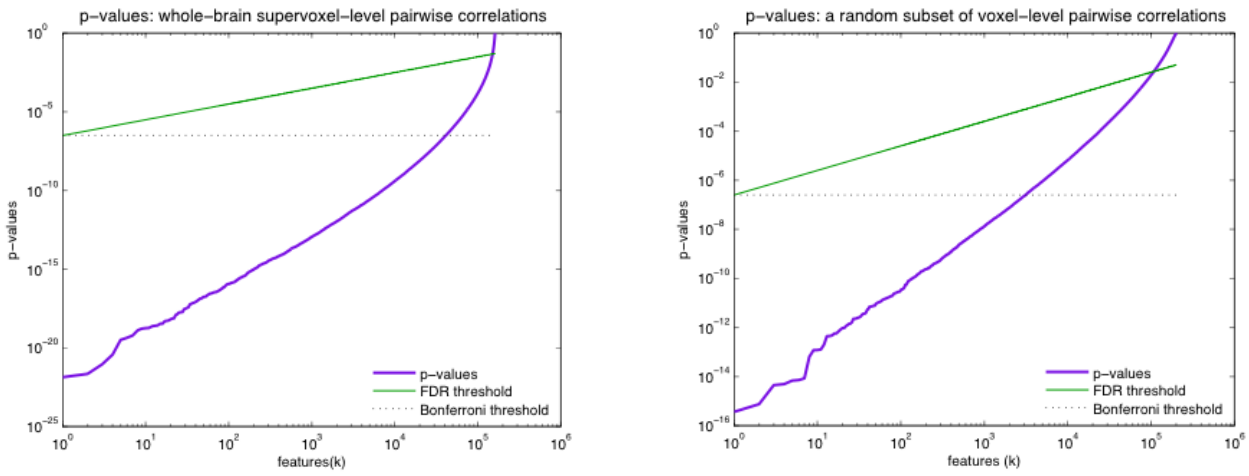
The feature stability analysis involved determining what fraction of features in each size-k top-ranking feature subset was common across all leave-one-subject-out data subsets. For this, we considered 95 different data subsets, each formed by leaving out one subject at a time. All four fMRI runs were excluded for the left-out subject. For various values of k ($k \in \{1, \ldots, N\}$, where N is the total number of features for each feature type and ks are arbitrarily-selected integers), the top-k features were computed in each of the 95 data subsets, then the intersection of these size-k subsets was computed, and its size was divided by k to obtain a ratio. This measure assumes that more stable feature subsets – i.e. those with larger fractions of overlapping features across different data subsets – might be more reliable biomarker candidates. (Future work involves developing a cross-validated feature stability analysis, which would allow for computing estimates of the expected feature overlap in novel datasets.)

2.6 Classification

We trained a range of classifiers on each of the functional network databases – both voxel- and supervoxel- level. The classifiers included Gaussian Naïve Bayes, SVM (linear and RBF kernels), Nearest Neighbors, Logistic Regression, Decision Tree, Random Forest and LDA (linear discriminate analysis). We considered various values of k (i.e. number of top-ranked feature), where feature ranking was based on univariate t-test-based significance. (To avoid biased estimates of the generalization error, feature ranking and subset formation was done in each cross-validation training data subset separately.) Leave-one-subject-out cross validation was used to evaluate generalization accuracy of each model (classifier/feature subset) on test samples. The hyperparameter space for different classifiers is presented in Supplementary Material.

3. RESULTS

Both feature types (whole-brain supervoxel-level pairwise correlations and random subset of voxel-level pairwise correlations) showed highly significant differences across the two groups. Supervoxel features showed significant differences after Bonferroni correction in 41,249 out of 161,596 supervoxel edge weights, that is approximately 26% of the total number of features. Voxel-level features showed significant differences after Bonferroni correction in 3,127 out of approximately 200,000 voxel-level edge weights[2], that is approximately 1.6% of total features. Much larger proportions of features survived FDR correction (approximately 93% and 54% for supervoxel and voxel-level features, respectively). Also see Figure 1.



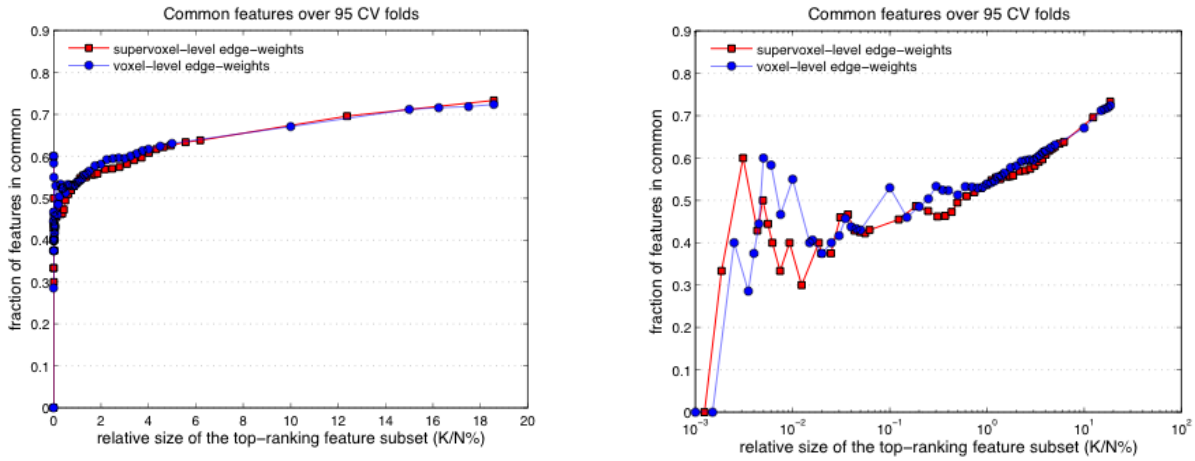
a) b)

Figure 1. t-test results for a) supervoxel-level and b) voxel-level edge-weight features (pairwise correlations). Features are ranked based on the p-values of their t-tests, from lowest p-value (the most significantly different feature) to the highest p-value (the least significant or insignificant feature). Both feature types are site-standardized.

[2] The exact number of valid random correlations was 199,993. This is because i and j indices in seven out of 200K (i, j) pairs were coincidentally the same, resulting in zero variance in pairwise correlation values among the samples for those particular features.

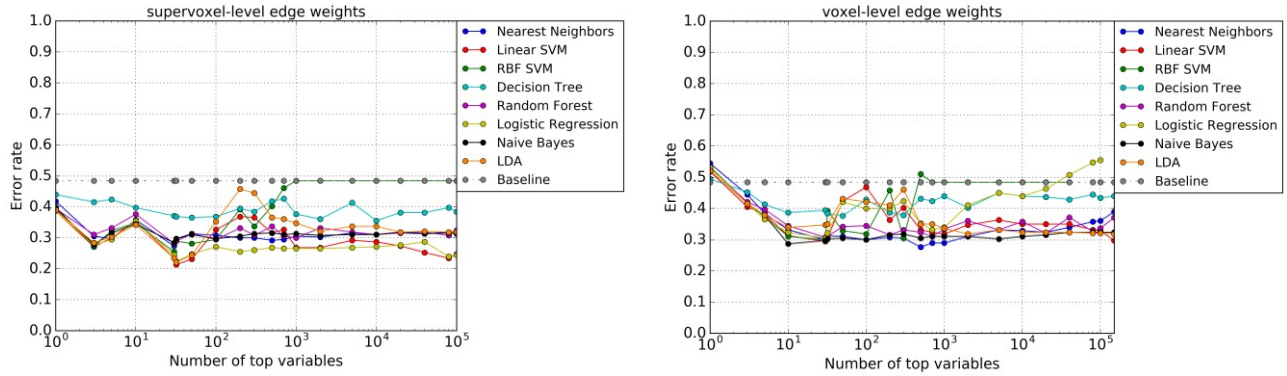
Both feature types were reasonably stable: there was over 70% feature overlap across data subsets in feature subsets with a relative size of less than 20% of the total number of features (Figure 2). Clearly, the fraction of overlapping features across data subsets reaches 1.0 once 100% of the features are used. Figure 2 displays the feature subset stability only over smaller relative sizes of feature subsets as this range is of particular interest if more interpretable models are desired - that is, models that use smaller number of features.

The supervoxel edge weights, however, demonstrated higher discriminative accuracy (up to 78.7% over 51.6% chance level) than the voxel-level edge weights (72.4%) (Figure 3). (It is noteworthy, however, that we only tested one random subset of voxel-level edge weights, as mentioned earlier in Section 2.2. Other random subsets may exist that yield higher prediction accuracies.) The FBIRN dataset has a higher level of variability compared to the more homogeneous data used in previous studies (e.g., Rish, et al.[6]) – primarily due to cross-site differences in image acquisition equipment (see Supplementary Table S1) as well as higher heterogeneity in our patient samples. Given this, we consider our classification accuracy quite encouraging, exceeding the results of similar multi-site studies (~70%[13,14]-75%[15]). Discriminative properties of the features were relatively stable across multiple classifiers, i.e. multiple classifiers showed similar performance for a per feature type (Figure 3).



a) b)

Figure 2. Stability of top-ranked feature subsets selected in cross-validation folds for supervoxel- vs. voxel-level edge-weight features, shown in a) linear scale and b) log-scale to highlight the small fractions zone. Plots show the fraction of features selected in all cross-validation folds for various-size subsets of top-ranked features. Both feature types are site-standardized.

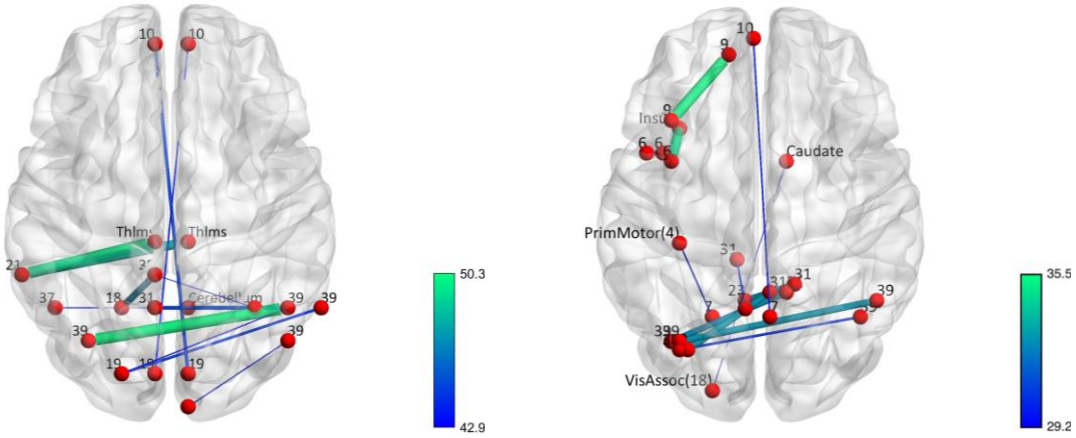


a) b)

Figure 3. Average cross-validation classification error for eight different classifiers using a) whole-brain supervoxel-level edge-weights and b) random subset of voxel-level edge-weights, as a function of number of variables included. False negative (FN) and false positive (FP) error rates for both feature types are presented in Supplementary Figure S1. Both feature types are site-standardized.

A large number of the most-predictive features remained stable across multiple folds (data subsets) in the cross-validating process. 12 links out of the top-ranked feature subsets of size 32 (resp., 30), for the supervoxel-level (resp., voxel-level) edge weight features, were common across all 95 folds. (25/32 and 22/30 links were common in >80% of the folds in the two feature types respectively.) Note that the above feature subset sizes yielded the lowest average classification errors (Figure 3), and thus the feature subset stability was particularly assessed for those feature subset sizes. Figure 4 visualizes, for both feature types, the top-ranking subsets of features that remained stable across all cross-validation subsets of the data. Interestingly, the set of the 12 stable most-discriminative links, for each feature type, were the same as the top 12 most-significant links (lowest p-values) on the full dataset.

The two brain maps in Figure 4 show that the locations of the stable most-discriminative subset of links at the two different spatial scales were very similar. Several common areas were involved including Brodmann areas 39, 31 and 10 corresponding to gross anatomical regions: angular gyrus, dorsal posterior cingulate, and dorsolateral prefrontal cortex (DLPFC), respectively. Most connections spanned across the left and right brain hemispheres, such as those between left and right angular gyrus, an area that has shown abnormal asymmetries in schizophrenia[16]. Indeed, the significant difference in functional connectivity involving BA39 between the healthy and schizophrenia groups might be a result of the abnormal asymmetry in angular gyrus in schizophrenia. (Note that besides being significantly different, these sets of features also contributed to prediction as a part of the size-k (k = 32 and k = 30 for the supervoxel- and voxel-level edge weight features respectively) feature subsets that yielded the lowest average classification error). DLPFC, an area in the prefrontal cortex known to be involved in executive functions, has also been repeatedly shown to be affected in schizophrenia, both in terms of activation[17] and functional connectivity to other brain regions, such as hippocampus[18], in response to memory tasks for example, as well as during rest[19].



a) b)

Figure 4. Map of the cross-validation-stable links common to all subsets of 32 (resp., 30) top-ranked (lowest p-value in univariate t-tests) features for a) whole brain supervoxel links and b) a random subset of voxel-level links from across the whole brain. Note that the linear SVM model using the top-32 (resp., 30) supervoxel-level (resp., voxel-level) links provided highest average accuracies of 78.7% (resp., 70.5%). Conversion of the nodes' x, y, z coordinates in the MNI space to Brodmann area labels was done using the BioImage Suite[20]. The numbers next to the nodes denote Brodmann area labels, e.g. 39 means Brodmann Area 39. Left and right hemispheres of the brain are represented on the left and right sides of the above brain images respectively. Thickness and color of the links are scaled based on -ln(p-value). A list of all 12 area-to-area connections for each feature type is presented in Supplementary Table S2.

4. CONCLUSIONS

We demonstrate that brain-wide functional connectivity features extracted from fMRI data are capable of discriminating between schizophrenic patients and controls with quite high accuracy, even in challenging datasets such as FBIRN that includes data from multiple sites and highly heterogeneous group of patients. We observe that, while both statistical significance and feature stability across cross-validation folds were high for both feature types (i.e. whole-brain supervoxel edge weights and a random subset of voxel-level edge weights), the former were more discriminative than

the latter, yielding a considerably higher accuracy of about 78.7% (vs. 72.4%). Thus, from a methodological perspective, we demonstrate that statistical testing alone may not be a sufficiently strong criterion for statistical biomarker discovery, and must be combined with a wider range of criteria, including generalization accuracy and stability. The above methodology, in combination with the fully data-driven approach employed here, which uses the information from the whole brain, have the potential to provide insights into possible biomarkers of schizophrenia.

ACKNOWLEDGEMENT

The authors would like to thank Biomedical Informatics Research Network (BIRN) for sharing FBIRN phase II dataset (supported by grants to the Function BIRN [U24-RR021992] Testbed funded by the National Center for Research Resources at the National Institute of Health, U.S.A.), and WestGrid (www.westgrid.ca) and Compute Canada for providing high performance computing resources and technical support. Financial support for this work was provided by Alberta Innovates – Health Solutions, IBM via the IBM Alberta Centre for Advanced Studies, the Natural Sciences and Engineering Research Council of Canada (NSERC) and the Alberta Machine Intelligence Institute (AMII).

REFERENCES

[1] Friston, K. J. & Frith, C. D. Schizophrenia: a disconnection syndrome? *Clin Neurosci* **3**, 89-97 (1995).
[2] Wernicke, C. Grundrisse der Psychiatrie. *Thieme* (1906).
[3] Bleuler, E. *Dementia Praecox or the Group of Schizophrenias*. (International Universities Press, 1911).
[4] Lo, A., Chernoff, H., Zheng, T. & Lo, S. H. Why significant variables aren't automatically good predictors. *Proc Natl Acad Sci U S A* **112**, 13892-13897, doi:10.1073/pnas.1518285112 (2015).
[5] Bedi, G. *et al.* Automated analysis of free speech predicts psychosis onset in high-risk youths. *NPJ Schizophr* **1**, 15030, doi:10.1038/npjschz.2015.30 (2015).
[6] Rish, I. *et al.* Schizophrenia as a network disease: disruption of emergent brain function in patients with auditory hallucinations. *PLoS One* **8**, e50625, doi:10.1371/journal.pone.0050625 (2013).
[7] Wolfers, T., Buitelaar, J. K., Beckmann, C. F., Franke, B. & Marquand, A. F. From estimating activation locality to predicting disorder: A review of pattern recognition for neuroimaging-based psychiatric diagnostics. *Neurosci Biobehav Rev* **57**, 328-349, doi:10.1016/j.neubiorev.2015.08.001 (2015).
[8] Potkin, S. G. & Ford, J. M. Widespread cortical dysfunction in schizophrenia: the FBIRN imaging consortium. *Schizophr Bull* **35**, 15-18, doi:10.1093/schbul/sbn159 (2009).
[9] Keator, D. B. *et al.* The Function Biomedical Informatics Research Network Data Repository. *Neuroimage* **124**, 1074-1079, doi:10.1016/j.neuroimage.2015.09.003 (2016).
[10] Jenkinson, M., Beckmann, C. F., Behrens, T. E., Woolrich, M. W. & Smith, S. M. Fsl. *Neuroimage* **62**, 782-790, doi:10.1016/j.neuroimage.2011.09.015 (2012).
[11] Behzadi, Y., Restom, K., Liau, J. & Liu, T. T. A component based noise correction method (CompCor) for BOLD and perfusion based fMRI. *Neuroimage* **37**, 90-101, doi:10.1016/j.neuroimage.2007.04.042 (2007).
[12] Jenkinson, M., Bannister, P., Brady, M. & Smith, S. Improved optimization for the robust and accurate linear registration and motion correction of brain images. *Neuroimage* **17**, 825-841 (2002).
[13] Gollub, R. L. *et al.* The MCIC collection: a shared repository of multi-modal, multi-site brain image data from a clinical investigation of schizophrenia. *Neuroinformatics* **11**, 367-388, doi:10.1007/s12021-013-9184-3 (2013).
[14] Anderson, A. & Cohen, M. S. Decreased small-world functional network connectivity and clustering across resting state networks in schizophrenia: an fMRI classification tutorial. *Front Hum Neurosci* **7**, 520, doi:10.3389/fnhum.2013.00520 (2013).
[15] Cheng, W. *et al.* Voxel-based, brain-wide association study of aberrant functional connectivity in schizophrenia implicates thalamocortical circuitry. *NPJ Schizophr* **1**, 15016, doi:10.1038/npjschz.2015.16 (2015).
[16] Niznikiewicz, M. *et al.* Abnormal angular gyrus asymmetry in schizophrenia. *Am J Psychiatry* **157**, 428-437, doi:10.1176/appi.ajp.157.3.428 (2000).
[17] Potkin, S. G. *et al.* Working memory and DLPFC inefficiency in schizophrenia: the FBIRN study. *Schizophr Bull* **35**, 19-31, doi:10.1093/schbul/sbn162 (2009).
[18] Meyer-Lindenberg, A. S. *et al.* Regionally specific disturbance of dorsolateral prefrontal-hippocampal functional connectivity in schizophrenia. *Arch Gen Psychiatry* **62**, 379-386, doi:10.1001/archpsyc.62.4.379 (2005).

[19] Zhou, Y. *et al.* Functional dysconnectivity of the dorsolateral prefrontal cortex in first-episode schizophrenia using resting-state fMRI. *Neurosci Lett* **417**, 297-302, doi:10.1016/j.neulet.2007.02.081 (2007).
[20] Lacadie, C. M., Fulbright, R. K., Rajeevan, N., Constable, R. T. & Papademetris, X. More accurate Talairach coordinates for neuroimaging using non-linear registration. *Neuroimage* **42**, 717-725, doi:10.1016/j.neuroimage.2008.04.240 (2008).

SUPPLEMENTARY MATERIAL

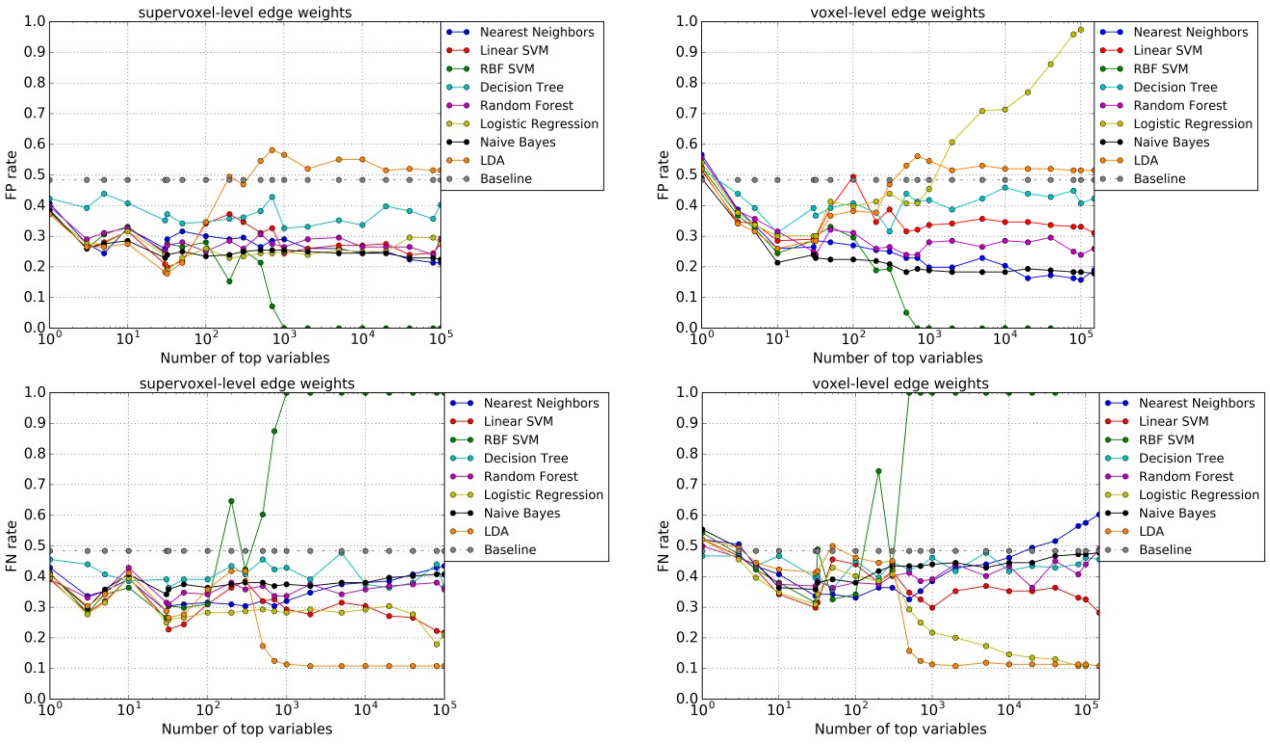
Classifiers hyperparameter space

Classifiers hyperparameters were selected based on a randomized search over the parameter space (iterations = 10) and optimized (with respect to accuracy) via an internal 5-fold cross validation. Parameter space for different classifiers were as follows:

Linear SVM: {C: 1};
RBF SVM: {C: [0.01, 0.1, 1.0, 10.0, 100.0], gamma: [0.01, 0.025, 0.063, 0.158, 0.398, 1.0, 1.778, 3.162, 5.623, 10.0]};
Logistic Regression: {C: [1.3, 3.8, 11.7, 35.5, 108.0, 328.6, 1000.0]};
Random Forest: {max_depth: [2, 3, 5, 7, 10, 16]};
Nearest Neighbors: {n_neighbors: [1, 5, 10, 20]}};
LDA: {};
Decision Tree: {};
Naive Bayes: {}.

The pipeline for our learning algorithm (including within-fold feature ranking/feature subset selection and classification) was implemeted in python 2.7.

Supplementary Figure: FP and FN error rates



a) b)

Figure S1. Average cross-validation classification FN and FP error rates for 8 different classifier using a) whole-brain supervoxel-level edge-weights; and b) random subset of voxel-level edge-weights. Average total error for each feature type is presented in Figure 3.

Supplementary Tables

Table S1. Scanner and image acquisition information per site.

Site code	MR Field strength	make	Head coil	k-space
3	1.5T, 4T	GE	Volume birdcage	(spiral)
9	1.5T	Marconi	Bird-cage	linear
10	1.5T	Siemens	CP for EPI; 8 channel for T1	linear
18	3T	Siemens	CP	linear
6	3T	Siemens	8-channel	linear

Table S2. Stable subset of 12 most-discriminative a) supervoxel-level and b) voxel-level links (see also Figure 4). (VisAssoc: Visual Association cortex; LOCsd: Lateral Occipital Cortex, superior division; PrimMotor: Primary Motor Cortex; L and R represent left and right hemispheres.)

a)

Link_index	From	↔	To	Link Strength
1	BA39.L	↔	BA39.R	50.35
2	BA21.L	↔	Thalamus.L	49.88
3	BA21.L	↔	Thalamus.R	48.47
4	VisAssoc(18).L	↔	BA30.L	47.06
5	BA39.R	↔	BA31.L	44.89
6	BA10.L	↔	BA19.R	44.58
7	BA39.R	↔	BA19.L	44.25
8	BA10.R	↔	BA19.L	43.52
9	BA39.R	↔	LOCsd.L	43.24
10	Vis.Assoc(18).R	↔	BA39.R	43.18
11	Cerebellum.R	↔	Fusiform(37).L	43.15
12	BA39.R	↔	BA23.L	42.9

b)

Link_index	From	↔	To	Link Strength
1	BA9.L	↔	BA9.L	35.56
2	Insula.L	↔	BA6.L	34.83
3	BA31.R	↔	BA39.L	33.05
4	BA39.L	↔	BA39.R	32.95
5	BA39.L	↔	BA31.R	32.62
6	BA23.L	↔	BA31.R	32.56
7	BA6.L	↔	BA6.L	32.41
8	BA39.R	↔	BA39.L	30.39
9	BA31.L	↔	BA7.L	29.76
10	BA10.L	↔	BA7.R	29.75
11	BA7.L	↔	PrimMotor(4).L	29.69
12	VisAssoc(18).L	↔	Caudate.R	29.24