Author's response to reviews

Title: Pattern of neural responses to verbal fluency shows diagnostic specificity for schizophrenia and bipolar disorder

Authors:

Sergi G Costafreda (sergi.1.costafreda@kcl.ac.uk)
Cynthia HY Fu (cynthia.fu@kcl.ac.uk)
Marco Picchioni (marco.picchioni@kcl.ac.uk)
Timothea Touloupoulou (timothea.touloupoulou@kcl.ac.uk)
Colm McDonald (colm.mcdonald@kcl.ac.uk)
Muriel Walshe (muriel.walshe@kcl.ac.uk)
Diana Prata (diana.prata@kcl.ac.uk)
Robin M Murray (robin.murray@kcl.ac.uk)
Philip K McGuire (philip.mcguire@kcl.ac.uk)

Version: 2 Date: 3 December 2010

Author's response to reviews: see over
1st December 2010

Dr. Shukwinder Shergill
Editor, BMC Psychiatry

Dear Dr. Shergill

RE: MS: 2109675770445390
Pattern of neural responses to verbal fluency shows diagnostic specificity for schizophrenia and bipolar disorder. Costafreda et al.

Thank you for the reviews of our submission. We have addressed each of the Reviewers comments and have modified our manuscript with a point-by-point reply to each comment. We have also introduced in the revised manuscript the changes requested by your office.

We hope this revised manuscript is acceptable for publication in BMC Psychiatry.

Sincerely yours,

Sergi G. Costafreda
Response to Reviewers’ comments:

Pattern of neural responses to verbal fluency shows diagnostic specificity for schizophrenia and bipolar disorder. Costafreda et al.

RE: MS: 2109675770445390

1 December 2010

Reviewer: Stephen Lawrie

Costafreda et al. use machine learning to discriminate between schizophrenia and bipolar disorder and healthy controls using verbal fluency fMRI data. Altered function in plausible brain networks achieved high (80-90%) group discrimination. The manuscript is well written and the scientific message clear.

I only have minor essential revisions:

In the results section, the machine learning classification analysis sensitivity and specificity figures are given but these presumably can be increased or decreased by altering settings during the machine learning. In other words, is the machine learning classification generally taken to optimally balance sensitivity and specificity, or is it prioritising one over the other? Please comment in the methods, results or discussion.

We thank Prof. Lawrie for his comments. In our approach, the classification of each individual within a diagnostic category was based on the boundaries defined by the SVM algorithm. These boundaries are hyperplanes in the high-dimensional space defined by the training points. SVM finds an “optimal” hyperplane, defined as the one that maximises the margins, or distances between the hyperplane and the nearest training examples. The rationale for considering this an optimal separation hyperplane is that the larger the margins, the lower the expected error of the trained classifier in new, unclassified examples (the generalization error).

The procedure is therefore symmetrical with regards to the type of error (confusing a patient for a control or vice versa); in other words, it tries to minimize the expected generalization error (or equivalently, maximise the classification accuracy in new examples), without prioritising either sensitivity or specificity.

For some clinical applications though, errors may not be equivalent (the cost of misdiagnosing a patient for a control is not the same as misdiagnosing a control for a patient) and thus it would be advantageous to optimise either sensitivity or specificity. This, however, would carry the disadvantage of a potential drop in global accuracy. The present paper aims to make a proof of concept for the
potential of the neural correlates of verbal fluency as a diagnostic biomarker, and this goal in our view is best served by optimising the overall accuracy.

We have clarified this issue in the methods section by adding the following paragraph to the Methods (in page 7):

“As implemented here, the procedure finds the boundary that maximises the expected overall classification accuracy in new, unclassified examples. This boundary therefore treats as equivalent two types of errors: false positives (FP, e.g. labelling a control as patient) and false negatives (FN, misdiagnosing a patient as a control). For some clinical applications, such types of errors may not be equivalent. For example, if the clinical goal is to confirm the presence of a disorder, a better classification rule would be one that ensures a low FP rate (high specificity) while tolerating a higher FN rate (lower sensitivity) and potentially a lower overall classification accuracy. Our purpose in the present paper, though, was to establish the potential of the neural correlates of verbal fluency as a diagnostic biomarker, and this proof-of-principle goal benefits from optimising the overall diagnostic accuracy rather than sensitivity or specificity.”

The authors discuss the importance of current and previous psychotic symptoms in the bipolar group but as far as I can see have not mentioned how many of their subjects have suffered from psychotic symptoms in the past.

We can confirm that none of the bipolar subjects were actively psychotic at the time of the scan, however the presence of psychotic symptoms in past manic or depressive episodes was not consistently recorded during the assessment. We have extended the acknowledgement of this limitation to the Discussion (Page 12):

“It is also possible that past psychotic symptoms in bipolar subjects may have impaired their differentiation from schizophrenia subjects. While we can confirm that none of the bipolar subjects were actively psychotic at the time of the scan, the presence of psychotic symptoms in past manic or depressive episodes was not consistently recorded during the assessment.”

The main limitation of this paper and indeed the field as a whole is the lack of an independent test set of patients to evaluate the machine learning parameters in another group. This should be acknowledged.

We agree with Professor Lawrie’s comment that a complete assessment of the diagnostic utility of our approach requires testing in a fully independent sample. We have attempted to partially mitigate this difficulty by using cross-validation, as is common in the field. To acknowledge this limitation, we have added the following to the Limitations section in the Discussion (page 12): “Finally, although we used leave-one-out cross-validation to ensure that the classification algorithm was tested in different subjects from the ones on which it was developed, a complete assessment of the clinical utility of the diagnostic algorithm should include testing in a fully independent set of patients, recruited in a different clinical setting.”
The discussion mentions that this is an unusual sample, being comprised of approximately 50% twins. This might reduce variance within each of the groups and thereby increase separation between diagnoses artificially. The other difference between the groups which should also be acknowledged is that the bipolar disorder group are approximately 6 years older on average than either of the other 2 groups, which again might have been expected to facilitate the differentiation.

We agree with Professor Lawrie. We have added the following sentence in the Discussion (page 12) to acknowledge the difference in age for bipolar subjects: “Bipolar subjects were also on average 6 years older than either of the other two groups, which may have facilitated diagnostic classification.”

Regarding the presence of twins, we ensured that only 1 subject from each twin set was included in the sample through random selection, so that all subjects can be considered statistically independent from all the other subjects in the final sample. This selection process ensures that the variance within each of the groups is similar to what would be found in a sample of non-twin subjects, and should not be a factor in the classification. We have clarified this process and its implications in the Methods (page 4): “from the Twin studies, only 1 subject from each twin set was included to ensure that each individual could be considered statistically independent from the other subjects in the final sample; the inclusion of non-independent subjects could have reduced the variance within each of the groups thereby increasing separation between diagnoses artificially.”

**Reviewer: Bert G Park**

This is an interesting paper which shows the use of support vector machines to classify functional imaging data obtained during a verbal fluency task from patients with schizophrenia and bipolar disorder, both from each other and from healthy controls with reasonable sensitivity and specificity.

Major compulsory revisions

1. The paper would benefit from further consideration of the other neuroimaging modalities and methods which have been applied in trying to find a diagnostic tool for schizophrenia and bipolar disorder. Nenadic et al (2009 Neuroimage) found a very similar classification performance in the subsyndromes of schizophrenia against controls using structural MRI images alone. A question which then needs to be asked is what is the additional advantage gained by using fMRI given it’s additional costs? Some discussion is also warranted as to what levels of sensitivity and specificity would be required for such a test to be of clinical use.

We thank Dr. Park for his comments and for pointing us to the recent papers of Nenadic et al and Schmah et al.

We view the modalities of functional and structural MRI as complementary sources of information to achieve clinical prediction. While there is likely to be some overlap between the regions reported as abnormal by both structural and
functional MRI, the findings are not identical. For instance, while both Nenadic et al. and our study point to abnormalities in prefrontal areas (dorsolateral, inferior frontal and anterior cingulate), our findings additionally stress the importance of reduced deactivation in default-mode network regions (precuneus, posterior cingulate, angular gyri) which do not appear to be affected in the Nenadic et al. paper. Our work in depression also seems to support the notion that functional and structural MRI may tap into different abnormalities with different predictive potential (refs 8,17,59). These differences of abnormalities could potentially be exploited to produce a diagnostic test utilizing both structural and functional MRI that may be superior to either of them in isolation. This is important because as suggested by Dr. Park, clinical utility would require very high levels of diagnostic accuracy. As the field of neuroimaging-based classification is in its infancy, our view is that we should focus in maximising its potential, and multimodality integration seems a promising way forward.

We have addressed this issue in the Discussion (page 10):
“...The classification analysis revealed over 90% sensitivity and specificity for the detection of schizophrenia relative to both bipolar subjects and matched healthy controls. Similarly high diagnostic utility has been reported for the diagnosis of schizophrenia based on the fMRI neural correlates of an auditory oddball task [56], and VBM-derived structural differences [57, 58]. Notably the basis for such accurate diagnostic decision has not been identical across studies and tasks: for instance, while prefrontal deficits have been prominent in both VBM-based and fMRI-based classification studies, abnormalities in posterior regions such as precuneus and posterior cingulate have only been reported in fMRI-based classification [56, and the present paper]. Our work on neuroimaging-based prediction in depression has also shown that functional and structural MRI may convey complementary predictive information [8,17,59]. A promising way to further optimize diagnostic performance may therefore be the fusion of complementary information from structural and functional MRI that may be superior to either of them in isolation. Increased performance, even above the encouraging figures reported so far, is likely to be necessary to achieve clinical utility. ”

2. More detail in the methods section needs to be given to the authors’ use of classification method. Applying machine learning methods to fMRI data is a developing field (Schmah et al 2010 Neural Computation 22:2729) and from the range of available machine learning algorithms and indeed variants of support vector machine it would be interesting to know why the authors chose the particular approach they did. Similarly it would also be good to know which R libraries/functions were actually used.

The choice of the SVM approach was motivated by two factors: SVM has been proven to be a high-performance, robust and versatile classification method, not only for neuroimaging-based classification but also in other areas where highly-dimensional datasets are used for clinical prediction. In head-to-head comparative studies, SVM consistently achieves one of the best performances. In
the Schmah et al. paper this is shown for fMRI-based prediction, and similar results have been shown for proteomics (Lee et al, 2005, ref [30]) and genomics (Wu et al, 2003, ref [31]) among other fields. We have developed expertise in the use of SVM for neuroimaging-based patient classification, as reflected in recent publications that use this methodology (references [8],[17] and [59]).

We have clarified this in the Methods (pages 6-7): “We employed Support Vector Machines (SVM) classification analysis [28], which has been shown to be a powerful tool for statistical pattern recognition. SVM has proven to be a robust and versatile approach for clinical prediction, as demonstrated by its consistently high performance in head-to-head methodological comparisons of diverse machine learning methods performed with fMRI data [29] and other high-dimensional clinical datasets such as proteomics [30] and genomics [31]. Our group has also demonstrated the potential of linear SVM for neuroimaging-based prediction in depression [8, 17].”

We have also clarified the libraries used in the analysis in the Methods section (page 7): “Additional analyses were performed using the following packages of the R statistical software [33]: AnalyzeFMRI which offers input/output, visualisation and analysis functions for fMRI data and the e1071 package, which supplies an interface to the libsvm library ([http://www.csie.ntu.edu.tw/~cjlin/libsvm/]).”

It was good to see that the authors included the initial feature extraction (the second level ANOVA) within the leave one out cross validation, but a discussion of why the authors chose to use an initial feature extraction of verbal fluency data rather than say a full dataset of five minutes of resting state data would again be informative.

We thank Dr Park for this comment. We chose the verbal fluency task because its performance is affected in both schizophrenia and in bipolar disorder. It is a task that is well placed to tap into potential abnormalities in executive function and language, which are central to schizophrenia and relevant for bipolar disorder. It was therefore possible to elucidate whether the neural correlates of the task showed any commonalities in bipolar and schizophrenia patients relative to controls and whether there were any difference according to the specific disorder. By using machine learning techniques, we were able to directly evaluate its diagnostic specificity, which, as expected, was higher for schizophrenia than for bipolar disorder. Both our study and that of Calhoun and colleagues (2008) coincide in finding that reduced deactivation in the default mode network is present in both schizophrenia and bipolar disorder, and has diagnostic potential. At least in our case this was an unexpected finding, as we would have expected functional differences to be limited to prefrontal, and perhaps temporal, regions. Applying machine learning classification to resting state data is also a promising line of enquiry.

We have added this in the Discussion (page 11) “An observation in both Calhoun and colleagues and the present work, is the relevance of default mode network abnormalities for diagnostic purposes. We had anticipated that functional
differences would be largely confined to prefrontal regions. This convergence of findings across two different tasks suggests that applying machine learning classification to resting state data may also be a promising line of enquiry.

**Minor Essential Revisions**

1. For a paper which is looking to assess the performance of a diagnostic test it is a little unusual that consensus methods were used rather than the gold standard of structured diagnostic interview. This should be acknowledged.

   We agree with the comment and have added the following sentence to the Limitations section in the Discussion (page 12): “Patient diagnoses were ascertained through consensus methods by consultant psychiatrists, rather than with a structured diagnostic interview, potentially leading to lower diagnostic certainty.”

2. The authors rightly acknowledge the potentially important confound of medication status. Given recent concern about the extent to which this confound has been overlooked in neuroimaging studies (Moncrieff & Leo 2010 Psych Med) this is all the more important given we are likely to be wanting to use diagnostic imaging prior to starting any such treatment.

   We agree with Dr. Park, and we have included this in the Discussion (page 12): “A limitation of the present study was the medication status of the patients. Although we did not find any significant effects of antipsychotic drug dose in our sample, there is some evidence of modulatory effects of psychoactive drugs on brain activation as antipsychotic and lithium treatment affect frontal activation (58,59) and antipsychotic medication has been linked to functional and structural changes, particularly in prefrontal areas and the striatum (59-61). If present, such confounding may result in increased brain function differences between patients and controls, and also between schizophrenia and bipolar patients, as the latter are less likely to require long-term antipsychotic treatment. For classification, this medication effect could result in increased separation between groups and therefore increased classification accuracy than would be the case in unmedicated samples. Replication of our findings in patients who are medication-free is thus necessary to exclude these potentially confounding effects, particularly as any diagnostic tool would be most useful prior to the initiation of medication.”