Reviewer's report
Title: Machine Learning Classification of First-Episode Schizophrenia Spectrum Disorders and Controls using Whole Brain White Matter Fractional Anisotropy

Version: 0 Date: 24 Nov 2017
Reviewer: Raymond Salvador

Reviewer's report:
Please include all comments for the authors in this box rather than uploading your report as an attachment. Please only upload as attachments annotated versions of manuscripts, graphs, supporting materials or other aspects of your report which cannot be included in a text format. Please overwrite this text when adding your comments to the authors.

Dear Editors of BMC Psychiatry,

It has been a pleasure to review the work by Mikolas et al. The authors apply a linear support vector machine (i.e. a support vector classifier [SVC]) to evaluate the discriminative power of this methodology to differentiate between first episode patients and controls. Here there are my comments on their work:

1.- Although statistically significant, the discriminant power found between both groups is rather low. This may be a realistic estimate but there are some aspects which could have optimized the strength of the classifier:

A) As explained by the authors, to avoid over-fitting in situations where the number of variables p is clearly higher than the number of samples N (p >> N situation) the SVC requires fitting a regularizing parameter C. In the text, the authors argue that results are rather insensitive to changes in C and apply C = 1. However, my personal experience with SVC and MRI data is that performance is highly dependent on such value. Indeed, to achieve competitive results an exhaustive search should be carried out on potential C values.
Of course, this will require the further division of the training data on training-validation subsets (in a way that models with all C values are generated in the training subset and validated in the validation subset. Then, the C value delivering the best validation is applied on the test data (which has not been used yet).

B) The authors say that both N = 77 samples are matched for gender and age. Still, if gender and sex have some effect on FA values (something quite likely) their effect be present in the images as noise and will lower the performance of the classifier. To avoid this, the most sensible approach is to fit a model on the skeleton FA as a function of gender and age, prior to the classification, and to classify the residuals of such a fit. Before the classifier is applied to the test data, residuals from the test data should be obtained by using the model with the parameters generated by the training data (of course, test data can not be used for the fitting of the model).

2.- In order to classify the individuals in one of the two classes the SVC delivers a score (akin to a probability) that quantifies its affinity to the target class (I guess the FE). This score can be used to evaluate, in the most direct way, the potential relation between the classifier and the covariates (i.e. the amount of medication and the PANSS scores). I think this approach should be used instead of those presented in the manuscript.

3.- On the other hand, although it is fancier to think of white matter abnormalities feeding the classifying algorithm, there is also the possibility of other less elegant factors taking that role. One of them is differential movement between groups. Since Diffusion images are taken in a sequential way, they allow for a quantification of movement during the acquisition. Please check if there are significant differences in movement levels between both groups.

4.- Finally, it should be acknowledged that, although classifying FE and controls is closer to the real setting than classifying chronics and controls, it is still different from the psychiatric unit situation (where individuals are coming with psychiatric symptoms but only some of them will develop schizophrenia). In the current study, there is no guarantee that the healthy sample is comparable to such non-schizophrenic individuals.
Are the methods appropriate and well described?
If not, please specify what is required in your comments to the authors.
Yes

Does the work include the necessary controls?
If not, please specify which controls are required in your comments to the authors.
Yes

Are the conclusions drawn adequately supported by the data shown?
If not, please explain in your comments to the authors.
Yes

Are you able to assess any statistics in the manuscript or would you recommend an additional statistical review?
If an additional statistical review is recommended, please specify what aspects require further assessment in your comments to the editors.
I am able to assess the statistics

Quality of written English
Please indicate the quality of language in the manuscript:
Acceptable

Declaration of competing interests
Please complete a declaration of competing interests, considering the following questions:

1. Have you in the past five years received reimbursements, fees, funding, or salary from an organisation that may in any way gain or lose financially from the publication of this manuscript, either now or in the future?
2. Do you hold any stocks or shares in an organisation that may in any way gain or lose financially from the publication of this manuscript, either now or in the future?
3. Do you hold or are you currently applying for any patents relating to the content of the manuscript?
4. Have you received reimbursements, fees, funding, or salary from an organization that holds or has applied for patents relating to the content of the manuscript?
5. Do you have any other financial competing interests?
6. Do you have any non-financial competing interests in relation to this paper?
If you can answer no to all of the above, write 'I declare that I have no competing interests' below. If your reply is yes to any, please give details below.

I declare that I have no competing interests

I agree to the open peer review policy of the journal. I understand that my name will be included on my report to the authors and, if the manuscript is accepted for publication, my named report including any attachments I upload will be posted on the website along with the authors' responses. I agree for my report to be made available under an Open Access Creative Commons CC-BY license (http://creativecommons.org/licenses/by/4.0/). I understand that any comments which I do not wish to be included in my named report can be included as confidential comments to the editors, which will not be published.

I agree to the open peer review policy of the journal