Author’s response to reviews

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

Authors:
Pavol Mikolas (pavol.mikolas@gmail.com)
Jaroslav Hlinka (jaroslav.hlinka@nudz.cz)
Antonin Skoch (ansk@medicon.cz)
Pitra Zbynek (z.pitra@gmail.com)
Thomas Frodl (thomas.frodl@med.ovgu.de)
Filip Spaniel (filip.spaniel@nudz.cz)
Tomas Hajek (tomas.hajek@dal.ca)

Version: 1 Date: 20 Dec 2017

Author’s response to reviews:

Reviewer reports:

1. Gabriele Sachs (Reviewer 1):

The manuscript "Machine Learning Classification of First-Episode Schizophrenia Spectrum Disorders and Controls using Whole Brain White Matter Fractional Anisotropy" deals with the important issue of applying machine learning to MRI data to differentiate participants with a first episode of schizophrenia-spectrum disorder (FES) from healthy controls based on diffusion tensor imaging data.

In this study the authors found that machine learning applied to whole brain maps differentiated patients with FES and healthy controls with above change accuracy of 62.34%. The white matter
regions, which contributed to the correct identification of participants with FES, showed lower fractional anisotropy in patients relative to controls.

It is important to mention that in this study the intentions of the authors were primarily clinical. A simple, standardized approach potentially suitable for clinical application was used (an "out of the box" approach).

A particular strength of the study presented in this manuscript is that the method is accurate, relevant literature is included. The authors discussed limitations appropriately.

There are minimal concerns and shortcomings:

1.1 Why did the author not include DSM-IV criteria for the recruitment?

We used a standardized diagnostic instrument, the Mini-International Neuropsychiatric Interview (MINI) [1], which provides both ICD and DSM diagnoses. As the study was performed in Europe, where ICD-10 is the official diagnostic system used in the clinical as well as research setting, we utilized the ICD diagnoses.

1.2 The authors mentioned that participants were hospitalized shortly after developing symptoms, some of them did not meet the criteria for schizophrenia. The patients received the working diagnosis of acute and transient psychotic disorder. It is not clear if these persons were included in the study. How many persons got this diagnosis?

We were primarily interested in subjects early in the course of illness, in order to limit the effects of confounding variables (previous psychotic episodes, exposure to medications), which could alter classification accuracy. Consequently, as many of the participants were hospitalized shortly
after developing symptoms, some of them did not meet the duration criteria for schizophrenia at the time of scanning. Specifically, there were 46 patients with the diagnosis of schizophrenia (F20.0) and 31 patients with the diagnosis of acute and transient psychotic disorders (F23.x), which is congruent with DSM-IV defined brief psychotic disorder. We expanded the Methods and Results section to clarify this and included information about the proportion of diagnoses in our sample (page 3, section 2.1; page 7, table 1).

1.3 The patients were treated with antipsychotics. The authors should mention the medication that patients were taking at the time of scanning. Were there differences in the results between typical and atypical antipsychotics?

The overwhelming majority of patients were on atypical antipsychotics at the time of scanning. Only 4 participants were treated with typical antipsychotics, so comparison between atypical and typical antipsychotics was not possible. The detailed numbers of patients on each medication was included in the Methods section (page 4, section 2.1) as follows: olanzapine 29, risperidone 24, quetiapine 5, amisulprid 3, aripiprazole 4, clozapine 2, ziprasidone 1, haloperidol 3, flupenthixol 1, medication naive 1, n/a 4.

1.4 How were the results with regard to the duration of illness?

There was no association between the duration of untreated psychosis (r(74)=0.113, p=0.331) or the duration of illness (r(72)=0.70, p=0.56) and the classification accuracy. It is relevant to note, that already participants in their first episode demonstrated significantly lower FA relative to healthy controls. This suggests that the classifier was based on the trait marker of FES.

2. Raymond Salvador (Reviewer 2):
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:

2.1 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).
Thank you for suggesting this potential improvement. Indeed, the C parameter optimization is a relevant and highly discussed issue. It is important to mention that in this study our goal was primarily geared towards potential clinical utility. Thus, we used a simple, standardized approach potentially suitable for clinical application (an "out of the box" approach).

Any optimization of the C parameter would make it more difficult to compare the results to other studies, would introduce methodological heterogeneity, might increase the risk of overfitting and would be incongruent with potentially clinical use. Thus in keeping with previous studies and with our main goal, we used C=1 [3,4].

Also, as you correctly point out, the hyper parameter optimization would require a nested cross validation with an inner loop for parameter optimization and an outer loop for testing of the model. Such approach would result in smaller effective training set sizes and training test size is critical for performance of the model. In addition, with 77 FES participants, the number of folds that could be used would be relatively small, thus leading to wide confidence intervals. Also, while model performance can be highly sensitive to hyperparameter choices for some techniques, it is much less important for the SVM. So on balance we decided against using hyperparameter optimization.

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).
Thank you for a relevant comment on one of the essential issues in machine learning analyses. Whereas some authors propose the removal of covariates [5] or including the covariates in the model itself [6], matching has been shown to perform at least equally well [7]. We used the one-to-one matching with regards to the most relevant covariates which are known to affect the FA - age and sex in our previous studies [8,9]. The classification itself was performed in a leave-two-out manner, so that the testing set contained always one patient and one control of the same age and sex. Therefore the classification itself was always performed on comparable participants, who did not differ in relevant demographic variables, but only differed in the presence or absence of psychiatric diagnosis.

There are also some specific FA related issues, which speak against the á priori removal of covariates. Firstly, the association between DTI measures and age in adults is likely non-linear [10,11]. Secondly, the association between DTI measures and age or sex likely differs between regions/tracts [10,12,13]. Thirdly, the sample size is such that it would not allow us to get a precise model of association between DTI and age or sex, which may furthermore differ between the FES and controls. Furthermore, the requirement to keep the testing and training sets separate and use a nested cross validation scheme would result in sample size reduction and even lower precision of estimating the association between demographic variables and DTI measures. In addition, this approach, which would require a fitting of a new regression model at each iteration, would go against the main logic of this study, which was to use a simple, standardized approach potentially suitable for clinical application (an "out of the box" approach).

While we agree that methodological refinements are very much needed, our goals here were not towards improvements of methodology, but rather towards applying currently standard methods.

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.

Thank you for your comment. We believe that it concerns the procedure proposed by John Platt (known as Platt scaling, [14]), that transforms the outputs of a classification model into a probability distribution over classes. Based on your suggestion, we performed this procedure and calculated the association between clinical variables and the probabilistic estimates of group membership derived from the SVM classification. There was no association between the probabilistic estimates of the prediction function value and medication or symptoms: CPZ $r(74)=0.09$, $p=0.42$; PANSS Total $r(75)=0.13$, $p=0.25$; PANSS Positive $r(75)=0.072$, $p=0.54$; PANSS Negative $r(75)=0.13$, $p=0.25$; PANSS General $r(75)=0.12$, $p=0.29$).

We added this information into the revised Methods and Results sections (page 6, section 2.5; page 7, section 3.2).

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.

There was no difference in the mean dislocation parameter between the two groups ($t(152)=0.711$, $p=0.727$).

We added this information into the revised Methods section (page 4, section 2.2).
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.

Yes, we agree that in a clinical inpatient setting, where all patients present with marked symptoms, it is more relevant to differentiate between psychiatric diagnoses than between patients and controls. We included the following section into limitations:

In a clinical inpatient setting, where all patients present with marked symptoms, it is more relevant to differentiate between psychiatric diagnoses than between patients and controls. We did not recruit a comparison group of participants with for example first episode of mania. Only few studies have addressed differential diagnosis between major classes of psychosis [5,17]. More future studies should focus on this important and clinically relevant issue, specifically among participants early in the course of illness.

References


