Contents lists available at ScienceDirect

Journal of Affective Disorders

journal homepage: www.elsevier.com/locate/jad

Natural speech algorithm applied to baseline interview data can predict which patients will respond to psilocybin for treatment-resistant depression

Facundo Carrillo[a,b,*], Mariano Sigman[c], Diego Fernández Slezak[a,b], Philip Ashton[d], Lily Fitzgerald[d], Jack Stroud[d], David J. Nutt[d], Robin L. Carhart-Harris[d]

[a] Applied Artificial Intelligence Lab, Computer Science Department, School of Science, Buenos Aires University, CONICET, Buenos Aires 1428, Argentina
[b] CONICET-Universidad de Buenos Aires, Instituto de Investigación en Ciencias de la Computación (ICC), Buenos Aires, Argentina
[c] Integrative Neuroscience Lab, Universidad Torcuato Di Tella, CONICET, Buenos Aires 1428, Argentina
[d] Psychedelic Research Group, Centre for Psychiatry, Dept of Medicine, Imperial College London, London, UK

ARTICLE INFO

Keywords:
Psilocybin treatment
Natural speech analysis
Machine learning
Depression
Treatment-resistant depression
Computational psychiatry
Predict therapeutic effectiveness

ABSTRACT

Background: Natural speech analytics has seen some improvements over recent years, and this has opened a window for objective and quantitative diagnosis in psychiatry. Here, we used a machine learning algorithm applied to natural speech to ask whether language properties measured before psilocybin for treatment-resistant can predict for which patients it will be effective and for which it will not.
Methods: A baseline autobiographical memory interview was conducted and transcribed. Patients with treatment-resistant depression received 2 doses of psilocybin, 10 mg and 25 mg, 7 days apart. Psychological support was provided before, during and after all dosing sessions. Quantitative speech measures were applied to the interview data from 17 patients and 18 untreated age-matched healthy control subjects. A machine learning algorithm was used to classify between controls and patients and predict treatment response.
Results: Speech analytics and machine learning successfully differentiated depressed patients from healthy controls and identified treatment responders from non-responders with a significant level of 85% of accuracy (75% precision).
Conclusions: Automatic natural language analysis was used to predict effective response to treatment with psilocybin, suggesting that these tools offer a highly cost-effective facility for screening individuals for treatment suitability and sensitivity.
Limitations: The sample size was small and replication is required to strengthen inferences on these results.

1. Introduction

Quantitative analyses of natural speech have undergone significant advances in recent years (Mikolov et al., 2013; Michel et al., 2011; Carrillo et al., 2015; Cambria and White, 2014)and are beginning to be applied in psychiatry (Wang and Krystal, 2014; Wiecki et al., 2015; Mota et al., 2016; Carrillo et al., 2014; Huys et al., 2011; Mundt et al., 2012). For example, automatic analysis of speech incoherence has been used as a biomarker of schizophrenia; in a proof-of-concept experiment with a small sample size, a machine learning algorithm predicted conversion to psychosis in 'at-risk' individuals with a 100% accuracy (Bedi et al., 2015).

In mood disorders, a recently developed measure of emotion in speech was found to accurately sort between bipolar patients and control subjects (Carrillo et al., 2016) and related tools have proved effective in identifying depression in interview-based speech (Pestian et al., 2008) and social media-based text (De Choudhury et al., 2013, 2013). These studies highlight the power of natural language analytics to diagnose and prognose mental illness and its response to treatment.

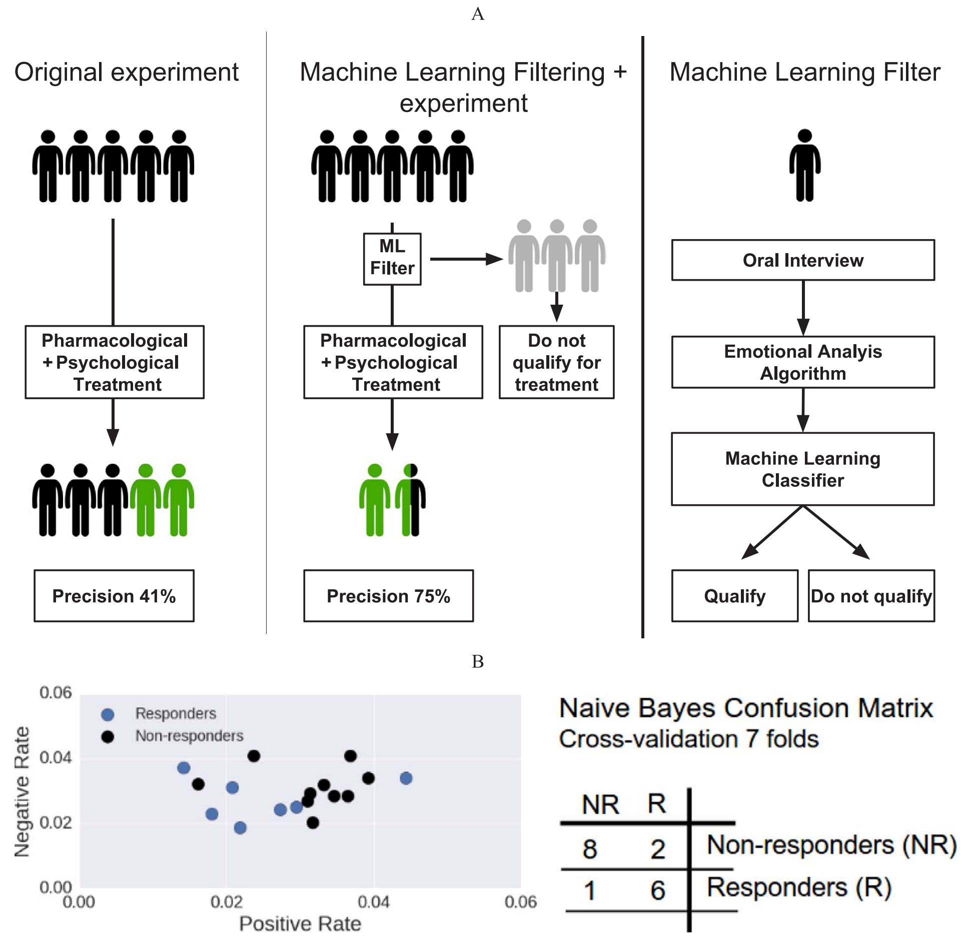
In the present study, we sought to build on this work by testing whether natural speech analytics combined with machine learning could predict clinical responses to psilocybin in patients with treatment-resistant depression (TRD), defined here as failure of at least two different antidepressants of differing pharmacology, within the same depressive episode. Psilocybin is a serotonin 2A receptor agonist and classic psychedelic drug that is currently showing promise for the treatment of a range of psychiatric conditions, including depression (Carhart-Harris and Goodwin, 2017).

2. Methods

This trial received a favorable opinion from NRES London-West

[*] Corresponding author at: Applied Artificial Intelligence Lab, Computer Science Department, School of Science, Buenos Aires University, CONICET, Buenos Aires 1428, Argentina.
E-mail address: fcarrillo@dc.uba.ar (F. Carrillo).

https://doi.org/10.1016/j.jad.2018.01.006
Received 8 June 2017; Received in revised form 5 December 2017; Accepted 22 January 2018


A
Original experiment
Pharmacological + Psychological Treatment
Precision 41%
Machine Learning Filtering + experiment
ML Filter
Pharmacological + Psychological Treatment
Do not qualify for treatment
Precision 75%
Machine Learning Filter
Oral Interview
Emotional Analyis Algorithm
Machine Learning Classifier
Qualify
Do not qualify
B
Responders
Non-responders
Negative Rate
Positive Rate
0.00
0.02
0.04
0.06
Naive Bayes Confusion Matrix
Cross-validation 7 folds

NR	R	
8	2	Non-responders (NR)
1	6	Responders (R)

Fig. 1. Panel A shows: 1) The original protocol where subjects take the treatment and only 41% response well. 2) The modification of the protocol where before taking the treatment subjects are evaluated by the Machine Learning Filter. Some subjects pass the Filter and start with the treatment and others leave the protocol. This modification increase the precision of the treatment to 75% of precision. 3) A description of the Machine Learning Filter with its 3 steps: Oral interview, Emotional Analysis algorithm and Machine Learning Classifier. Panel B shows: 1) The relation by subject of the AVG P and AVG N, black dots represent subjects which the treatment originally did not work, blue ones correspond to subjects that the original treatment worked. 2) The confusion matrix of the Naive Bayes Classifier with 7 folds cross-validation schema. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

London, was sponsored by Imperial College London, and was carried out in accordance with Good Clinical Practice Guidelines. It was an open-label design in which patients with TRD received two doses of psilocybin (10 mg and 25 mg) one week apart. The autobiographical memory test (AMT) (Williams and Scott, 1988) was performed by patients (n = 17) and age and sex matched matched controls (n = 18), who were recruited separately. For more details on the design and procedures of the main trial, see Carhart-Harris et al. (2016).

The AMT is a structured interview in which participants are asked to provide specific autobiographical memories in response to specific cue words. For example, the cue word “newspaper” may be read to a participant, who might then reply “When I was about 8 years old, I remember a dog biting my arm as I tried to pick-up a newspaper” etc. Two different but balanced versions of the task, with a different set of word cues, were completed across the sample but there were no between-group differences in the completed versions. Patients completed their AMT interviews approximately 2 weeks prior to receiving their first dose of psilocybin and the matched controls did theirs at their convenience. All AMT interviews were audio recorded and transcribed (by L.F, J.S and P.A). More details on study procedures can be found in the online supplement and the main outcomes of the trial are published elsewhere (Carhart-Harris et al., 2016).

The sample comprised of 17 patients (mean age = 44.59 (SD = 10.97), 5 females) and 18 healthy control subjects (mean age = 36.44 (SD = 17.23), 7 females). The primary outcome measure, the Quick Inventory of Depressive Symptoms (QIDS-16), was rated by both sets of participants at baseline, and patients rated it again 5 weeks after the 25 mg psilocybin dose. Treatment response was defined as ≥ 50% reduction in QIDS scores at 5 weeks. There were 7 treatment responders and 10 non responders at 5 weeks (41%).

2.1. *Analysis on subject speech*

Emotional Analysis (EA) (Carrillo et al., 2016) is an automated algorithm for quantifying the emotional content of spoken or written text (Esuli and Sebastiani, 2007). As it was employed here, positive and negative emotional sentiment scores were assigned to each word in the transcribed AMT interviews. EM scores are decimal values between 0 and 1. We defined the average positivity (AVG P) of a text as the mean positive score over all words in the text, and did the same for its average negativity (AVG N). For example, the sentence “It is a happy day, but I am sad” yields an AVG P value of 0.104 and an AVG N value of 0.041. In this example, positive and negative scores (respectively) are: “it” (0,0), “a” (0,0), “happy” (0.5,0), “day” (0.125,0), “I” (0,0) and “sad” (0,0.25).

2.2. *Machine learning*

We used a Gaussian Naive Bayes classifier (John and Langley, 1995) and then represented every subject with 2 features: AVG P and AVG N. This approach was subsequently used to classify patients versus controls and separately, responders versus non-responders. A cross-validation (cv) schema was used to test for the significance of classification accuracy (see online supplement for details).

3. Results

Before we addressed our main question, we asked whether our method can distinguish between controls and patients. A significant between-group difference was found in the rate of positive words used in participants’ AMT interview responses, with patients using significant fewer positive words: controls AVG P = 0.0532 ± 0.013 and patients AVG P = 0.0384 ± 0.011 (*t*-test p = 0.0011). The AVG N did not differ significantly between both groups (p = 0.4). Using a machine learning classifier, with a 7 folds cross-validation scheme, to identify patients (versus controls) based on a combination of AVG P and N values, we obtained a mean accuracy of 82.85%, (precision = 0.82, recall = 0.82, sensitivity = 0.82, specificity = 0.83). A control experiment using random permutation testing (1000 trials) (see online supplement), confirmed that this accuracy was significantly greater than chance ($p < 0.05$).

Next, we tackled the main study question: whether pre-treatment speech could be predictive of subsequent treatment success. To do this, we employed the same machine learning approach as described above to identify responders from non-responders. AVG P and N values were not significantly different for responders (P: 0.0334 ± 0.0132, N: 0.0368 ± 0.0088) and non-responders (P: 0.0418 ± 0.0091, N:0.041 ± 0.0082); however, using the same input formula as above, we were able to predict treatment response with an above chance accuracy of 85% (precision 0.75 with a 7 folds cross-validation scheme and Gaussian Naive Bayes as classifier algorithm).

As can be discerned from in Fig. 1, AVG P, was the most sensitive variable for distinguishing patients from controls, and for predicting responder versus non-responders. On closer inspection of the data, it was found that responders used fewer emotional words at baseline (and fewer positive words especially) than non-responders, potentially reflecting a greater capacity for change in the responders that rendered them particularly sensitive to this treatment.

Permutation testing revealed that the 85% accuracy was in the upper 97-percentile of the distribution and therefore significantly greater than chance ($p < 0.05$).

4. Discussion

In the present study, natural speech analytics combined with machine learning was able to differentiate depressed patients from healthy controls and predict responders versus non-responders in a clinical trial of psilocybin for treatment-resistant depression. The AMT interviews that produced the data on which these analyses were performed took little longer than 10 min to perform, yet were able to identify depression from health and predict treatment response with a significant level of precision. Psilocybin, like other psychedelics, has idiosyncratic acute effects, and the quality of the acute drug experience has been found to be strongly predictive of subsequent long-term clinical outcomes (Carhart-Harris and Goodwin, 2017). Psilocybin is currently being studied as a treatment for a range of different psychiatric disorders, and particularly depression (Carhart-Harris et al., 2016). As well as providing further support for the diagnostic potential of natural speech analytics, the present results – combined with the near-to-zero application cost of this software methods – suggest that these tools offer a highly cost-effective facility for screening individuals for treatment suitability and sensitivity. Future work may test the specificity of the highlighted relationships and whether they generalize to other interventions and outcomes.

Acknowledgments

Robin L. Carhart-Harris is supported by the Alex Mosley Charitable Trust and that the UK MRC (MR/J00460X/1) funded the trial. Diego Fernández Slezak is sponsored by the Microsoft Faculty Fellowship 2014. MS is sponsored by James McDonnell Foundation 21st Century Science Initiative in Understanding Human Cognition.

Appendix A. Supplementary data

Supplementary data associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.jad.2018.01.006.

References

Bedi, G., et al., 2015. Automated analysis of free speech predicts psychosis onset in high-risk youths. NPJ Schizophr. 1, 15030.

Cambria, E., White, B., 2014. Jumping nlp curves: a review of natural language processing research. IEEE Comput. Intell. Mag. 9, 48–57.

Carhart-Harris, R.L., Goodwin, G.M., 2017. The therapeutic potential of psychedelic drugs: past, present and future. Neuropsychopharmacology.

Carhart-Harris, R.L., et al., 2016. Psilocybin with psychological support for treatment-resistant depression: an open-label feasibility study. Lancet Psychiatry 3, 619–627.

Carrillo, F., et al. 2014. Automated speech analysis for psychosis evaluation. In: International Workshop on Machine Learning and Interpretation in Neuroimaging, 31-39 (Springer International Publishing).

Carrillo, F., Cecchi, G.A., Sigman, M., Slezak, D.F., 2015. Fast distributed dynamics of semantic networks via social media. Comput. Intell. Neurosci. 2015 (50).

Carrillo, F. et al. 2016. Emotional intensity analysis in bipolar subjects. arXiv preprint arXiv:1606.02231.

De Choudhury, M., Gamon, M., Counts, S., Horvitz, E., 2013. Predicting depression via social media. ICWSM 2.

De Choudhury, M., Counts, S., Horvitz, E., 2013. Social media as a measurement tool of depression in populations. In: Proceedings of the 5th Annual ACM Web Science Conference, 47-56 (ACM).

Esuli, A., Sebastiani, F., 2007. Sentiwordnet: a high-coverage lexical resource for opinion mining. Evaluation 1–26.

Huys, Q.J., Moutoussis, M., Williams, J., 2011. Are computational models of any use to psychiatry? Neural Netw. 24, 544–551.

John, G.H., Langley, P., 1995. Estimating continuous distributions in bayesian classifiers. In: Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence, 338–345 (Morgan Kaufmann Publishers Inc.).

Michel, J.-B., et al., 2011. Quantitative analysis of culture using millions of digitized books. Science 331, 176–182.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J., 2013. Distributed representations of words and phrases and their compositionality. Adv. Neural Inf. Process. Syst. 3111–3119.

Mota, N., Carrillo, F., Slezak, D., Copelli, M., Ribeiro, S., 2016. Characterization of the relationship between semantic and structural language features in psychiatric diagnosis. In: Proceedings of the 50th Asilomar Conference on Signals, Systems and Computers, 2016, 836-838 (IEEE).

Mundt, J.C., Vogel, A.P., Feltner, D.E., Lenderking, W.R., 2012. Vocal acoustic biomarkers of depression severity and treatment response. Biol. Psychiatry 72, 580–587.

Pestian, J.P., Matykiewicz, P., Grupp-Phelan, J., 2008. Using natural language processing to classify suicide notes. In: Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing, 96-97 (Association for Computational Linguistics).

Wang, X.-J., Krystal, J.H., 2014. Computational psychiatry. Neuron 84, 638–654.

Wiecki, T.V., Poland, J., Frank, M.J., 2015. Model-based cognitive neuroscience approaches to computational psychiatry clustering and classification. Clin. Psychol. Sci. 3, 378–399.

Williams, J., Scott, J., 1988. Autobiographical memory in depression. Psychol. Med. 18, 689–695.