NeuroImage: Clinical 4 (2014) 98–111

Contents lists available at ScienceDirect

NeuroImage: Clinical

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

Dissecting psychiatric spectrum disorders by generative embedding☆,☆☆

Kay H. Brodersen [a,b,*], Lorenz Deserno [c,d], Florian Schlagenhauf [c,d], Zhihao Lin [a,b], Will D. Penny [e], Joachim M. Buhmann [b], Klaas E. Stephan [a,e,f]

[a] Translational Neuromodeling Unit (TNU), Institute for Biomedical Engineering, University of Zurich & ETH Zurich, Switzerland
[b] Machine Learning Laboratory, Department of Computer Science, ETH Zurich, Switzerland
[c] Department of Psychiatry and Psychotherapy, Charité-Universitätsmedizin Berlin, Germany
[d] Max Planck Institute for Cognitive and Brain Sciences, Leipzig, Germany
[e] Wellcome Trust Centre for Neuroimaging, University College London, United Kingdom
[f] Laboratory for Social and Neural Systems Research (SNS), University of Zurich, Switzerland

ARTICLE INFO

Article history:
Received 4 July 2013
Received in revised form 6 October 2013
Accepted 7 November 2013
Available online 16 November 2013

Keywords:
Clustering
Clinical validation
Balanced purity
Schizophrenia
Variational Bayes

ABSTRACT

This proof-of-concept study examines the feasibility of defining subgroups in psychiatric spectrum disorders by generative embedding, using dynamical system models which infer neuronal circuit mechanisms from neuroimaging data. To this end, we re-analysed an fMRI dataset of 41 patients diagnosed with schizophrenia and 42 healthy controls performing a numerical *n*-back working-memory task. In our generative-embedding approach, we used parameter estimates from a dynamic causal model (DCM) of a visual–parietal–prefrontal network to define a model-based feature space for the subsequent application of supervised and unsupervised learning techniques. First, using a linear support vector machine for classification, we were able to predict individual diagnostic labels significantly more accurately (78%) from DCM-based effective connectivity estimates than from functional connectivity between (62%) or local activity within the same regions (55%). Second, an unsupervised approach based on variational Bayesian Gaussian mixture modelling provided evidence for two clusters which mapped onto patients and controls with nearly the same accuracy (71%) as the supervised approach. Finally, when restricting the analysis only to the patients, Gaussian mixture modelling suggested the existence of three patient subgroups, each of which was characterised by a different architecture of the visual–parietal–prefrontal working-memory network. Critically, even though this analysis did not have access to information about the patients' clinical symptoms, the three neurophysiologically defined subgroups mapped onto three clinically distinct subgroups, distinguished by significant differences in negative symptom severity, as assessed on the Positive and Negative Syndrome Scale (PANSS). In summary, this study provides a concrete example of how psychiatric spectrum diseases may be split into subgroups that are defined in terms of neurophysiological mechanisms specified by a generative model of network dynamics such as DCM. The results corroborate our previous findings in stroke patients that generative embedding, compared to analyses of more conventional measures such as functional connectivity or regional activity, can significantly enhance both the interpretability and performance of computational approaches to clinical classification.

1. Introduction

Psychiatry has experienced a long-standing and ongoing discussion about the validity of pathophysiological concepts and clinical classification schemes. One central problem is that despite all progress in neuroscience, there has been an almost complete lack of mechanistic insights that would allow for the development of diagnostic tests for detecting pathophysiological mechanisms in individual patients. As a result, with the exception of excluding 'external' causes such as brain lesions or metabolic disturbances (Kapur et al., 2012), psychiatric diagnosis

* Corresponding author at: Translational Neuromodeling Unit (TNU), Institute for Biomedical Engineering, University of Zurich & ETH Zurich, Wilfriedstrasse 6, CH 8032 Zurich, Switzerland. Tel.: +41 76 760 84 08.
E-mail address: brodersen@biomed.ee.ethz.ch (K.H. Brodersen).

2213-1582/$ – see front matter © 2013 The Authors. Published by Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.nicl.2013.11.002

still relies on symptom-based definitions of disease, such as the classifications proposed by the Diagnostic and Statistical Manual Of Mental Disorders (DSM) or the International Classification of Diseases (ICD).

For example, despite high initial hopes, genetic tests have not entered clinical practice so far (Braff and Freedman, 2008; Tansey et al., 2012). This is not only because most diseases appear to be highly polygenetic, with each candidate polymorphism possibly conveying only a modest increase in risk (International Schizophrenia Consortium, 2009) and for more than one disease (Cross-Disorder Group Of the Psychiatric Genomics Consortium, 2013). More importantly, genetic tests are impeded by the presence of strong gene–environment interactions (Caspi and Moffitt, 2006). These interactions mean that even when the genome is identical the influence of different environmental factors can lead to the occurrence of different disease mechanisms and symptoms (Dempster et al., 2011; Petronis et al., 2003).

Beyond genetics, neuroimaging is another discipline which has, so far, struggled to fulfil its promise with regard to establishing practically useful diagnostic tests for psychiatry (cf. Borgwardt et al., 2012). This is despite the fact that over the past few years, neuroimaging has seen a veritable explosion in the application to psychiatric questions.

For example, numerous studies have applied *machine-learning* techniques, such as support vector machine (SVM) classification, to structural or functional magnetic resonance imaging (MRI) data. The majority of previous studies have tried to discriminate patients with a particular DSM/ICD diagnosis from healthy controls, or to disambiguate between patients from different DSM/ICD-defined diseases (see Klöppel et al., 2012, for a recent review of the application of machine-learning methods to neuroimaging data of patients). However, for many psychiatric diseases, diagnosis with respect to DSM/ICD criteria is not the key clinical problem (with some notable exceptions, such as distinguishing between unipolar and bipolar affective psychosis in first-episode patients). Therefore, machine-learning approaches which use diagnostic labels from DSM/ICD for training a classifier applied to neuroimaging data can at best reproduce the presently established diagnostic classification, but using a considerably more expensive and complicated procedure.

Instead, it seems more fruitful to develop statistical techniques for predicting future variables which are important for clinical decision making, e.g., whether a particular patient with mild cognitive impairment will develop Alzheimer's disease within a certain period or not (Davatzikos et al., 2008; Lehmann et al., 2012). One prominent hope is that biological markers derived from neuroimaging procedures may enable more accurate predictions of treatment response or disease trajectory than the behavioural and cognitive symptoms on which current DSM/ICD diagnoses are based.

This approach is logistically considerably more challenging than the attempt of reproducing DSM-based disease definitions since it requires longitudinal studies. Nevertheless, a few recent studies have been able to demonstrate that it may be possible to predict individual treatment response (e.g., Costafreda et al., 2009; Liu et al., 2012; Szeszko et al., 2012) or clinical outcome (e.g., Koutsouleris et al., 2012; Mourao-Miranda et al., 2012; Siegle and Thompson, 2012) from structural or functional MRI data, using multivariate classification. If such a procedure could be established that allowed, with sufficient sensitivity and specificity, for clinically relevant decisions, it might indeed become a cost-effective tool for clinical decision-making. Still, however, any such approach would effectively remain a 'black-box' classifier, providing very limited insights, if any, into disease mechanisms. This is a fundamental limitation, since without mechanistic interpretability no diagnostic procedure can inform a change in disease concepts or guide the development of future therapies.

A potential alternative to black-box classification is to embed classification into a space spanned by the parameters of a generative model which explains how the measured data could have arisen from underlying neurophysiological mechanisms (e.g., synaptic connections between distinct neuronal populations). This is the generative-embedding approach which we recently introduced to neuroimaging (Brodersen et al., 2011a).

In this previous work, we demonstrated that a six-region dynamic causal model (DCM) of the early auditory system during passive speech listening could predict, with near-perfect accuracy (98%), the absence or presence of a 'hidden' (i.e., outside the field of view) lesion in aphasic patients compared to healthy controls. Critically, this model-based classification approach not only significantly outperformed conventional approaches, such as searchlight classification on the raw fMRI data or classification based on functional connectivity between the same regions; more importantly, it also highlighted network mechanisms which distinguished the two groups. In this case, the connections from the right to the left hemisphere were particularly informative for enabling this subject-by-subject classification, suggesting that the remote lesion prominently affected interhemispheric transfer of language information to the dominant hemisphere.

Mechanistically interpretable approaches like generative embedding have potential for significantly enhancing model-based predictions of clinically relevant variables such as outcome or treatment response. However, these approaches are of equal importance for addressing a second fundamental problem in psychiatry: the nature of psychiatric nosology itself, i.e., the disease definitions that determine clinical diagnostics and classification. As described above, DSM defines diseases purely on the basis of symptoms that can be assessed by means of structured interviews. This approach was introduced a few decades ago to ensure the reproducibility of diagnostic statements across clinicians and institutions. However, the consequence of its entirely phenomenological nature is that the resulting disease concepts are completely agnostic about underlying mechanisms. Furthermore, many empirical studies have questioned the clinical validity of this classification scheme, demonstrating problematic predictive validity with regard to treatment and outcome (e.g., Johnstone et al., 1988; Johnstone et al., 1992). It is therefore not surprising that this phenomenological definition of diseases has received substantial criticism, and alternatives are being sought, such as the Research Domain Criteria (RDoC; Insel et al., 2010) which aim to redefine psychiatric diseases based on pathophysiological mechanisms. Key challenges for this endeavour are how such pathophysiological mechanisms can be detected in the individual, and how they are combined to produce a meaningful classification.

We have previously argued that a pathophysiologically informed dissection of psychiatric spectrum diseases, such as schizophrenia, into physiologically defined subgroups should be guided by model-based estimates of synaptic physiology from neuroimaging and electrophysiological data (Stephan, 2004; Stephan et al., 2009). This approach requires modelling techniques which can be applied to non-invasive measures of brain activity in individual patients and which are capable of inferring neurophysiological mechanisms at the circuit level. One such method is dynamic causal modelling (DCM; Friston et al., 2003), a Bayesian framework for inferring neurophysiological mechanisms from neuroimaging data.

Previous electrophysiological studies have demonstrated that DCM can provide valid information on mechanisms which represent potential key dimensions of psychiatric disease, e.g., excitation–inhibition balance, synaptic plasticity by NMDA receptors, or its regulation by neuromodulatory transmitters such as dopamine or acetylcholine (Moran et al., 2011a,b; Schmidt et al., 2012). When applied to fMRI data, DCM allows for less fine-grained representation of physiological mechanisms and is largely restricted to inferring on synaptic coupling between large, undifferentiated neuronal populations. Nevertheless, even this coarse physiological representation has proven useful for distinguishing groups with different cognitive or disease states, such as: the presence vs. absence of a 'hidden' lesion (see above; Brodersen et al., 2011a,b); Parkinson patients on vs. off dopaminergic medication (Rowe et al., 2010); individuals with different types of synaesthesia (Van Leeuwen et al., 2011); patients suffering from depression vs.