Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Behavioral Sciences

What big data can do for treatment in psychiatry

Claire M .Gillan[1,2,3] and Robert Whelan[1,2,3]

Treatments for psychiatric disorders are only as effective as the precision with which we administer them. We have treatments that work; we just cannot always accurately predict who they are going to work for and why. In this article, we discuss how big data can help identify robust, reproducible and generalizable predictors of treatment response in psychiatry. Specifically, we focus on how machine-learning approaches can facilitate a move beyond discovery studies and toward model validation. We will highlight some recent exemplary studies in this area, describe how one can assess the merits of studies reporting treatment biomarkers, and discuss what we consider to be best practice for prediction research in psychiatry.

Addresses
[1] School of Psychology, Trinity College Dublin, Dublin 2, Ireland
[2] Trinity College Institute of Neuroscience, Trinity College Dublin, Dublin 2, Ireland
[3] Global Brain Health Institute, Trinity College Dublin, Dublin 2, Ireland

Corresponding author: .Gillan, Claire M (claire.gillan@tcd.ie)

Current Opinion in Behavioral Sciences 2017, **18**:34–42

This review comes from a themed issue on **Big data in the behavioral sciences**

Edited by **Michal Kosinski** and **Tara Behrend**

http://dx.doi.org/10.1016/j.cobeha.2017.07.003

Introduction

Treatment response in psychiatry is highly variable. Most patients do not remit following their first course of treatment [1], and in practice multiple therapies are often trialed (and failed) before finding one that works. This means that patients may endure long periods of untreated symptoms before finding something effective — if at all. Unfortunately, we do not possess prognostic tools that can accurately predict a patient's response to a specific treatment. A key goal of modern psychiatry research is to remedy this; to identify the right treatment for each patient, first time around.

Although there are myriad existing research studies that purport to distinguish for example antidepressant treatment responders from non-responders using cognitive, neuroimaging, neurochemical, demographic or clinical measurements, we remain without a single viable biomarker for treatment prediction in psychiatry. Why is this?

Issues surrounding reproducibility in science have never been more visible [2,3]. It is being increasingly recognized that there is an over-reliance on null hypothesis significance testing, and a related focus on *p*-values [4,5]. Although social psychology has been the sacrificial lamb for the reproducibility debate in psychology thus far, neuroimaging has been subjected to similar critiques [6], where statistical power and researcher degrees of freedom have been identified as key problem areas. Most neuroimaging studies test small samples (typically <50), often having many more variables (e.g., tens of thousands of voxels) than subjects. This makes finding spurious results (i.e., overfitting) highly likely, because the ratio of cases to predictors for ordinary least-squares is directly related to the overestimation of a model's performance (see [7]).

If held to the same standard, biomarker research studies in psychiatry would fare no better — failing the most crucial test: reproducibility. Although numerous brain measurements have been identified as potential biomarkers of treatment response (see Jollans and Whelan, 2016 for a review [8]), there is almost no consistency across studies. This is because these studies, with a few exceptions described later, do not incorporate adequate external validation steps — that is, testing the ability of a putative predictor to classify *unseen* data.

This paper seeks to emphasize the importance of big data and robust statistical methodologies in treatment prediction research. How, in the absence of pre-registration, methods like internal cross-validation and the use of 'hold-out data' or external data (both are considered unseen data, but the latter is recruited independently of the training data for example, taken from another study) are crucial tools for prediction research in psychiatry, and particularly for studies involving neuroimaging [9,10]. We will start by introducing the concept of machine-learning and discussing how it complements theory-driven approaches to understanding treatment response. Then we will describe recent exemplars that have successfully applied machine learning to treatment prediction in psychiatry. We will close with some best-practice guidelines for research in this area and some recommendations for collaborative research strategies (Box 1, Figure 1). These approaches align nicely with initiatives like that of the National Institute of Mental Health (Research Domain Criteria (RDoC)), which aim to establish biologically-grounded alternatives to our current system of psychiatric diagnostic classification.

Box 1 Best practice for treatment prediction research

Large samples. Other considerations being equal, bigger sample sizes improve model reliability by reducing the tendency to overfit [10]. It is difficult for the model to fit to random noise in the training data as the sample size increases. Larger effect sizes protect against overfitting, while high-dimensional feature spaces (e.g. voxels in neuroimaging) promote it.

Validation. The performance of any model should be referenced against unseen data. One approach is to split a dataset into three parts (nomenclature of Hasti, Tibshirani & Friedman [12] but note terminology use can differ): a training set, and smaller validation and test ('hold-out') sets. The training set is used for model fitting, the validation set is used to measure generalization error of the model, for example using nested cross-validation. The training and validation data can also be used for model selection, for example assessing how well competing models perform on validation data by varying model parameters (e.g., regularization constants) or type of algorithm (e.g., Elastic Net or SVM). In this case, a test set is critical, otherwise overfitting due to 'researcher degrees of freedom' will occur (e.g., choosing the best performing algorithm). The 'hold-out' test set is one that has been kept entirely separate from the rest of the data, and is only used to test the performance of one final selected model. Where possible, the test set would come from a different sample with similar characteristics (an external test), which is an even stronger test of generalization.

Appropriate metrics. No one metric can capture model performance because factors such as differences in base rates of response to treatment affect the interpretation of these metrics. Thus, a range of metrics are necessary and should include sensitivity, specificity, positive predictive value, negative predictive value (see Figure 2 for detailed explanation).

Regret. Psychiatric treatments have a range of side-effects and financial burdens, and therefore some misclassifications are worse than others in terms of patient safety and expense. Machine learning algorithms that include 'regret' by incorporating a different cost for error types are likely to be useful for treatment response prediction. For example, we may want to predict response to treatment with the goal of stopping ineffective treatment early for financial reasons. False positives (predicting no improvement when the patient does improve) is more risky in terms of patient health than false negatives (predicting improvement when the patient does not improve). Thus, a classifier could be constructed so that, during training, false positives are twice as costly as false negatives [48].

Interpretability. Woo and colleagues make [49] three excellent suggestions improving interpretability of models. Briefly, the models should firstly, be summarized (e.g., by applying a data reduction method) to present the most predictive features secondly, be evaluated for neuroscientific plausibility (e.g., is it concordant with known pathology) and thirdly, consider the potential for confounding variables to contribute to the model.

Open Science. Data and models should be shared to firstly, facilitate comparisons with previous and future models and secondly, to provide datasets for external model validation.

Data-driven and theory-driven approaches

Machine-learning (essentially synonymous with 'data-mining' or 'statistical learning') refers to a class of approaches that focus on prediction rather than interpretation or mechanism. Typically, an outcome variable such as responder/non-responder status is used to train an algorithm to identify some combination of features (e.g., self-report, demographic, cognitive or brain data) that are associated with the outcome. This type of question — responder vs. non-responder as the outcome variable — is often treated as a classification problem in machine learning. But the outcome variable can also be continuous, and in these cases a regression analysis is often used. From a machine-learning perspective, the same principles apply to both regression and classification questions. Models must balance the need to accommodate the complexity of the data (i.e., to be flexible) and the need for interpretability (see [11], section 2.1.3 for a more detailed exposition). For example, linear regression is inflexible because only linear relationships are allowed, as a result the output is easy to interpret — the outcome is a weighted linear combination of the features (e.g., younger people respond better to a specific treatment). By contrast, a support vector machine or random forest approach yields more accurate predictions, but interpretation is more difficult [12]. The choice of method depends on the goal of the analysis, as does the choice of metric for quantifying model performance. For example, the model can be optimized for real-world implementation for example using terms like 'regret' which take into account the fact that in the context of patient-care, some misclassification errors are worse than others (e.g., false positive errors might be worse than false negative errors in some situations; Box 1).

Machine learning approaches are often contrasted with theory-driven approaches, such as those promoted by the computational psychiatry movement, which endeavor to explain psychiatric phenomena in terms of detailed models of brain function [13,14]. This theory-driven strategy might help improve treatment outcomes in one of two ways. First, it is thought that by linking clinical symptoms directly to theory-driven computational models of neural processes, new treatments could be designed to more precisely and effectively target these neural processes [15,16]. Alternatively, it is possible that the heterogeneity of response to existing treatments within diagnostic categories might be resolved if we redefine those categories based on commonalities in well-defined neurobiological processes rather than symptomatology. Enthusiasm for the latter approach has been borne out in work that has found new ways of parsing symptoms in ways that link more closely to neural [17] and cognitive [18,19] processes than existing diagnostic categories.

The computational psychiatry approach is appealing, yet it remains to be seen how these insights will transfer to the clinic. Aside from issues of scalability and implementation in terms of both reach and cost-effectiveness of its mainstay tools, like functional imaging, seeking a near perfect computational characterization of the brain processes linked to a given clinical symptom cluster might be a dead end, because there is no guarantee that this will produce insights for improving treatments. This is in part because, much like in general medicine, different underlying causes can produce similar symptoms (e.g., jaundice