Evaluation of machine learning models in psychiatry

Dinga, R.

2020

document version
Publisher's PDF, also known as Version of record

SUMMARY OF FINDINGS

This thesis evaluated specific problems of statistical validation of machine learning models, identified shortcomings of the current practice, and provided solutions to avoid them. The general introduction section introduced machine learning methods and their application to psychiatry, together with specific problems of their evaluation.

Chapter 2 contains the development of a machine learning predictive model of two-year depression remission and chronicity in the NESDA dataset on 804 subjects with a depressive disorder. Subjects were assessed based on 81 clinical, psychological, and biological variables, which were used to create a predictive model using an elastic net logistic regression. The focus of this chapter was to evaluate which of the wide range of variables were important for prediction. This was assessed using a stability selection approach. Stability selection is the probability that a variable will be selected in a machine learning model, and it provides a family-wise type I error control for the selected variables(Meinshausen and Bühlmann 2010). Only one variable, IDS score, was statistically significant, and no other variable substantially improved model predictions. This finding is in line with results reported by (Chekroud et al. 2016). For a machine learning prediction of remission of depressive symptoms followed by 12-week citalopram treatment. Their model selected a shortened version of the IDS questionnaire as the most important predictor form a range of sociodemographic and clinical features. Our findings contrast findings from other cross-sectional studies that showed group level association vitamin D (Milaneschi et al. 2014) and cortisol (Vreeburg et al. 2013) to depression chronicity. This demonstrates that group-level associations, as discovered, does not necessarily translate to better machine learning predictions.

Chapter 3 contains a methodological replication of a prominent study identifying biotypes of depression (Drysdale et al. 2017). We identified shortcomings of the methods used in this study, questioned the validity of some of the results, and provided recommendations for future studies. In the original study, authors performed canonical correlation analysis (CCA) between resting-state fMRI features and clinical symptoms of depression in a sample of currently severely depressed, treatment-resistant participants. CCA identified two biological-clinical factors related to anhedonia and anxiety. Next, they used hierarchical clustering to identify four subtypes of depression according to these two factors. These

subtypes had different clinical profiles, including different average response to transcranial magnetic stimulation treatment in an independent dataset.

Our replication was conducted on 187 participants with depression, anxiety, or depression anxiety comorbidity from NESDA and MOTAR datasets. The original analytical pipeline was followed as closely as possible, including finding a low-dimensional representation of clinical and resting-state data using canonical correlation analysis, and hierarchical clustering to identify biotypes. In this chapter, several problems of the original analytical pipelines that lead to biased and overly confident results were identified. Mainly, resting-state features were selected based on their correlation with the clinical variables, which leads to overly optimistic p-values for the subsequent canonical correlation analysis between the selected resting-state features and clinical symptoms. Next, we demonstrated that the criteria for defining "biotypes" would produce spurious biotypes even if there are no real clusters in the data. Last, we showed that due to the overfitting of the canonical correlation analysis, the biological and clinical canonical variates and, therefore, subtypes could be extremely unstable. Due to these methodological limitations, we concluded that the presented analysis does not provide sufficient evidence for biotypes of depression.

Chapter 4 contains an evaluation of performance measures used in machine learning studies. We showed that the most commonly used measure has the worst statistical properties, and it is also suboptimal for clinical settings compared to alternatives. We highlighted that there is no one best performance measure, but that the selection of appropriate performance measures should be based on the specific goals of the machine learning model. We reviewed four types of performance measures focusing on their applications in neuroimaging and clinical settings. Next, using simulated and real datasets, we evaluated the statistical properties of these measures, including statistical power, detecting model improvement, selecting informative features, and reliability of results. Accuracy, although the most commonly used performance measure, had the worst statistical properties compared to alternatives, therefore it should be avoided when the statistical inference is the primary goal of the machine learning model. Furthermore, accuracy should also be avoided when evaluating machine learning models in a clinical setting because it does not take into account the uncertainty of predictions and the relative cost of false positive and false negative misclassifications.

Chapter 5 focuses on an evaluation of machine learning models in the presence of confounding variables. We showed that the most commonly used method does not sufficiently control for confounding variables and thus can lead to wrong results and introduced an alternative approach that does not have this problem. Confounding is an important problem in machine learning. For the translation of machine learning models to clinical practice, it is important to know that the machine learning model is predicting the clinical variable of interest and not some other variable such as age, gender, or scan-site. The most common method to correct for confounds in machine learning models is to regress out confounding variables from each input variable separately(Fortin et al. 2017; Snoek et al. 2019). We show that this method cannot sufficiently correct for confounds in machine learning studies, because machine learning methods can learn information from the data that cannot be removed using the traditional method. We propose a simple method where confounds are controlled for on the level of machine learning predictions and not input variables. We show in simulated and real datasets that this method correctly controls for confounds even in situations where the traditional approach fails.