Predicting therapeutic success in clinical depression: Are machine learning algorithms a good use?

23 May 2023, Julia Pöschko

Clinical depression is one of the most common clinical disorders worldwide. When dealing with depressive disorders, timely diagnoses and adequate treatment plans play a crucial role. However, patients do not respond equally well to different methods of treatment, and in fact, one third of patients are resistant to standard antidepressant treatments. Thus, developing knowledge about patients' characteristics and their influence on treatment success becomes more and more important. To predict therapeutic success in depressed populations, studies have begun to investigate the use of machine learning algorithms. But are they actually capable of making accurate predictions in this clinical context? With this question in mind, Lee and colleagues conducted the meta-analysis »Applications of machine learning algorithms to predict therapeutic outcomes in depression: A meta-analysis and systematic review«, published in 2018, to summarize previous findings on this research topic.

Overview: Meta-Analysis

Focus of the study: Prediction of depression-related therapeutic outcomes using machine learning algorithms

Target group: Adults ages 18 or above with unipolar or bipolar depression, fulfilling the criteria provided by a diagnostic manual

Average effect size: Machine learning models predicted therapeutic outcomes in depressed adults with a total classification accuracy of 0.82 (95% CI: [0.77, 0.87]).

Additional results:

- Machine learning models that integrate multiple predictive factors (phenomenological, neuroimaging, genetic) showed better prediction accuracy than models that include just one type.
- In some studies, machine learning algorithms were reported to be more powerful to detect effects compared to traditional methods.
- Publication bias was found.

Download short review

Download PDF

Read Short Review Online

Introduction

Clinical depression, or, more formally, major depressive disorder (MDD), is a mental disorder with persistent symptoms such as overwhelming feelings of sadness, worthlessness, guilt, hopelessness, and anhedonia. MDD is one of the most prevailing clinical disorders and the most common affective disorder. Generally, patients who remit earlier tend to have better, non-lasting outcomes. However, timely diagnoses and treatments are rare, and **treatment selection** is a

trial-and-error process, since patients do not respond equally well to different methods of treatment. In fact, about one-third of MDD-patients are not sensitive to the standard antidepressant treatments.

Thus, the identification of **reliable predictors of therapeutic outcomes** constitutes an important factor to improve on-time diagnoses and treatment. However, previously conducted studies have not found single variables that robustly predict therapeutic outcomes, but researchers emphasize the need to **integrate multiple factors** and sources to make accurate predictions. Suitable for integrating a large number of variables and analyzing their relevance for specific outcomes is the approach of machine learning.

Machine learning denotes a set of computational methods that make use of algorithms to recognize patterns in data and to predict output from certain input (Helm et al., 2020). In the recent past, studies have begun to investigate predictors of therapeutic results in depression using such machine learning methods. However, no meta-analysis has yet been conducted to review and summarize the findings of these studies. This was set as the aim of the current meta-analysis.

What is this study about?

The goal of this study is to review and summarize previous research that investigated the **application of machine learning models for prediction of therapeutic results** in depressed adults. By doing this, the meta-analysis wants to identify how and to what extent prior research has used machine learning algorithms to establish knowledge about the selection and individualization of therapeutic treatments for clinical depression.

To conduct the meta-analysis, the authors searched through various databases to find suitable research articles. Articles were selected according to following broad **inclusion criteria**:

- Publication date prior to February 8, 2018
- A machine learning algorithm was applied
- Individual-level or group-level data as predicting variables
- Sample consisting of adults aged 18 or above with unipolar or bipolar depression, diagnosed by a diagnostic manual
- Intervention being at least one evidenced-based treatment for depression (e.g., psychotherapy, pharmacotherapy)
- Longitudinal change in depression outcome, with a standardized measure of therapeutic improvement

- Prospective or retrospective study design, open-labelled or controlled study, with or without randomization

The initial search, conducted by searching through the databases with topic-related keywords like ‘mood disorder’ and ‘machine learning’, yielded an overall amount of 639 articles. However, after excluding articles that did not meet the inclusion criteria, the authors were left with 26 studies for the qualitative analysis (review), and from those, **20 for the quantitative analysis** (meta-analysis). The overall sample of the meta-analysis includes 6325 participants. Most of the considered studies are peer-reviewed research articles published in journals. The majority was published after 2015, suggesting an upward trend in the number of publications since then. The oldest study was from 2004. The articles were conducted in diverse countries, including the USA, Canada, UK, Australia, Austria, Italy, Germany, Netherlands, China, and Malaysia.

Concerning **machine learning**, most of the articles utilized **classification**, a supervised machine learning approach. A classification model is trained on a labeled dataset and in this research context, learns to correctly classify patients as treatment responders or non-responders based on predicting variables. Besides this, two studies utilized a **clustering** method, an unsupervised machine learning approach. In clustering approaches, patients are clustered in groups regarding their therapeutic outcomes, with the aim to subsequently identify differentiating factors. However, the studies that made use of clustering were not included in the quantiative analysis.

A range of variables were used in the considered studies as **predictors of therapeutic outcomes**. To group them for the meta-analyis, the authors allocated them to specific classes, being **‘neuroimaging’, ‘phenomenological’, ‘genetic’ or ‘combined’**. Neuroimaging predictors refer to neuroanatomical structural or functional connectivity aspects, measured by electroencephalography (EEG) and structural (MRI) or functional Magnetic Resonance Imaging (fMRI). Phenomenological variables include sociodemographic, anthropometric , psychometric, or neurocognitive aspects or aspects about the psychiatric background. Genetic factors relate to specific genetic markers that are found to be associated with therapeutic responses, by for example having an impact on serotonin brain receptors.

To summarize the findings of the included research articles, the authors analyzed the **overall classification accuracy** of the machine learning models, expressed as a proportion between 0 and 1 (0-100%). This corresponds to the percentage of correctly classified subjects into treatment responders or non-responders. Beyond this meta-analysis of classification accuracy, the authors also conducted a **meta-regression analysis**. This technique is used to investigate how certain characteristics of the incorporated studies relate to variation of the classification accuracy. Such characteristics are called **moderators** (moderating variables). Moderators that

were investigated by the authors - in terms of whether they have an impact on the classification accuracy - were:

- Type of predictor variable (‘neuroimaging’, ‘phenomenological’, ‘genetic’, or ‘combined’).
- Machine learning technique
- Date of study publication
- Country of the affiliation of the primary study author
- Impact factor, which represents a ranking of scientific journals regarding their number of citations and published articles, to determine journals with the highest scientific influence within its research area (Casadevall & Fang, 2014; Garfield, 2006)

What does this study find?

The **meta-analysis of the classification accuracy** came to following results: In consideration of all the included studies, the overall classification accuracy of the machine learning models was 0.82 (95% CI: [0.77, 0.87]). This implies that in 82% of the cases, the machine learning algorithm was able to correctly classify the therapeutic outcome (treatment responders vs. non-responders). This result suggests that overall, machine learning models are capable of predicting therapeutic outcomes.

When considering the different types of predictor variables, the use of combined predictors was associated with higher accuracy outcomes (accuracy = 0.93, 95% CI: [0.86, 0.97]) compared to the other types (neuroimaging: 0.85 [0.81, 0.88], phenomenological: 0.76 [0.63, 0.87], genetic: 0.68 [0.62, 0.74]), and this difference in accuracies was statistically significant ($\chi 2 = 31.39$, $df = 3$, $p < 0.0001$). These findings imply that the use of several predicting variables might increase the accuracy of prediction.

The **moderator analysis** (in the meta-regression analysis) yielded similar results: The classification accuracy was moderated by the type of predictor variable ($QM = 8.70$, $df = 3$, $p = 0.0335$; $QE = 137.90$, $df = 17$, $p < .0001$), suggesting that depending on the type of variable used for making predictions, the classification accuracy increases or decreases. However, no other moderating effects were found.

In the **qualitative analysis**, the authors additionally reviewed findings of studies which compared machine learning techniques to more conventional statistical methods. Overall, differences in the findings between both approaches were found. Three studies failed to find effects using conventional methods, while machine learning approaches predicted therapeutic outcomes with an accuracy between 78% and 91%. In another study, conventional analyses

were able to demonstrate some of the demographic and clinical predictors that an artificial neural network was able to establish. These findings emphasize the superiority of machine learning methods in specific cases, especially in complex research contexts like the prediction of therapeutic outcomes based on various factors.

One final aspect that the authors evaluated is the presence of **publication bias**. Publication bias refers to the selective publishing of studies: Findings that are statistically significant and preferred in their direction are generally more likely to be published than nonsignificant and unwanted results (Marks-Anglin & Chen, 2020). The funnel plot is a commonly used method to check for publication bias, assessing whether the published studies are symmetric around the expected effect size. In the current meta-analysis, the funnel plot showed asymmetry, and the corresponding statistical test (Egger's test) yielded a significant result (slope = 0.94, p = 0.0026). These results both suggest the presence of a publication bias. The trim and fill-method, which adjusts for the asymmetry of studies, added 10 missing studies, leading to an adjusted estimation of the classification accuracy of 0.71 (95% CI: [0.64, 0.78]).

How does the Digital Psychology Lab Teaching evaluate this study?

How substantial is the outcome? The overall classification accuracy was 0.82, and the largest accuracy was 0.93, reported for combined predictor variables. Although there is no universally used criterion or reference point that indicates the goodness of accuracy values, the accuracies stated here are in this context viewed as considerable, since 82% (93%) of subjects were classified correctly into treatment responders or non-responders. Thus, the findings of this meta-analysis be considered as substantial.

How precise are the outcomes? The classification accuracy was analyzed overall and for each predictor type ('neuroimaging', 'phenomenological', 'genetic', 'combined'). Furthermore, meta-regression analyses were conducted to assess whether certain study-specific variables have an impact on the classification accuracy outcome. Here, it was reported that 'combined' predictors lead to more accurate predictions, while 'genetic' predictors have lower accuracy. No other study-related aspects however displayed moderating effects on the classifcation accuracy. Overall, the results can be considered as precise.

Is the outcome generalizable? For study search and selection, broad inclusion criteria were utilized, beneficial for generalizability. The core results of the meta-analysis were derived from a sample of 20 studies, overall including a large sample of over 6000 subjects. The studies included in the meta-analysis analyzed various predictor variables and made use of different

machine learning algorithms. Except for the predictor variable type, no moderation effects were found, suggesting the results are very robust in terms of classification technique, time frame, country, and impact factor. However, differences in accuracies were found for predictor variable type, and the results might not apply when using genetic predictor types due to their lower accuracy scores. Moreover, since the subjects were required to fulfill specific depression-related diagnostic criteria, the results can only be generalized within this psychiatric subgroup. In addition, generalization to young age groups is not possible, since the participant requirement was age 18 or above. Therapeutic outcome in patients was measured with overall depressive symptom severity, which might not directly generalize to patients' psychosocial or occupational functional recovery. Thus, overall, the results of the meta-analysis can be viewed as generalizable within some restrictions.

How relevant is the outcome, scientifically and therapeutically? The meta-analysis identified factors that predict therapeutic success robustly and showed that machine learning algorithms are capable of making predictions. This result has valuable implications for scientific research and therapeutic treatment of depression in the field. In research, machine learning models can be applied to gain more knowledge about predictors of therapeutic success – to develop an understanding of patients' characteristics and their associations to treatment outcomes. Such research findings in turn provide evidence that is useful in real-life therapeutical settings and can be used to improve individual treatment selection to be timely and less cost intensive.

How high is the credibility of the outcome? The meta-analysis was conducted by considering several official guidelines, including the Cochrane Handbook for Systematic Reviews of Interventions, the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) statement, and the Joanna Briggs Institute's Manual for Scoping Reviews. In their article, the authors clearly described their search strategy to find relevant research articles and included a detailed list of search terms. Moreover, a PRISMA flow diagram was depicted, summarizing the search strategy. The authors stated which databases were used, and what the inclusion and exclusion criteria were for selecting articles. Non-refereed (grey) literature was not excluded from the analysis, beneficial to counteract publication bias, but possibly disadvantageous for research quality. Regarding publication bias, the authors conducted several assessments to obtain an estimation, including funnel plots, statistical tests and the trim and fill-method. However, no information about interrater reliability for the process of rating the studies was given. Overall, the quality of the meta-analysis can be considered as high, despite some minor points of criticism. However, the presence of publication bias should not be neglected. According to the trim and fill results, the classificatoin accuracy might in fact be lower than

reported in the meta-analysis (0.71 instead of 0.82), having an impact on the credibility of the outcome.

Conclusion for researchers, therapists, and depression-affected individuals

The results of the meta-analysis suggest that machine learning can be a **powerful tool** to predict depression-related therapeutic outcomes. By using machine learning approaches, large numbers of variables can be analyzed regarding their prediction power, and patterns in data can be found that might not be detectable with conventional statistical methods.

Thus, applying this technology can be of **good use**, beneficial to researchers, therapists, and depression-affected individuals in general. For clinical research, the application of machine learning can help to improve models to predict treatment responses and deepen the knowledge about influential patient characteristics. For therapeutic settings, the findings emphasize the importance of considering multiple patient-related characteristics for treatment selection. By conducting further studies that utilize machine learning approaches, individualization of depression-related therapy can be further advanced, representing a promising outlook for depression-affected populations.

However, some **limitations** have to be mentioned in addition. Since publication bias was detected, the overall classification accuracy may be lower than reported in the analysis. Although machine learning is a powerful tool, some machine learning techniques do not give much insight into how results are obtained, or information about the individual variables' relevance for prediction. This can result in missing information or lead to more serious problems, like discrimination against humans.

Overall, machine learning represents a promising technology in the field of predicting depression-related therapeutic outcomes, when not neglecting potentially negative or harmful aspects.

Study example

Previous studies have demonstrated that in patients with major depressive disorder (MDD), **structural brain changes** and abnormalities in grey matter (GM) and white matter (WM) structures occur, possibly being the cause for depressive symptoms. To investigate which neurological markers in patients’ brain structures might reveal valuable information about

individual treatment sensitivity, Liu et al. (2012) conducted a study focusing on magnetic resonance imaging (MRI) scans.

For this purpose, Liu and colleagues recruited clinically depressed patients who were diagnosed with MDD according to standardized (DSM-4) criteria, in addition to 17 healthy controls (control-group). 17 subjects of the clinical group were affected by treatment-sensitive depression (TSD), while 18 patients had treatment-resistant depression (TRD). Severity of depressive symptoms was additionally measured, using the clinician-administered Hamilton Rating Scale for Depression (HRSD).

To analyze GM and WM brain structures in the MRI scans, three methods were used. Firstly, traditional **voxel-based analysis** was conducted, in which volumes of specific brain regions are compared between groups (with simple t-tests). Secondly, a modified **multivariate pattern analysis (MVPA)** was utilized. The MVPA combines several machine learning approaches, including a classification algorithm that learns to classify patients into groups based on brain scans. Thirdly, **correlational analyses** were performed to investigate associations between specific brain region volumes and the severity of depressive symptoms.

The results revealed that the MVPA method was able to **classify depressed patients as either treatment-sensitive or treatment-resistant with considerable accuracy** (up to 82.9%), based on GM and WM information from certain brain areas. Furthermore, the method was able to classify subjects as treatment-sensitive or healthy, and treatment-resistant or healthy, with accuracies up to 91.2%. These results imply that specific GM and WM areas in the brain reveal certain information about whether a person is depressed, and whether a depressed patient is sensitive or resistant to treatment.

When correlating the established brain regions in their volumes with the severity of depression (in all patients), statistically significant associations were found. However, in contrast to the findings above, the conventional **voxel-based analysis did not yield significant volumetric differences** in any brain areas between the TRD and TSD group.

Overall, the results of this study suggest that structural MRI and MVPA might be valuable techniques to investigate neuroanatomical factors in their ability to distinguish between MDD patients and healthy controls, and between TRD and TSD patients. According to the authors, the **difference in findings between the conventional voxel-based analysis and MVPA** could be ascribed to the inability of the voxel-based analysis to identify subtle differences between the TRD and TSD groups. This constitutes a case that demonstrates the superiority of machine learning methods over traditional techniques.

References

Meta-analysis:

Lee, Y., Ragguett, R.-M., Mansur, R. B., Boutilier, J. J., Rosenblat, J. D., Trevizol, A., Brietzke, E., Lin, K., Pan, Z., Subramaniapillai, M., Chan, T. C. Y., Fus, D., Park, C., Musial, N., Zuckerman, H., Chen, V. C.-H., Ho, R., Rong, C., & McIntyre, R. S. (2018). Applications of machine learning algorithms to predict therapeutic outcomes in depression: A meta-analysis and systematic review. *Journal of Affective Disorders, 241* , 519–532. https://doi.org/10.1016/j.jad.2018.08.073

Study example:

Liu, F., Guo, W., Yu, D., Gao, Q., Gao, K., Xue, Z., Du, H., Zhang, J., Tan, C., Liu, Z., Zhao, J., & Chen, H. (2012). Classification of different therapeutic responses of major depressive disorder with multivariate pattern analysis method based on structural MR scans. *PLoS ONE, 7*(7). https://doi.org/10.1371/journal.pone.0040968

Additional references:

Casadevall, A., & Fang, F. C. (2014). Causes for the persistence of impact factor mania. *mBio, 5*(2), e00064-14. https://doi.org/10.1128/mbio.00064-14

Garfield, E. (2006). The history and meaning of the journal impact factor. *JAMA, 295*(1), 90. https://doi.org/10.1001/jama.295.1.90

Helm, J. M., Swiergosz, A. M., Haeberle, H. S., Karnuta, J. M., Schaffer, J. L., Krebs, V. E., Spitzer, A. I., & Ramkumar, P. N. (2020). Machine Learning and Artificial Intelligence: Definitions, applications, and future directions. *Current Reviews in Musculoskeletal Medicine, 13*(1), 69–76. https://doi.org/10.1007/s12178-020-09600-8

Marks-Anglin, A., & Chen, Y. (2020). A Historical Review of Publication Bias. https://doi.org/10.31222/osf.io/zmdpk

Picture: uploaded by cottonbro studio on Pexels for free use (https://www.pexels.com/de-de/foto/mann-paar-menschen-buro-4101143/)