The promise of machine learning in predicting treatment outcomes in psychiatry

Adam M. Chekroud[1,2], Julia Bondar[2], Jaime Delgadillo[3], Gavin Doherty[4], Akash Wasil[5], Marjolein Fokkema[6], Zachary Cohen[7], Danielle Belgrave[8], Robert DeRubeis[5], Raquel Iniesta[9], Dominic Dwyer[10], Karmel Choi[11,12]

[1]Department of Psychiatry, Yale School of Medicine, New Haven, CT, USA; [2]Spring Health, New York City, NY, USA; [3]Clinical Psychology Unit, Department of Psychology, University of Sheffield, Sheffield, UK; [4]School of Computer Science and Statistics, Trinity College Dublin, Dublin, Ireland; [5]Department of Psychology, University of Pennsylvania, Philadelphia, PA, USA; [6]Department of Methods and Statistics, Institute of Psychology, Leiden University, Leiden, The Netherlands; [7]Department of Psychiatry and Biobehavioral Sciences, University of California, Los Angeles, Los Angeles, CA, USA; [8]Microsoft Research, Cambridge, UK; [9]Department of Biostatistics and Health Informatics, Institute of Psychiatry, Psychology and Neurosciences, King's College London, London, UK; [10]Department of Psychiatry and Psychotherapy, Section for Neurodiagnostic Applications, Ludwig-Maximilian University, Munich, Germany; [11]Harvard T.H. Chan School of Public Health, Boston, MA, USA; [12]Department of Psychiatry, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA

For many years, psychiatrists have tried to understand factors involved in response to medications or psychotherapies, in order to personalize their treatment choices. There is now a broad and growing interest in the idea that we can develop models to personalize treatment decisions using new statistical approaches from the field of machine learning and applying them to larger volumes of data. In this pursuit, there has been a paradigm shift away from experimental studies to confirm or refute specific hypotheses towards a focus on the overall explanatory power of a predictive model when tested on new, unseen datasets. In this paper, we review key studies using machine learning to predict treatment outcomes in psychiatry, ranging from medications and psychotherapies to digital interventions and neurobiological treatments. Next, we focus on some new sources of data that are being used for the development of predictive models based on machine learning, such as electronic health records, smartphone and social media data, and on the potential utility of data from genetics, electrophysiology, neuroimaging and cognitive testing. Finally, we discuss how far the field has come towards implementing prediction tools in real-world clinical practice. Relatively few retrospective studies to-date include appropriate external validation procedures, and there are even fewer prospective studies testing the clinical feasibility and effectiveness of predictive models. Applications of machine learning in psychiatry face some of the same ethical challenges posed by these techniques in other areas of medicine or computer science, which we discuss here. In short, machine learning is a nascent but important approach to improve the effectiveness of mental health care, and several prospective clinical studies suggest that it may be working already.

Key words: Computational psychiatry, machine learning, treatment outcomes, prediction, external validation, pharmacotherapies, psychotherapies, electronic health records, smartphone data

(World Psychiatry 2021;20:154–170)

Treatment interventions in psychiatry are far from being effective in all cases in which they are indicated. In depression, for example, only 30-50% of individuals achieve remission after whatever initial treatment they receive, even in the context of a well-conducted clinical trial[1]. Eventually, after trying some number or combination of treatments, most patients do attain remission. What if, rather than iterating through the available treatments that a patient *might* benefit from, we could predict the right treatment for each individual from the start?

Researchers have wanted this for decades. Historically, they have tried to understand specific factors involved in treatment response based on theoretical groundings, leading to many studies focusing on single variables such as early childhood stress, suicidality, major life events, or comorbid diagnoses. Since then, the ongoing search for one (or a few) true explanatory variables has included many levels of analysis, including: the patient (clinical characteristics, blood marker levels), his/her brain (structural and functional neuroimaging, cerebral blood flow, scalp electrical recordings), his/her genes (single nucleotide polymorphisms, mutations/rare genetic variants, copy number variations, gene expression), and intervention characteristics (the medication or psychotherapy selected, the way it was delivered, the provider, the therapeutic alliance). If one variable alone could accurately predict treatment response, our field would probably have found it by now. Instead, most characteristics identified so far have shown small explanatory power over treatment outcomes, and researchers' attention naturally turned towards multivariable models that can incorporate many smaller effects.

Machine learning is a collection of statistical tools and approaches that are extremely well suited to this goal of detecting and aggregating small effects in order to predict an outcome of interest[2]. It allows researchers to go from evaluating a small number (~10) of predictor variables to many hundreds or thousands of variables or variable combinations. There are many potential pitfalls when applying these techniques, but, when implemented well, they afford many opportunities for psychiatric research[3,4]. They allow us to examine many variables, even correlated ones, simultaneously. They move away from exclusively additive models and allow us to identify more complex non-linear patterns in data. They more naturally bridge disparate data types, potentially incorporating clinical assessments, geospatial information, and biological findings into a single analysis. By unlocking powerful hypothesis-free approaches, they enable us to discover factors that are less intuitive but nonetheless predictive of outcomes.

The introduction of machine learning in psychiatry is more than just adding an analysis tool for combining and exploring bigger data sets – it marks a paradigm shift[5]. For years, we used classical statistical approaches to confirm or refute specific hypotheses. Now, machine learning studies shift the focus toward the overall predictive power of a model, particularly how accurately it predicts the desired outcome in a new, unseen dataset. Studies in this field are evaluated primarily by their potential clinical impact: what our model can reliably tell us about the prognosis of new patients in the future, and what we can do with that information to improve clinical practice.

With this in mind, this paper explores the promise of machine learning in predicting treatment outcomes in psychiatry. There are many things that we do not focus on. This is not a primer on machine learning[6], an explanation of how it works[2], or a debate about what counts as machine learning versus traditional statistics or "non-machine-learning". We do not explain how to build predictive models[7] or how to validate them. We are not formally comparing different algorithmic approaches, how each one works, or circumstances where one may be more appropriate than another. We also avoid a distinction between moderators versus mediators of treatment outcomes, or whether a model predicts outcomes specifically for a treatment versus others or predicts outcomes more generically for multiple treatments[8]. Finally, we do not aim to review the many sociodemographic and clinical variables that have been or can be used for prediction of treatment response in psychiatry, which generally have the most predictive power and are cheapest to collect[9,10].

We begin by discussing machine learning methods, how they compare to traditional statistical approaches, and to what extent is machine learning specifically adding value. Next, we provide an overview of the interventions for which researchers have tried to use machine learning methods to predict outcomes, ranging from medications and psychotherapy to digital interventions and neurobiological treatments. In doing so, we highlight characteristics that made them gold standard examples, and discuss the different goals that can be achieved in each context. Next, we focus on the potential utility of electronic health records, smartphone and social media data, and of data from genetics, electrophysiology, neuroimaging and cognitive testing for the development of predictive models based on machine learning. Finally, we help the reader understand the broader context: how close have we come to implementing these prediction tools in real-world clinical practice; and what are the ethical challenges that these tools carry. The intent of this paper is to review studies throughout psychiatry; any emphasis on depression is not intentional, but it does reflect the fact that the majority of research in this field has been conducted in people with that mental disorder.

IS MACHINE LEARNING ADDING VALUE OVER TRADITIONAL STATISTICS?

Machine learning studies generally differ from traditional research in two ways. The first is a focus on prediction (explanatory power of the model) rather than inference (hypothesis testing). The second is a shift towards model flexibility, with the ability to handle large numbers of predictors simultaneously.

Prediction can be performed without machine learning algorithms, and many studies still use traditional statistical techniques such as logistic regression. In fact, when assumptions and sample size requirements are reasonably met, the number of predictors is small (≤25), and non-linear effects are relatively weak, traditional parametric models will likely predict well. Several studies found no benefit of machine learning over traditional logistic regression, for example in predicting treatment resistance in major depression[11], brain injury outcomes[12], or major chronic diseases[13].

One recent systematic review of clinical prediction models found no difference in performance between machine learning and logistic regression[14], although the authors considered in the category of logistic regression some advanced frameworks that could be included within machine learning, such as penalization (e.g., lasso, ridge or elastic net) and splines (which capture non-linearities). In areas of medicine such as diabetes and heart failure, simple logistic models have performed well and have been externally validated more than machine learning models[15,16].

The added value of machine learning approaches emerges when the number of potential predictors is large and/or their effects are non-linear. Many machine learning algorithms are capable of handling large numbers of predictors, even in cases where there are more predictor variables than observations, due to built-in overfitting control. For example, ridge, lasso and elastic net regression[17] include penalization, which forces the regression coefficients to be closer to zero than in the traditional linear or logistic regression models. Machine learning approaches are also good at capturing complex, interactive, or non-linear effects. For example, tree-based models are able to evaluate many possible variables and variable combinations to identify subgroups that could not be captured by traditional linear models. Another common technique adopted by machine learning approaches is "ensembling". Here, several models are fitted on random samples of the original dataset, and then an average is taken amongst the predictions from each model. This approach is a key element of many popular machine learning techniques today, especially gradient boosting machines and random forests[18-20].

Several recent treatment outcome prediction studies in psychiatry demonstrated the added value of machine learning. Random forests and/or elastic net regression[21-24], as well as support vector machines[25], were found to outperform traditional regression methods. Large-scale comparisons on benchmark datasets consistently found machine learning to outperform traditional methods[26-29]. Overall, boosted trees (random forests and gradient boosting machines), regularized regression, support vector machines, and artificial neural networks can all perform well, but no one method will have the best performance across all situations.

While researchers typically aim to maximize predictive performance, practical aspects such as explainability or the cost of including more variables should also be considered. In some cases, simpler models with slightly lower predictive accuracy or higher generalizability might be preferred, because they already capture most of the effects[30,31]. There is no silver bullet in statistics, and all prediction algorithms face the so-called bias-variance tradeoff[2,32,33], where flexibility needs to be balanced with the risk of overfitting. For machine learning methods to capture increasingly complex effects, much larger sample sizes are still needed. Although these methods can deal with large numbers of potential predictor variables, careful pre-selection of variables likely improves predictive accuracy.

While traditional research approaches focused on p values for specific coefficients in a model, prediction studies focus on the overall explanatory power of the model, often in terms of R^2, balanced accuracy, or area under the receiver operating characteristic curve (AUC). Predictive studies require a keen focus on validation approaches, to examine whether the model is learning patterns that are substantive and consistent from one dataset to another, or whether the model has simply learned idiosyncrasies of the initial training data. Table 1 discusses various kinds of validation that are conducted in predictive studies, from internal approaches that use just one dataset, to external validation approaches that use data from independent sites, studies, trials, countries, or consortia to test model generalizability. Validation frameworks, especially external validation, are critical for developing models that are reliable and useful, and understanding whether the fitted model is likely to generalize to unseen data in the future[34-36].

PREDICTING TREATMENT OUTCOMES IN PSYCHIATRY BY USE OF MACHINE LEARNING

Medications

Predicting treatment outcomes for psychiatric medications is the most active area of research in the field, primarily because they were the easiest place to start. Machine learning studies require large volumes of data to build predictive models, ideally with clearly labelled outcomes, control over the intervention, and relevant data about the patients before treatment. Since this describes most large clinical trials, and most large clinical trials in psychiatry are conducted to evaluate efficacy of a medication, most machine learning efforts began by investigating treatment responses to medications treating depression, schizophrenia or bipolar disorder.

These studies mostly used information from demographic intake forms and clinical symptom scales common in clinical trials, although more recently genetic and neuroimaging data have also been incorporated (discussed later in this paper). Despite being the most active area of research, most resulting models have not yet been validated in external samples. Relatively few prediction tools generated by mental health researchers so far have advanced through implementation studies and into clinical practice[37-39]. Here we focus on examples of studies that were adequately powered, underwent external validation, or are notable for other reasons.

Most treatment prediction studies have focused on antidepressants commonly used in the acute phase of depression. For example, Chekroud et al[40] determined a small group of 25 pretreatment variables that were most predictive of remission with citalopram in the Sequenced Alternative Treatments for Depression (STAR*D) trial. This model achieved an accuracy of 64.6%. The model was then applied to data from another clinical trial to examine whether it can generalize to patients from an entirely independent population. The model was able to predict response to two similar antidepressant regimens (escitalopram plus placebo, and escitalopram plus bupropion, each with an accuracy of around 60%), but the model did not predict remission better than chance for patients who took venlafaxine plus mirtazapine (51%).

The five most important variables identified by the model in predicting remission were baseline depression severity, employ-

Table 1 Common validation approaches used in clinical prediction studies

Generalizability test	Description
None, p value testing	The entire sample is used to predict an outcome, and a p value indicates the probability of obtaining the result in the absence of a true effect. The study cannot make any claims concerning translation or generalizability because they have not been tested.
Leave-one-out cross-validation	One subject is randomly chosen and left out. A model is trained on the remaining subjects and applied to the left-out subject to assess generalizability. This procedure is repeated for every subject in the dataset. This is the simplest form of cross-validation. It produces optimistic biased results.
K-fold cross-validation	The sample is randomly divided into subsamples (called "folds"). One fold is left out and statistical models are trained on the remaining subjects. The models are applied to the subjects in the left-out fold to assess generalizability. This is a common technique to reduce overfitting. However, when the data are from one sample (even if collected at multiple sites), generalizability claims need to be tempered.
Leave-one-site-out cross-validation	Instead of randomly leaving out subjects, sites are now randomly left out. Models are fitted on the remaining sites, and applied to the left-out site. This assesses cross-site generalizability, and the same technique can be extended to any other group definition, such as blocks of time, gender or ethnicity. Generalizability and translational claims still need to be tempered.
External validation	A model is created in one study and applied to a completely separate sample. This approach reflects a high degree of generalizability capacity. Demonstrations can be increasingly close to real-life circumstances, which strengthens the evidence of generalizability and translational potential (but does not guarantee it). The approach may still be subject to poor sociodemographic representation, sampling biases, or study designs that do not reflect clinical reality.
Prospective validation	A previously-created model is evaluated in a prospective study that is ideally randomized and in conditions as close to clinical reality as possible, in order to test whether the tool is safe and effective in practice. Prospective validation studies are still susceptible to the same concerns around external validity as all other clinical trials (e.g., participant compensation and meaningful endpoints), and require large sample sizes, a broad and unbiased recruitment process, and good clinical practices. As with other clinical trials, a phased process may be necessary to first evaluate feasibility and safety in a smaller sample before proceeding to broad evaluation of effectiveness.

ment status, feeling restless during the past seven days (psychomotor agitation), reduced energy level during the past seven days, and Black or African American ethnicity. The study was later replicated by Nie et al[41], who similarly trained a model to predict citalopram treatment outcomes using information easily obtainable at baseline. The team trained and tested the model in the STAR*D dataset and validated it in data from a different open-label citalopram trial, using 22 predictor variables that overlapped between the two trials. Despite minor differences depending on the specific algorithm used, the balanced accuracy of the models was roughly 64-67%.

An earlier study by Perlis[11] showed that eventual treatment resistance might also be predictable from the outset. The author developed a model using STAR*D data that was able to predict at baseline whether an individual would not reach remission after two antidepressant treatment trials, with an AUC of 0.71. Early proofs of concept like the Perlis study did not include external validation, at least partly due to the lack of independent datasets with similar trial designs that could be used for that validation.

The above antidepressant studies selected predictors in a purely data-driven way, including all data that could be extracted at baseline and then using machine learning methods that discard irrelevant information or are amenable to including many variables at once. However, the choice of predictors is not always hypothesis-free, and *a priori* knowledge from scientific literature can also guide the choice of variables and yield useful results. Iniesta et al[42] aimed to predict remission of depression in patients treated with escitalopram or nortriptyline using only variables that had previously been confirmed as individual predictors or moderators of response to treatment. Their models predicted overall response to medication with an AUC of 0.74 and response to escitalopram with an AUC of 0.75, but prediction of nortriptyline outcomes was not statistically significant. In subsequent work incorporating genetic data to the models[43], these authors predicted response to escitalopram and nortriptyline with an AUC of 0.77.

A second use of machine learning to predict medication outcomes is to better define subgroups of patients, symptoms, or symptom trajectories, and then use these subgroups to make more nuanced predictions. Drysdale et al[44] used clustering to identify four "subtypes", or groups, amongst 1,188 depressed patients based on patterns of dysfunctional connectivity in limbic and frontostriatal networks. They developed classifiers for each depressive subtype using support vector machines and later tested these models on an independent dataset, accurately classifying 86.2% of the testing sample. As a next step, the team used the subtypes to predict response to transcranial magnetic stimulation, but did not validate these predictions in any independent sample. Although the biotypes approach is interesting, subsequent methodological research has highlighted concerns and limitations[45].

Chekroud et al[46] used clustering to identify groups of symptoms and mixed-effects regression to determine if they had different response trajectories. Three symptom clusters (core emotional, sleep and atypical) emerged consistently from two independent medication trials – STAR*D and Combining Medications to Enhance Depression Outcomes (COMED) – across two commonly used symptom scales. The authors subsequently used data from STAR*D to train gradient boosting machines (one for each combination of cluster and medication arm), finding modest improvements in the ability of clusters of symptoms to predict total severity outcomes. The same symptom clustering approach was also effective in a study of treatments for adolescents[47].

Other researchers first used techniques like growth mixture modeling[48] or finite mixture modeling[49] to identify trajectories of symptom response such as "fast and stable remitter", "sustained response", or "late relapse". Machine learning models were then developed to try and predict the specific response trajectory a patient will have for a given treatment. This approach is potentially more robust to the noise that is naturally present amongst individual patient trajectories and less affected by the way that outcomes are defined in trials – e.g., whether remission is defined as a score of 5 on the Patient Health Questionnaire-9 (PHQ-9) or a score of 5 or 6 on the Quick Inventory of Depressive Symptomatology (QIDS)[48,49]. However, the approach relies on the availability of repeated measures.

Medication treatment outcomes have been most widely studied in depression, due to the prevalence of the condition and extant available data, but the approach has also been proven in other psychiatric conditions. For schizophrenia, Koutsouleris et al[25] used data from the European First Episode Schizophrenia Trial (EUFEST, N=344) to predict good and bad outcomes based on global functioning scores over time using a support vector machine, and validated the ten most predictive features on an unseen sample of 108 patients with a balanced accuracy of 71.7%. The most valuable predictors identified were largely psychosocial variables, rather than symptom data: unemployment, poor education, functional deficits, and unmet psychosocial needs.

Again in schizophrenia, Leighton et al[50] were not only successful in predicting response to medication treatment in first episode psychosis, but also in validating findings in two independent samples. They first identified predictors that were available across three studies – the Evaluating the Development and Impact of Early Intervention Services (EDEN) study in England, two cohorts recruited from the National Health Service (NHS) in Scotland, and the Danish clinical trial called OPUS. This allowed them to build and test harmonized models across the three studies to predict four outcomes capturing different aspects of recovery: symptom remission, social recovery, vocational recovery and quality of life. Next, they used logistic regression with elastic net regularization to identify the most relevant predictors in the EDEN study (N=1027) – much like Chekroud et al[40] – to determine a smaller subset of variables that could still predict outcomes but require less effort for future data collection and improve clinical applicability. These regularized models trained in the EDEN sample reached internal validation AUCs of 0.70 to 0.74 (depending on the outcome measure). When tested in the second Scottish cohort, the AUC ranged from 0.68 to 0.87. In the OPUS trial, it ranged from 0.57 to 0.68.

Predicting medication response in other mental disorders is still in early stages. Two studies[51,52] used baseline sociodemographic, clinical and family history information to predict response to medications commonly used in bipolar disorder: lithium and quetiapine. Although both obtained models with performance above chance, neither was validated in independent samples, and one used 180 variables for prediction[51], which limits its clinical applicability.

Psychotherapies

Historically, efforts to predict treatment outcomes in psychotherapies have focused on theoretically-motivated single variables that might moderate treatment outcomes. Only relatively recently have psychotherapy researchers applied machine learning approaches to predict treatment outcomes[53]. Even amongst these studies, the historical focus on moderators of psychotherapeutic effects has persisted, leading researchers to distinguish between "prognostic" and "prescriptive" models. Prognostic models are those that predict whether a patient will recover with a given treatment. Prescriptive models instead predict which of two (or more) treatments is best suited for a particular patient[54]. Both kinds of model can clearly have clinical utility, even if they answer slightly different questions. The differences continue to blur further with more recent attempts to build prescriptive models by developing multiple prognostic models for different treatments and then comparing their outputs[55].

In an early effort, Lutz et al[53] used nearest neighbor modeling to predict rate of symptom change and session-by-session variability. Models were based on age, gender and baseline symptom scores. Compared to non-machine learning models, the nearest neighbor predictions were more highly correlated with actual values of rate of change, but not session-by-session variability.

Since then, other approaches to prediction in psychotherapy proliferated. DeRubeis et al[56] developed a multivariable modeling method, known as the "personalized advantage index" (PAI), that uses interaction effects between baseline variables and treatment condition, to predict whether a patient will respond better to antidepressants versus cognitive behavioral therapy (CBT). Amongst their small sample of 154 individuals, a clinically meaningful advantage (PAI ≥3), favoring one of the treatments relative to the other, was predicted for 60% of the patients. When these patients were divided into those randomly assigned to their "optimal" treatment versus those assigned to their "non-optimal" treatment, outcomes in the former group were better (d=0.58, 95% CI: 0.17-1.01). Similar approaches have been developed by other groups[55,58], and more recently improved further by the use of machine learning approaches[59] to generate better predictions and incorporate more variables.

Several studies since then have tried to predict which evidence-based psychotherapy is most likely to benefit a specific patient[55,59], including efforts to identify which of two (or more) psychotherapies may be most effective[60,61], and whether a given patient is predicted to respond better to psychotherapy or medications[56]. A recent scoping review[62] identified a total of 44 studies that developed and tested a machine learning model in psychotherapy, but only seven of them reported on the feasibility of the tool. Since psychotherapy trials are often expensive and rarely have large sample sizes, some have argued that predictive models may need to be developed initially with large observational datasets[63].

PAI-style approaches that calculate treatment by variable interactions quickly lead to high-dimensionality prediction analyses that are prone to overfitting (or require very large sample sizes). Using data from two Dutch randomized trials, van Bronswijk et al[60] examined whether PAI models developed in one clinical trial dataset were able to successfully generalize to an independent dataset. Although the models performed statistically above chance in the trial used to train them, they did not generalize to the other clinical trial when predicting benefit for CBT versus interpersonal therapy (IPT) for depression.

The psychotherapy literature has generated several other prediction models, potentially optimizing significant aspects of patient care. For example, models have been developed[64,65] that would enable mental health providers to select low- or high-intensity treatments for patients on the basis of their expected prognosis. Other studies have tried to deconstruct the content that is traditionally combined to form a course of psychotherapy treatment, in order to predict which treatment components should be delivered within a given intervention, as well as the order in which the components should be implemented[66-68]. Other novel directions include using machine learning to match patients to specific therapists[69], replicating human ratings and judgements[70,71], and using natural language processing techniques to discover patterns of therapist-patient interactions that predict treatment response[72,73].

In general, many machine learning approaches to predict responses to psychotherapies are in the early stages of development[62]. However, a notable exception is found in the well-developed literature on routine outcome monitoring and "progress feedback". This involves tracking a patient's response to treatment in real time by entering his/her self-reported outcome/symptom measures into a computerized system that compares his/her response to predicted trajectories of improvement derived from clinical data using conventional statistical analyses (e.g., longitudinal multilevel/mixed models and growth curve modelling). There are now over 20 randomized controlled trials and several meta-analyses indicating that such clinical prediction models can help to improve treatment outcomes[74].

In addition to models investigating differential response to treatment and treatment optimization, the psychotherapy literature also includes adequately powered studies predicting overall response to treatment based on sociodemographic and clinical variables, much like the literature on response to medication. Buckman et al[75] built nine different models, using depression and anxiety symptoms, social support, alcohol use, and life events to predict depressive symptom response after 3-4 months of treatment in primary care settings. Models were trained on data from three clinical trials (N=1,722) and tested on three independent trials (N=1,136). All models predicted remission better than a null model using only one post-baseline depression severity measurement. Green et al[76] also predicted depressive symptom response to psychotherapy in 4,393 patients from community health services. They found that a model with only five pre-treatment variables (initial severity of anxiety and depression, ethnicity, deprivation and gender) predicted reduction of anxiety and depression symptoms with an accuracy of 74.9%. The number of sessions attended/missed was also an important factor affecting treatment response.

Digital CBT

In recent years, online delivery of mental health interventions has been seen as a promising approach to reducing barriers to care, with growing evidence for the effectiveness of both guided and unguided delivery[77,78]. Interventions such as internet-based CBT (iCBT) may be particularly amenable to the use of machine learning techniques, due to the possibility of longitudinal standardized collection of outcome data at scale, and the potential to directly incorporate machine learning outputs into online or app-based interventions. For example, in guided treatments, machine learning tools could provide feedback to therapists or alerts regarding risk. They could also be used to drive just-in-time adaptive interventions[79]. Smartphone delivery also opens up the possibility of automated collection of sensor data to derive behavioral markers[80], which would open up many possibilities for tailored interventions, while also raising a number of privacy and ethical concerns.

Machine learning-derived outcome predictions for iCBT may have advantages with regard to ease of deployment, for example by providing integrated decision support for case management. However, most existing work focused on predicting outcomes has been exploratory in nature and based on modest sample sizes. A key distinction is between approaches that use only baseline pre-treatment data, and hence may be applied to direct the choice of treatment, and approaches which use data gathered during the course of treatment, such as regular outcome measures or ecological momentary assessment (EMA).

As an example of the former, Lenhard et al[81] examined how clinical baseline variables can be used to predict post-treatment outcomes for 61 adolescents in a trial of iCBT for obsessive-compulsive disorder. Whereas multivariable logistic regression detected no significant predictors, the four machine learning algorithms investigated were able to predict treatment response with a 75 to 83% accuracy.

In a study which included, in addition to demographic and clinical data, therapy-related predictors of treatment credibility and working alliance, assessed at week 2 of treatment, Flygare et al[82] used a random forest algorithm to predict remission from body dysmorphic disorder after iCBT in a sample of 88 patients, comparing the results to logistic regression. Random forests achieved a prediction accuracy of 78% at post-treatment, with lower accuracy in subsequent follow-ups. The most important predictors were depressive symptoms, treatment credibility, working alliance, and initial severity of the disorder.

van Breda et al[83] added EMA data to models using baseline measures in a study predicting outcomes for patients who were randomized to blended therapy (face-to-face CBT and iCBT) or treatment as usual. This approach did not improve prediction accuracy.

The effectiveness of digital CBT interventions is mediated by patient engagement[84]. Detailed patient engagement data can be gathered automatically in online or app-based interventions; this may include data such as content views, completion of exercises, and interactions with clinical supporters[85]. Engagement data may be used within predictive models, providing interpretable and actionable outputs (e.g., the need for more frequent therapist contact in order to motivate greater engagement). Chien et al[86] analyzed engagement data from 54,604 patients using a supported online intervention for depression and anxiety. A hidden Markov model was used to identify five engagement subtypes, based on patient interactions with sections of the intervention. Interestingly, while in general patients who engaged more achieved better outcomes, the best outcomes were found in those who were more likely to complete content belonging to key components of CBT (i.e., cognitive restructuring and behavioral activation) within the first two weeks on the program, despite not spending the highest amount of time using the intervention. This work demonstrates the feasibility of gathering detailed engagement and outcome data at scale.

Interactions between patient and therapist, and the content of text in patient exercises, may also be analyzed using sentiment analysis techniques[87]. Analysis of patient texts might be embedded in therapist feedback tools for guided interventions, or as features within predictive models. Ewbank et al[73] conducted an analysis of 90,934 session transcripts (specifically, CBT via real-time text messages). Deep learning was used to automatically categorize utterances from the transcripts into feature categories related to CBT competences, and then multivariable logistic regression was applied to assess the association with treatment outcomes. A number of session features, such as "therapeutic praise", were associated with greater odds of improvement.

Chikersal et al[88] analyzed 234,735 messages sent from clinical supporters to clients within an iCBT platform, examining how support strategies correlate with clinical outcomes. They used k-means clustering to identify supporters whose messages were linked with "high", "medium" or "low" improvements in client outcomes, as measured by PHQ-9 and Generalized Anxiety Disorder-7 (GAD-7). The messages of more successful supporters were more positively phrased, more encouraging, more often used first person plural pronouns, were less abstract, and referenced more social behaviors. Association rule mining was then applied to linguistic features in the messages in order to identify contexts in which particular support strategies were more effective. For less engaged patients, longer, more positive and more supportive messages were linked with better outcomes. For more engaged clients, messages with less negative words, less abstraction, and more references to social behaviors were associated to better outcomes. Such results could ultimately be used in the design of supporter training materials.

One could also try to predict whether a patient engages or drops out of care. Wallert et al[89] aimed to predict adherence to an online intervention targeting symptoms of depression and anxiety in people who had experienced a myocardial infarction. The analysis included linguistic features of the homework texts as well as demographic and clinical characteristics. The strongest predictors of adherence were cardiac-related fear, gender, and the number of words in the first homework assignment.

Neurobiological treatments

Numerous neurobiological options have emerged as potential treatments for severe and treatment-resistant depression, such as transcranial magnetic stimulation (TMS) and electroconvulsive therapy (ECT). Given the potential risks and side effects of these treatments, as well as their higher financial costs, there is an especially strong interest in identifying for whom they are safe and effective[90-92].

Recent reviews have examined predictors of treatment response and relapse among depressed patients receiving TMS[92-94]. TMS studies with more female patients tend to have higher effect sizes, suggesting that gender may be a predictor of TMS outcomes[95]. Although several studies have attempted to examine neurobiological predictors of response to TMS, the findings are currently inconsistent[92]. Small sample size generally means that machine learning study designs are likely to overfit and produce results that will not replicate later.

Efforts to predict treatment outcomes for ECT are still primarily traditional association studies. Some of them identified a few variables that appear to replicate across studies. Better outcomes have been found for older patients, those with psychotic depression, those with high suicidal intent, and those who exhibit early symptom changes[90,96]. However, due to the small sample size in most ECT trials, and the typically non-randomized study designs, this area has not seen much progress. These are also obstacles to the application of machine learning techniques.

THE UTILITY OF ELECTRONIC HEALTH RECORDS, SMARTPHONE AND SOCIAL MEDIA DATA

Electronic health records (EHR) are increasingly widely adopted in health care systems. They comprise data routinely collected and maintained for individual patients over the course of their clinical care. As such, these data may be particularly useful for building predictive models in psychiatry that could be readily integrated into points of care within clinical settings[97]. EHR data can be divided into two major types: coded structured data, including diagnostic codes, procedure codes, laboratory and medication prescription codes; and unstructured data, including clinical notes and other text-based documentation, which can be mined using natural language processing.

Recent studies have tested the potential of EHR data to predict treatment outcomes in psychiatry, with the bulk of efforts to date focused on depression, though examples exist for bipolar disorder[98] and schizophrenia[99]. Machine learning-based efforts using EHR data have sought to identify those individuals who are likely to drop out after initiating antidepressants[100], those who will show a stable treatment response to antidepressants[101], and those who may transition to a bipolar diagnosis after starting antidepressants for depression[102]. Such applications have shown promising, though still modest and not yet clinically actionable, results.

Applying logistic regression and random forest approaches, Pradier et al[102] used demographic and structured EHR data (i.e., diagnostic, medication and procedure codes) available at the time of initial prescription to predict treatment dropout after initiating one of nine most common antidepressants. Although mean AUC was below 0.70, they found that incorporating EHR data significantly improved prediction of treatment dropout compared to demographic information alone, and that predictive performance varied by type of antidepressant (AUC as high as 0.80 for escitalopram) and provider type (higher accuracy among psychiatrist-treated individuals).

Hughes et al[101] applied logistic regression and extremely randomized trees with demographic and structured EHR data to predict general and drug-specific treatment continuity in patients receiving any of 11 antidepressants, observing a mean AUC of 0.63-0.66 and similar performance when evaluated at a separate site.

Where symptom score (e.g., PHQ-9) data have been available for smaller EHR cohorts (e.g., N<2,500)[103], LASSO models incorporating demographic information, structured and unstructured EHR data, and baseline symptom scores have shown modest-to-adequate performance in predicting improvements in depressive symptom severity, for both medication treatment (AUC=0.66) and psychotherapy (AUC=0.75). However, the most important predictor in these models was baseline symptom scores. Only when symptom scores are routinely integrated into EHR treatment workflows will such models be relevant for outcome prediction in large-scale health systems.

When using EHR data for predicting treatment outcomes in psychiatry, a key challenge is how to operationalize the outcome of interest using available clinical information. This usually involves establishing a set of rules around which relevant EHR features are observed, or not observed, in a cohort of patients over a defined period. For example, treatment dropout was defined by Pradier et al[100] as less than 90 days of prescription availability after index antidepressant initiation, with no evidence of alternative psychiatric treatment procedures. Antidepressant treatment stability, on the other hand, has been defined as two or more antidepressant medication prescription codes at least 30 days apart over a period of at least 90 days, with additional rules about the maximum time gap between adjacent prescription codes, and other medication possession indicators[101].

EHR data are also highly dimensional, with tens of thousands of possible diagnostic codes in addition to possible medication and procedure codes. Machine learning methods may be particularly suitable for modeling complex signals across a diverse set of EHR-based predictors, but also for reducing their dimensions prior to modeling. In their study of antidepressant treatment stability, Hughes et al[101] applied supervised topic modeling using latent

Dirichlet allocation to reduce 9,256 coded EHR features into 10 interpretable empirically derived topics, finding that a classifier for continuous treatment based on this lower-dimensional set of predictors showed comparable performance to a logistic regression based on a higher-dimensional set of features. Simpler methods, such as selecting only diagnostic codes that meet a frequency threshold in the patient population, have also been used[100].

Smartphones can provide various kinds of data that are difficult to acquire through other means. Their first and biggest feature is that they contain many sensors that can passively collect data across a variety of domains. Passive smartphone data include dynamic measures of sleep quality, exercise, heart rate, geospatial locations, language use, and communication patterns[80,104]. Machine learning methods are indispensable for dealing with complex patterns in these sensor data[105]. Currently available studies applying machine learning to predict mental health outcomes using sensor data have generally employed modest samples of 7 to 70 participants, yielding proofs-of-principle more than generalizable results[80,106-108]. Mobile phones also facilitate the collection of EMA data, allowing investigators to perform measurements at frequent intervals (e.g., several times a day). Furthermore, smartphone-based neurocognitive assessments appear to be a promising way to scalably collect cognitive data[109,110].

Few studies have used smartphone data to predict treatment outcomes. These include studies using text data from emails to predict treatment response in patients with social anxiety[111], EMA data to predict changes in self-esteem from an online intervention[112], and EMA data to predict treatment response in patients with depression[83]. In the study predicting depression outcomes, a model including EMA data did not outperform a model using baseline characteristics[83], showing that the former data do not always provide incremental value.

Social media allow investigators to access large amounts of data relating to language use and online activity. However, to our knowledge, these data have not yet been used to predict treatment responses. One of the tradeoffs between incorporating different types of data is the cost and quantity versus quality of data: very often these data present with noise which may hinder the ability to identify meaningful patterns and signals. Novel methods of topological machine learning are robust to noise, and allow to extract descriptors of the shape and structure of data that can augment performance for the analysis of intensive time-point measurements[113]. Such data with repeated measures may be useful for testing hypotheses, since sample size may compensate for the increased noise of data[114].

THE USE OF DATA FROM GENETICS, ELECTROPHYSIOLOGY, NEUROIMAGING AND COGNITIVE TESTING

Genetics

Machine learning methods are an appealing analytical approach for bridging genetic data with the prediction of treatment response in psychiatry. They put the focus on prediction rather than association, are able to detect interactions between loci, wisely handle correlation, and do not assume a pre-defined statistical model or additivity[115].

Machine learning has been used with the objective to improve prediction of treatment outcomes from genetics alone in many diseases, including cancer[116,117] and hypertension[118].

The question of whether an individual's genetic background could affect how he/she responds to medication treatment has been investigated in pharmacogenomics. An earlier study applying genome-wide complex trait analysis in a sample of roughly 3,000 depressed patients suggested that common genetic variation could explain up to 42% of observed individual differences in antidepressant treatment response[119], suggesting that modeling common genetic variation could be useful for prediction. However, results of pharmacogenomic studies have so far, in general, been underwhelming[120].

Polygenic scores are a common method for quantifying the overall contribution of common genetic variation to particular traits[121]. Polygenic associations with treatment response have been investigated in relatively small patient cohorts (most N<1000) to date, with mixed findings[122-125]. For example, polygenic scores for major depression and schizophrenia did not significantly predict antidepressant efficacy (based on symptom improvement) in classic treatment studies such as Genome-Based Therapeutic Drugs for Depression (GENDEP) and STAR*D[123]. However, these scores were built on earlier genome wide association studies (GWAS) and were likely underpowered. Well-powered GWAS of antidepressant response have produced mixed results, with one study identifying gene sets of relevance for bupropion response[126] and another observing no significant findings for antidepressant resistance[127]. Larger-scale GWAS meta-analysis efforts are needed and ongoing. Even fewer studies have examined common genetic variation associated with responses to other treatment modalities such as psychotherapy[125] or ECT[128].

DNA methylation and gene expression data have been explored in combination with phenotypic datasets of demographic and clinical variables on their ability to predict response to multiple medications. A recent review[129] pointed out genetic prediction of therapeutic outcomes in depression as the most promising[43,130-133], with an overall accuracy of 0.82 (95% CI: 0.77-0.87)[134]. Models combining multiple data types, such as peripheral gene expression data, neuroimaging and clinical variables, achieved significantly higher accuracy[134].

Tree-based approaches were the most popular machine learning methods, followed by penalized regression, support vector machines and deep learning[129]. Studies were quite heterogeneous in design, methods, implementation and validation, limiting our capacity to elucidate the extent to which machine learning integrated with genetics can predict antidepressant drug response.

Evidence for polygenic risk scores versus support vector machines for the prediction of treatment-resistant schizophrenia from GWAS data have been reviewed[135]. Although support vector machines might be more suitable to take into account complex genetic interactions, the traditional polygenic risk score approach showed higher accuracy for classifying treatment-resistant individuals[115].

Despite many efforts to use many kinds of genetic information in many different ways, results so far have not been sufficiently compelling or accurate to support the use of these approaches to guide clinical care[136,137]. In the future, until novel analytic techniques become available to extract signal from the genome, or a better understanding of the genetic basis for mental illness emerges, the most promising avenue in this context is to integrate genetic information into multivariable analyses to potentially improve broader model performance[133,137].

Electrophysiology and neuroimaging

Tailoring treatment decisions based on brain measures is intuitively appealing and empirically well-justified. Systematic reviews and meta-analyses indicate that therapeutic outcomes are often related to pre-treatment brain differences and that the brain changes as a result of therapy[138-145]. However, in previous research using traditional statistical methods, effect sizes were too low to make the jump from statistical significance to clinical relevance, external validation was rare, sample sizes were small, methodological and site-related variance was high, and in many cases the techniques were not suited to an integration into clinical routine due to their cost-benefit ratio (e.g., positron emission tomography) or reliance on experimental protocols that are unavailable in most clinical settings[138,139,143,145,146]. Machine learning approaches offer hope in overcoming these barriers to clinical implementation. Preliminary reviews comparing accuracies support this optimism by suggesting superiority for treatment prediction with respect to traditional statistical methods[134].

Early studies in this area applied machine learning to detect outcomes such as response to clozapine in psychosis[147] and to selective serotonin reuptake inhibitors (SSRIs) in depression[148-150], but the majority of research has focused on predicting brain stimulation outcomes for depression[148,151-155]. For example, Corlier et al[156] found that alpha spectral correlation could be used to measure EEG connectivity, which then predicted response to repetitive TMS (rTMS), using cross-validated logistic regression, with an accuracy of 77% in a subgroup of depressed individuals. This increased to 81% when adding clinical symptoms of depression. Most studies report predictive accuracies of >80% on the basis of pilot samples consisting of approximately 50 cases or less[155], reflecting the strong likelihood of bias and overfitting that is also seen with magnetic resonance imaging (MRI)[157].

Task-related functional MRI (fMRI) has been used for treatment prediction[158]: for example, by modelling amygdala engagement interactions with early life stress during an experimental task to predict antidepressant outcome[159] or by using fear conditioning responses to predict panic disorder treatment outcome[160,161]. Similar task-related predictive models have been built in a number of studies of CBT[162] or antidepressant responses[162-164]. In task-based fMRI, however, the translational potential is limited due to the use of lengthy and methodologically complicated experimental paradigms. Resting-state fMRI is a popular alternative, because it measures behaviourally-relevant, synchronized brain network activity at rest, and the imaging protocols can be more easily harmonized across scanners[165]. Studies in this field have demonstrated similar accuracies for CBT[166], trauma-focused psychotherapy[167], antidepressant treatment[168], and antipsychotic therapy[169], while also showing predictive accuracy for ECT[165,170].

A challenge of functional imaging is reliability across scanners, especially in non-experimental clinical settings. Structural neuroimaging may provide an opportunity for faster implementation into existing clinical routines. Most studies have involved grey matter measurements, and ECT treatment prediction has been a frequent focus, with studies using whole-brain approaches[171], regional measurements[172], and combinations of neuroimaging modalities[173]. White matter measurements (e.g., with diffusion tensor imaging) have been relatively less commonly considered.

Overall, the lack of multi-site studies and external validation reflects the pilot-study stage of research in this area, where results can be interpreted as promising but highly experimental. Whether the machine learning results will ultimately agree with the low effect sizes found with classical statistical approaches remains an open question[143,145].

Cognitive testing

Cognitive testing is a straightforward method to indirectly assess brain functioning that has been historically linked to treatment outcomes. Although such testing can be time-consuming and costly when performed by a trained neuropsychologist, more recent computerized methods can facilitate efficient digital assessments that lend themselves especially well to machine learning, including from passively collecting smartphone measurements as described above[80,114,174].

Etkin et al[175] conducted an early study in this area, as part of the international Study to Predict Optimized Treatment in Depression (iSPOT-D), aimed to predict response to antidepressant treatment using a battery of computerized cognitive tasks assessing attention, processing speed, memory, and executive and emotional functions. In order to obtain accurate predictive estimates, they first classified depressed individuals into a subgroup with particularly poor cognition before training a supervised discriminant function to predict remission. Results demonstrated that remission following escitalopram could be predicted with 72% accuracy, but this was not confirmed with sertraline or venlafaxine.

Subtyping or unsupervised learning approaches have also been helpful to identify response trajectories to cognitive training. A recent study found that self-organizing maps detecting multivariate relationships between cognitive functions associated with working memory task performance could identify individuals who differentially responded to the training[176].

HOW CLOSE WE HAVE COME TO REAL-WORLD IMPLEMENTATION

Not all prediction models will translate readily for use in clinical or other real-world settings. In evaluating the readiness of predictive models for real-world implementation, key criteria include external validation, empirical support from implementation trials, and acceptability to users (e.g., clinicians).

External cross-validation remains the gold standard for evaluating real-world performance, as it quantifies performance loss when a trained model is applied to a completely independent sample. In addition, it guards against increased researcher degrees-of-freedom that may result from the many tuning parameters of more complex machine learning methods. A review focusing on machine learning in psychotherapy research reported that only 3 of 51 studies had performed external validation[62].

Studies without external validation are at high risk of overconfidence, as demonstrated by Van Bronswijk et al[60], who developed and then tested a treatment selection model across two randomized controlled trials comparing CBT and IPT. They found that the estimated effect size for the benefit of receiving the model-recommended treatment (generated through internal cross-validation) shrunk by 77% when the model was tested using the second study's data (external validation).

Some prediction efforts using large naturalistic samples have reported positive results following external validation[65,177,178].

When a model undergoes external validation and successfully predicts outcomes, the next step towards real-world use is an implementation trial. These trials provide the most compelling evidence for the value of a decision support tool. Here, patients are usually allocated to algorithm-guided treatment (generally within a shared decision-making framework) or treatment as usual.

Trial-based efforts to evaluate the efficacy of treatment personalization tools have begun to emerge. One example is a multi-service cluster randomized trial[179], in which patients (N=951) were referred to either high- or low-intensity psychotherapy. In one arm, the choice of intensity was informed by an algorithm previously developed in a naturalistic dataset. In the other arm, most patients started on low-intensity psychotherapy and were later referred to high-intensity treatment in the case of non-response, as per usual stepped care. The study found higher depression remission rates in patients whose initial treatment was recommended by the algorithm compared to usual stepped care (52.3% vs. 45.1%, odds ratio, OR=1.40, p=0.025).

Another recent example comes from Lutz et al[180], who used archival data from an outpatient CBT clinic to develop a predictive decision support system providing therapists with treatment strategy recommendations and psychometric feedback enhanced with clinical problem-solving tools. They randomized therapist-patient dyads (N=538) to treatment as usual or to algorithm-informed treatment. They reported that, overall, outcomes for those who were randomized to the intervention did not differ from those who received usual care. However, there was significant variability in the extent to which therapists in the intervention condition followed the recommendations provided by the decision support tool. When the authors analyzed outcomes for patients whose therapists had followed the recommendations, significant benefits emerged.

Browning et al[181] conducted another trial randomizing depressed patients to either algorithm-informed care or usual care for depression. Their algorithm, called PReDicT, used information from symptom scales and behavioral tests of affective cognition to predict non-response to treatment with citalopram. After eight weeks of treatment, the rate of depressive symptom response in the PReDicT arm was 55.9%, versus 51.8% in the usual care arm (not significant, OR=1.18, p=0.25). Of all instances where the algorithm predicted non-response, only 65% prompted a change in treatment regimen, and most consisted of an increase in dosage only.

In combination, the above findings highlight that accurate algorithms are not enough to ensure the success of a decision support system for precision treatment[39]. When randomizing patients to algorithm-informed care or usual care, clinicians may override algorithm recommendations and choose alternative treatments. Patients may refuse the algorithm-recommended treatment, or have restrictions to its use that were not contemplated by the decision support tool (e.g., prohibitive cost of therapy). In light of this, effect sizes for these interventions will often vary when applied in different settings.

The use of predictive models may be uniquely challenging in psychotherapy research and practice. One challenge is that a given therapist is only trained to provide a limited subset of psychotherapies. Whereas a psychiatrist may be qualified to prescribe a large number of different medications or medication combinations, a psychotherapist is less likely to be able to competently provide many different psychotherapies. Another consideration is that predictions from a model may lead to self-fulfilling prophecies, in which clinicians treat "easy" patients (those with good prognoses) differently than "difficult" patients[182].

For both medications and psychotherapies, in real-world, treatment decisions are rarely going to be made solely based on model recommendations. Rather, these decisions will involve the preferences of patients, the recommendations of clinicians, the availability and costs of treatments, and several other considerations[183]. As such, the development of data-driven decision tools should be informed by extensive consultation and co-production with the intended users, in order to implement models that maximize acceptability and compatibility with other clinical guidelines (i.e., risk management procedures, norms about safe dosage or titration of medications).

Another crucial barrier to implementation is the interpretability of machine learning models. As algorithms become increasingly complex, sometimes called "black box" algorithms, they can become very difficult to interpret, and therefore unlikely to be acceptable to clinical users. Methods for explaining predictions of complex models have therefore been developed[184,185], but there is currently no agreed-upon measure for assessing the quality or accuracy of these explanations. In addition, black-box predictive models combined with (similarly complex) explanatory methods may yield complicated decision pathways that increase the likelihood of human error[186].

In order to ensure that algorithm recommendations are used in trials, additional thought and effort must be devoted to issues of dissemination and implementation, with the goal of making the recommendations simple to generate, easy to understand, trustworthy, ethical, cost-effective, and compelling enough to influence the decision-maker(s)[187].

A recent experiment was conducted with 220 antidepressant-prescribing clinicians to assess the impact of providing machine learning recommendations and accompanying explanations[188]. It was found that recommendations did not improve accurate selection of antidepressants in hypothetical patient scenarios, and that accuracy was even lower when incorrect recommendations were presented than when standard information was available. Prospective field-tests[181,189] are one method for identifying the myriad institutional, cultural and contextual factors that could affect the uptake and sustained use of a precision psychiatry tool, aiming to co-produce acceptable and interpretable decision tools with the intended users.

ETHICAL CHALLENGES

From the development of machine learning tools to their potential deployment into clinical care, we can identify several ethical challenges[190-193].

The first challenge concerns responsibility. With the implementation of machine learning programs into clinical practice, physicians and machine learning-based tools would become "teammates" that collaborate in selecting an optimal treatment[194,195]. In such a scenario, who will hold authority and ethical responsibility over the decision made? We believe that a competent human agent should check and take final responsibility on the machine learning-based suggestions[196], as only he/she is equipped with empathy, a good understanding of the contextual environment and, most uniquely, consciousness.

The second challenge is to avoid dehumanization[197]. Machine learning can incorporate a great variety of psychological, environmental and social variables, and there is some progress towards including subjective patient experience into machine learning models[198]. However, giving a patient the space to articulate his/her concerns is essential to ensure accurate diagnosis, health outcomes, and humane care[199].

Third, making decisions is an intricate part of physicians' activity. The non-expert tends to act as a "technician" and more likely relies on protocols, whilst the expert, after the observation of many cases, is more prone to making decisions based on tacit knowledge[200-202]. The ethical mandate is that practitioners use all of their capabilities, including those based on self-experience and observation, even if this is in discordance with a statistical model. Disagreements between physicians and machine learning-based decisions may lead to consultations with other clinicians[193]. However, in the context of modern health care systems, respecting clinicians' judgement is vital[193,203], and they should not be forced to act against their own criteria (freedom of action)[204].

Practitioners (especially those with less expertise) might be in danger of not developing/losing their own clinical judgement and become dependent on automatically deployed machine learning outcomes[205], particularly for those complex cases that they fear they are not competent enough to solve. This would risk disempowerment of clinicians. On the other hand, it is a physicians' duty to train themselves in the use, understanding and interpretation of machine learning applications, so that they can trust the system and its outputs, and contribute to patients' acceptance[206].

Machine learning tools need to be transparent to the human teammates to facilitate understanding[194,207]. The idea of transparency is opposite to that of "black-box" machine learning algorithms, in which the patterns the algorithm follows to make a decision for a given patient are opaque to the person and even to the developer, making very challenging (if not impossible) for the affected person to understand how the system worked out an output for him/her[190]. This risks not only increasing clinicians' resistance to use the tool, but also disempowering patients and disrespecting their autonomy. Developers should consider simpler algorithms that balance interpretability with accuracy[191].

Furthermore, a central issue in fair machine learning development arises when the training dataset is not a good representation of the phenomenon being studied[192,208]. A model trained in such data will predict erroneous outcomes for groups that were underrepresented[209]. For example, a widely used machine learning algorithm assigned the same level of disease risk to Black and White patients, even if Black patients were sicker than White patients[210]. As a consequence, the system was actively causing harm to Black patients by leading to allocation of fewer resources to them. Potentially discriminatory predictors should be left out of the model, but developers should be aware that surrogate variables correlated with the excluded set might still become relevant for prediction. Objective unbiased applications might help reduce discrimination in machine learning[211,212].

Finally, the risk of misuse of personal and sensitive data exchanged in machine learning is high[213]. For this reason, machine learning tools can be only used when data security and privacy are guaranteed.

CONCLUSIONS

This paper reviews several studies suggesting that it is possible to predict outcomes and personalize psychiatric treatment by using machine learning. Several gold standard prediction studies have shown that we can predict whether a depressed patient will respond to specific antidepressants[40,41], to specific psychotherapeutic techniques[177], and whether patients with first episode psychosis will have good prognosis after one year with certain antipsychotic medications[25,50]. At least three predictive models have even been tested in prospective clinical trials.

Despite this progress, the potential for machine learning in psychiatry has just begun to be explored. Predicting treatment response is just one relatively narrow use case where machine learning can add value and improve mental health care. Prediction can help with so many more clinical decisions and clinical processes. We could predict barriers that prevent an individual from engaging in care initially, or non-adherence or dropout from care after initiation. We could streamline patients to the appropriate level of care, such as self-guided programs vs. outpatient care, or intensive outpatient versus inpatient care, to maximize scarce health care resources. In selecting a specific treatment approach, we could optimize dosing or predict side effect profiles in order to improve symptoms but minimize impact on patient quality of life. Some psychiatric treatments carry high cost (e.g., ketamine, ECT) or unwanted side effects (e.g., metabolic disruption and weight gain for antipsychotics). Doing no harm is arguably more important than improving the probability of recovery, and so precision mental health efforts could be especially important in identifying which treatments are safest and most tolerable.

Machine learning could even help sequence treatments over time, or design specific treatment protocols for an individual. For example, modular psychological interventions can be personalized[66,68], or tailored health behavior change interventions can be customized for an individual. This form of personalization and customization has proven effective in contexts like smoking cessation, breast cancer screening, and physical activity[214,215].

Techniques like natural language processing, often using machine learning algorithms, give us the ability to draw insights from text-based data – e.g., social media posts, peer-support conversations, or conversation transcriptions – that might inform the content that is offered to an individual as part of his/her treatment to maximize future outcomes. In addition, the same analytic techniques may form the basis of interventions, such as chatbots, that could provide scalable support for loneliness, stress, or other subclinical psychological issues when human support is unavailable or not clinically warranted. This personalization of iCBT treatment may be particularly necessary for unguided interventions, where non-adherence is widespread and undermines the potential for symptom relief.

Machine learning is a powerful tool that can help sift through multi-modal predictors and model their complex/non-linear contributions. And it can identify specific subtypes of patients, e.g., through clustering, for more nuanced prediction of treatment outcomes. Machine learning techniques are allowing us to extract more knowledge from bigger datasets in a more efficient way – which is a good and promising thing.

However, the ultimate goal of psychiatry is to better treat mental illness. The path toward machine learning improving psychiatric care in real-life settings is not only governed by statistical, but also by implementation considerations. Recent seminal findings[180,181] highlight that accurate algorithms alone are not enough to ensure the success of a decision support system for precision treatment. This is because many things change in the transition from a research setting into real patient care[39]. In practice, clinicians may override algorithm recommendations and choose alternative treatments. Patients may refuse the algorithm-recommended treatment, or have restrictions to its use that were not contemplated by the decision support tool. Recommendations may be provided in a poorly-designed user interface, and thus may go unseen or be actively ignored. All of these factors contribute to a general phenomenon of reduced effect sizes when an algorithm is implemented in clinical practice.

In our own personal experience, patient concerns around privacy are a very real problem. Because mental health is particularly sensitive, capturing personal data can be challenging and we need to innovate ways of collecting these data so that we do not have a biased perspective of the landscape due to a poor sampling within certain groups. Data needs to be collected in such a way that participants are aware of how and for what purposes those data will be used[216].

Technology systems must implement careful logging processes to examine concept or data drift, where the underlying distribution of a predictor or an outcome changes over time, and to ensure that the inputs and outputs of the system are auditable. This is a collective exercise of building trust in predictive models and how these will be potentially used to enhance patient outcomes, and can avoid the introduction of harm or biases in decision-making processes.

This paper reviews many kinds of data that have been used to predict treatment outcomes in psychiatry. Ultimately, treatment responses emerge from multiple interacting biological, psychological and social factors. Therefore, in theory, multi-modal approaches using demographic, clinical and brain variables should result in the most accurate predictions[217]. However, to this date, it is clear that certain kinds of data – specifically sociodemographic, self-report, psychosocial and clinical data – consistently offer more meaningful and generalizable predictions. Other types of data that might be more scientifically appealing – such as neuroimaging and genetic data – have not yet shown compelling results in a large external sample, let alone in prospective implementation studies.

Ultimately, data types that can be easily integrated into clinical care in a cost-effective and ethical way, which is appropriate for the prevalence and invasiveness of the therapy, are most likely to show favorable return on investment for ultimate decision makers in health systems and health payers.

ACKNOWLEDGEMENTS

A.M. Chekroud holds equity in Spring Care Inc., Carbon Health Technologies Inc., and UnitedHealthcare. J. Bondar is an employee of Spring Care Inc. G. Doherty is co-founder of SilverCloud Health. R. Iniesta is supported by the Brain & Behavior Research Foundation and the National Institute for Health Research (NIHR) Maudsley Biomedical Research Centre.

REFERENCES

1. Rush AJ, Trivedi MH, Wisniewski SR et al. Acute and longer-term outcomes in depressed outpatients requiring one or several treatment steps: a STAR*D report. Am J Psychiatry 2006;163:1905-17.
2. Hastie T, Tibshirani R, Friedman J. The elements of statistical learning: data mining, inference, and prediction, 2nd ed. New York: Springer, 2009.
3. Chekroud AM. Bigger data, harder questions – opportunities throughout mental health care. JAMA Psychiatry 2017;64:44-50.
4. Iniesta R, Stahl D, McGuffin P. Machine learning, statistical learning and the future of biological research in psychiatry. Psychol Med 2016;46:2455-65.
5. Breiman L. Statistical modeling: the two cultures (with comments and a rejoinder by the author). Stat Sci 2001;16:199-231.

6. Delgadillo J. Machine learning: a primer for psychotherapy researchers. Psychother Res 2021;31:1-4.
7. Sidey-Gibbons JAM, Sidey-Gibbons CJ. Machine learning in medicine: a practical introduction. BMC Med Res Methodol 2019;19:64.
8. Cohen ZD, DeRubeis RJ. Treatment selection in depression. Annu Rev Clin Psychol 2018;14:209-36.
9. Kessler RC, van Loo HM, Wardenaar KJ et al. Using patient self-reports to study heterogeneity of treatment effects in major depressive disorder. Epidemiol Psychiatr Sci 2017;26:22-36.
10. Maj M, Stein DJ, Parker G et al. The clinical characterization of the adult patient with depression aimed at personalization of management. World Psychiatry 2020;19:269-93.
11. Perlis RH. A clinical risk stratification tool for predicting treatment resistance in major depressive disorder. Biol Psychiatry 2013;74:7-14.
12. Gravesteijn BY, Nieboer D, Ercole A et al. Machine learning algorithms performed no better than regression models for prognostication in traumatic brain injury. J Clin Epidemiol 2020;122:95-107.
13. Nusinovici S, Tham YC, Yan MYC et al. Logistic regression was as good as machine learning for predicting major chronic diseases. J Clin Epidemiol 2020;122:56-69.
14. Christodoulou E, Ma J, Collins GS et al. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. J Clin Epidemiol 2019;110:12-22.
15. Desai RJ, Wang SV, Vaduganathan M et al. Comparison of machine learning methods with traditional models for use of administrative claims with electronic medical records to predict heart failure outcomes. JAMA Netw Open 2020;3:e1918962.
16. Lynam AL, Dennis JM, Owen KR et al. Logistic regression has similar performance to optimised machine learning algorithms in a clinical setting: application to the discrimination between type 1 and type 2 diabetes in young adults. Diagn Progn Res 2020;4:6.
17. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. J Stat Softw 2010;33:1-22.
18. Strobl C, Malley J, Tutz G. An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. Psychol Meth 2009;14:323.
19. Bühlmann P. Bagging, boosting and ensemble methods. In: Gentle J, Härdle W, Mori Y (eds). Berlin: Springer, 2012:985-1022.
20. van der Laan MJ, Polley EC, Hubbard AE et al. Super learner. Stat Appl Genet Mol Biol 2007;6:4765-74.
21. Kessler RC, van Loo HM, Wardenaar KJ et al. Testing a machine-learning algorithm to predict the persistence and severity of major depressive disorder from baseline self-reports. Mol Psychiatry 2016;21:1366-71.
22. Pearson R, Pisner D, Meyer B et al. A machine learning ensemble to predict treatment outcomes following an Internet intervention for depression. Psychol Med 2019;49:2330-41.
23. Webb CA, Trivedi MH, Cohen ZD et al. Personalized prediction of antidepressant v. placebo response: evidence from the EMBARC study. Psychol Med 2019;49:1118-27.
24. Hilbert K, Kunas SL, Lueken U et al. Predicting cognitive behavioral therapy outcome in the outpatient sector based on clinical routine data: a machine learning approach. Behav Res Ther 2020;124:103530.
25. Koutsouleris N, Kahn RS, Chekroud AM et al. Multisite prediction of 4-week and 52-week treatment outcomes in patients with first-episode psychosis: a machine learning approach. Lancet Psychiatry 2016;3:935-46.
26. Fernández-Delgado M, Cernadas E, Barro S et al. Do we need hundreds of classifiers to solve real world classification problems? J Mach Learn Res 2014;15:3133-81.
27. Gacto MJ, Soto-Hidalgo JM, Alcalá-Fdez J et al. Experimental study on 164 algorithms available in software tools for solving standard non-linear regression problems. IEEE Access 2019;7:108916-39.
28. Wainer J. Comparison of 14 different families of classification algorithms on 115 binary datasets. arXiv 2016;1606.00930.
29. Zhang C, Liu C, Zhang X et al. An up-to-date comparison of state-of-the-art classification algorithms. Expert Syst Appl 2017;82:128-50.
30. Hand DJ. Classifier technology and the illusion of progress. Stat Sci 2006;21:1-14.
31. Holte RC. Very simple classification rules perform well on most commonly used datasets. Mach Learn 1993;11:63-90.
32. Yarkoni T, Westfall J. Choosing prediction over explanation in psychology: lessons from machine learning. Perspect Psychol Sci 2017;12:1100-22.
33. de Rooij M, Pratiwi BC, Fokkema M et al. The early roots of statistical learning in the psychometric literature: a review and two new results. arXiv 2018;1911.11463.
34. Larson SC. The shrinkage of the coefficient of multiple correlation. J Educ Psychol 1931;22:45-55.
35. Mosier CI. I. Problems and designs of cross-validation 1. Educ Psychol Meas 1951;11:5-11.
36. Wainer H. Estimating coefficients in linear models: it don't make no nevermind. Psychol Bull 1976;83:213-7.
37. Delgadillo J, Lutz W. A development pathway towards precision mental health care. JAMA Psychiatry 2020;77:889-90.
38. Browning M, Carter CS, Chatham C et al. Realizing the clinical potential of computational psychiatry: report from the Banbury Center meeting, February 2019. Biol Psychiatry 2020;88:e5-10.
39. Chekroud AM, Koutsouleris N. The perilous path from publication to practice. Mol Psychiatry 2018;23:24-5.
40. Chekroud AM, Zotti RJ, Shehzad Z et al. Cross-trial prediction of treatment outcome in depression: a machine learning approach. Lancet Psychiatry 2016;3:243-50.
41. Nie Z, Vairavan S, Narayan VA et al. Predictive modeling of treatment resistant depression using data from STAR*D and an independent clinical study. PLoS One 2018;13:e0197268.
42. Iniesta R, Malki K, Maier W et al. Combining clinical variables to optimize prediction of antidepressant treatment outcomes. J Psychiatr Res 2016;78:94-102.
43. Iniesta R, Hodgson K, Stahl D et al. Antidepressant drug-specific prediction of depression treatment outcomes from genetic and clinical variables. Sci Rep 2018;8:5530.
44. Drysdale AT, Grosenick L, Downar J et al. Resting-state connectivity biomarkers define neurophysiological subtypes of depression. Nat Med 2017;23:28-38.
45. Dinga R, Schmaal L, Penninx B et al. Evaluating the evidence for biotypes of depression: methodological replication and extension of. Neuroimage Clin 2019;22:101796.
46. Chekroud AM, Gueorguieva R, Krumholz HM et al. Reevaluating the efficacy and predictability of antidepressant treatments: a symptom clustering approach. JAMA Psychiatry 2017;74:370-8.
47. Bondar J, Caye A, Chekroud AM et al. Symptom clusters in adolescent depression and differential response to treatment: a secondary analysis of the Treatment for Adolescents with Depression Study randomised trial. Lancet Psychiatry 2020;7:337-43.
48. Gueorguieva R, Chekroud AM, Krystal JH. Trajectories of relapse in randomised, placebo-controlled trials of treatment discontinuation in major depressive disorder: an individual patient-level data meta-analysis. Lancet Psychiatry 2017;4:230-7.
49. Paul R, Andlauer Till FM, Czamara D et al. Treatment response classes in major depressive disorder identified by model-based clustering and validated by clinical prediction models. Transl Psychiatry 2019;9:187.
50. Leighton SP, Upthegrove R, Krishnadas R et al. Development and validation of multivariable prediction models of remission, recovery, and quality of life outcomes in people with first episode psychosis: a machine learning approach. Lancet Digit Health 2019;1:e261-70.
51. Nunes A, Ardau R, Berghöfer A et al. Prediction of lithium response using clinical data. Acta Psychiatr Scand 2020;141:131-41.
52. Kim TT, Dufour S, Xu C et al. Predictive modeling for response to lithium and quetiapine in bipolar disorder. Bipolar Disord 2019;21:428-36.
53. Lutz W, Leach C, Barkham M et al. Predicting change for individual psychotherapy clients on the basis of their nearest neighbors. J Consult Clin Psychol 2005;73:904-13.
54. Lambert MJ, Bergin AE. Bergin and Garfield's handbook of psychotherapy and behavior change. Chichester: Wiley, 2021.
55. Deisenhofer AK, Delgadillo J, Rubel JA et al. Individual treatment selection for patients with posttraumatic stress disorder. Depress Anxiety 2018;35:541-50.
56. DeRubeis RJ, Cohen ZD, Forand NR et al. The personalized advantage index: translating research on prediction into individualized treatment recommendations. A demonstration. PLoS One 2014;9:1-8.
57. Cloitre M, Petkova E, Su Z et al. Patient characteristics as a moderator of posttraumatic stress disorder treatment outcome: combining symptom burden and strengths. BJPsych Open 2016;2:101-6.
58. Wallace ML, Frank E, Kraemer HC. A novel approach for developing and interpreting treatment moderator profiles in randomized clinical trials. JAMA Psychiatry 2013;70:1241-7.
59. Keefe JR, Wiltsey Stirman S, Cohen ZD et al. In rape trauma PTSD, patient characteristics indicate which trauma-focused treatment they are most likely to complete. Depress Anxiety 2018;35:330-8.

60. van Bronswijk SC, Bruijniks SJE, Lorenzo-Luaces L et al. Cross-trial prediction in psychotherapy: external validation of the Personalized Advantage Index using machine learning in two Dutch randomized trials comparing CBT versus IPT for depression. Psychother Res 2021;31:78-91.
61. Cohen ZD, Kim TT, Van HL et al. A demonstration of a multi-method variable selection approach for treatment selection: recommending cognitive-behavioral versus psychodynamic therapy for mild to moderate adult depression. Psychother Res 2020;30:137-50.
62. Aafjes-van Doorn K, Kamsteeg C, Bate J et al. A scoping review of machine learning in psychotherapy research. Psychother Res 2021;31:92-116.
63. Luedtke A, Sadikova E, Kessler RC. Sample size requirements for multivariate models to predict between-patient differences in best treatments of major depressive disorder. Clin Psychol Sci 2019;7:445-61.
64. Lorenzo-Luaces L, DeRubeis RJ, van Straten A et al. A prognostic index (PI) as a moderator of outcomes in the treatment of depression: a proof of concept combining multiple variables to inform risk-stratified stepped care models. J Affect Disord 2017;213:78-85.
65. Delgadillo J, Huey D, Bennett H et al. Case complexity as a guide for psychological treatment selection. J Consult Clin Psychol 2017;85:835-53.
66. Fisher AJ, Bosley HG, Fernandez KC et al. Open trial of a personalized modular treatment for mood and anxiety. Behav Res Ther 2019;116:69-79.
67. Lutz W, Rubel JA, Schwartz B et al. Towards integrating personalized feedback research into clinical practice: development of the Trier Treatment Navigator (TTN). Behav Res Ther 2019;120:103438.
68. Rubel JA, Fisher AJ, Husen K et al. Translating person-specific network models into personalized treatments: development and demonstration of the dynamic assessment treatment algorithm for individual networks (DATA-IN). Psychother Psychosom 2018;87:249-51.
69. Delgadillo J, Rubel J, Barkham M. Towards personalized allocation of patients to therapists. J Consult Clin Psychol 2020;88:799-808.
70. Atkins DC, Steyvers M, Imel ZE et al. Scaling up the evaluation of psychotherapy: evaluating motivational interviewing fidelity via statistical text classification. Implement Sci 2014;9:49.
71. Goldberg SB, Tanana M, Imel ZE et al. Can a computer detect interpersonal skills? Using machine learning to scale up the Facilitative Interpersonal Skills task. Psychother Res 2021;31:281-8.
72. Ewbank MP, Cummins R, Tablan V et al. Quantifying the association between psychotherapy content and clinical outcomes using deep learning. JAMA Psychiatry 2020;77:35-43.
73. Ewbank MP, Cummins R, Tablan V et al. Understanding the relationship between patient language and outcomes in internet-enabled cognitive behavioural therapy: a deep learning approach to automatic coding of session transcripts. Psychother Res 2020;31:326-38.
74. de Jong K, Conijn JM, Gallagher RAV et al. Using progress feedback to improve outcomes and reduce drop-out, treatment duration, and deterioration: a multilevel meta-analysis. Clin Psychol Rev 2021;85:102002.
75. Buckman JEJ, Cohen ZD, O'Driscoll C et al. Predicting prognosis for adults with depression using individual symptom data: a comparison of modelling approaches. Open Sci Framework (in press).
76. Green SA, Honeybourne E, Chalkley SR et al. A retrospective observational analysis to identify patient and treatment-related predictors of outcomes in a community mental health programme. BMJ Open 2015;5:e006103.
77. Andersson G, Titov N, Dear BF et al. Internet-delivered psychological treatments: from innovation to implementation. World Psychiatry 2019;18: 20-8.
78. Andrews G, Basu A, Cuijpers P et al. Computer therapy for the anxiety and depression disorders is effective, acceptable and practical health care: an updated meta-analysis. J Anxiety Disord 2018;55:70-8.
79. Nahum-Shani I, Smith SN, Spring BJ et al. Just-in-Time Adaptive Interventions (JITAIs) in mobile health: key components and design principles for ongoing health behavior support. Ann Behav Med 2018;52:446-62.
80. Mohr DC, Zhang M, Schueller SM. Personal sensing: understanding mental health using ubiquitous sensors and machine learning. Annu Rev Clin Psychol 2017;13:23-47.
81. Lenhard F, Sauer S, Andersson E et al. Prediction of outcome in internet-delivered cognitive behaviour therapy for paediatric obsessive-compulsive disorder: a machine learning approach. Int J Methods Psychiatr Res 2018;27:1-11.
82. Flygare O, Enander J, Andersson E et al. Predictors of remission from body dysmorphic disorder after internet-delivered cognitive behavior therapy: a machine learning approach. BMC Psychiatry 2020;20:1-9.
83. van Breda W, Bremer V, Becker D et al. Predicting therapy success for treatment as usual and blended treatment in the domain of depression. Internet Interv 2018;12:100-4.
84. Yardley L, Spring BJ, Riper H et al. Understanding and promoting effective engagement with digital behavior change interventions. Am J Prev Med 2016;51:833-42.
85. Morrison C, Doherty G. Analyzing engagement in a web-based intervention platform through visualizing log-data. J Med Internet Res 2014;16:e252.
86. Chien I, Enrique A, Palacios J et al. A machine learning approach to understanding patterns of engagement with internet-delivered mental health interventions. JAMA Netw Open 2020;3:e2010791.
87. Provoost S, Ruwaard J, van Breda W et al. Validating automated sentiment analysis of online cognitive behavioral therapy patient texts: an exploratory study. Front Psychol 2019;10:1-12.
88. Chikersal P, Belgrave D, Doherty G et al. Understanding client support strategies to improve clinical outcomes in an online mental health intervention. Presented at the ACM Conference on Human Factors in Computing Systems, Honolulu, April 2020.
89. Wallert J, Gustafson E, Held C et al. Predicting adherence to internet-delivered psychotherapy for symptoms of depression and anxiety after myocardial infarction: machine learning insights from the U-CARE heart randomized controlled trial. J Med Internet Res 2018;20:e10754.
90. Pinna M, Manchia M, Oppo R et al. Clinical and biological predictors of response to electroconvulsive therapy (ECT): a review. Neurosci Lett 2018; 669:32-42.
91. Haq AU, Sitzmann AF, Goldman ML et al. Response of depression to electroconvulsive therapy: a meta-analysis of clinical predictors. J Clin Psychiatry 2015;76:1374-84.
92. Kar SK. Predictors of response to repetitive transcranial magnetic stimulation in depression: a review of recent updates. Clin Psychopharmacol Neurosci 2019;17:25-33.
93. Miljevic A, Bailey NW, Herring SE et al. Potential predictors of depressive relapse following repetitive transcranial magnetic stimulation: a systematic review. J Affect Disord 2019;256:317-23.
94. Nord CL. Predicting response to brain stimulation in depression: a roadmap for biomarker discovery. Curr Behav Neurosci Rep 2021;8:11-9.
95. Kedzior KK, Azorina V, Reitz S. More female patients and fewer stimuli per session are associated with the short-term antidepressant properties of repetitive transcranial magnetic stimulation (rTMS): a meta-analysis of 54 sham-controlled studies published between 1997-2013. Neuropsychiatr Dis Treat 2014; 10:727-56.
96. Yao Z, McCall WV, Essali N, et al. Precision ECT for major depressive disorder: a review of clinical factors, laboratory, and physiologic biomarkers as predictors of response and remission. Pers Med Psychiatry 2019;17-18:23-31.
97. Smoller JW. The use of electronic health records for psychiatric phenotyping and genomics. Am J Med Genet 2018;177:601-12.
98. Hayes JF, Marston L, Walters K et al. Lithium vs. valproate vs. olanzapine vs. quetiapine as maintenance monotherapy for bipolar disorder: a population-based UK cohort study using electronic health records. World Psychiatry 2016;15:53-8.
99. Wu C-S, Luedtke AR, Sadikova E et al. Development and validation of a machine learning individualized treatment rule in first-episode schizophrenia. JAMA Netw Open 2020;3:e1921660.
100. Pradier MF, McCoy TH Jr, Hughes M et al. Predicting treatment dropout after antidepressant initiation. Transl Psychiatry 2020;10:60.
101. Hughes MC, Pradier MF, Ross AS et al. Assessment of a prediction model for antidepressant treatment stability using supervised topic models. JAMA Netw Open 2020;3:e205308.
102. Pradier MF, Hughes MC, McCoy TH et al. Predicting change in diagnosis from major depression to bipolar disorder after antidepressant initiation. Neuropsychopharmacology 2021;46:455-61.
103. Huang SH, LePendu P, Iyer SV et al. Toward personalizing treatment for depression: Predicting diagnosis and severity. J Am Med Inform Assoc 2014; 21:1069-75.
104. Hallgren KA, Bauer AM, Atkins DC. Digital technology and clinical decision making in depression treatment: current findings and future opportunities. Depress Anxiety 2017;34:494-501.
105. Willetts M, Hollowell S, Aslett L et al. Statistical machine learning of sleep and physical activity phenotypes from sensor data in 96,220 UK Biobank participants. Sci Rep 2018;8:7961.
106. Gravenhorst F, Muaremi A, Bardram J et al. Mobile phones as medical devices in mental disorder treatment: an overview. Pers Ubiquitous Comput 2015;19:335-53.
107. Saeb S, Lattie EG, Schueller SM et al. The relationship between mobile phone location sensor data and depressive symptom severity. PeerJ 2016;4:e2537.

108. Thakur SS, Roy RB. Predicting mental health using smart-phone usage and sensor data. J Ambient Intell Humaniz Comput (in press)..
109. Germine L, Reinecke K, Chaytor NS. Digital neuropsychology: challenges and opportunities at the intersection of science and software. Clin Neuropsychol 2019;33:271-86.
110. Passell E, Dillon GT, Baker JT et al. Digital cognitive assessment: results from the TestMyBrain NIMH Research Domain Criteria (RDoC) Field Test Battery Report. PsyArXiv 2019;10.31234.
111. Hoogendoorn M, Berger T, Schulz A et al. Predicting social anxiety treatment outcome based on therapeutic email conversations. IEEE J Biomed Health Inform 2017;21:1449-59.
112. Bremer V, Funk B, Riper H. Heterogeneity matters: predicting self-esteem in online interventions based on ecological momentary assessment data. Depress Res Treat 2019;2019:3481624.
113. Riihimäki H, Chachólski W, Theorell J et al. A topological data analysis based classification method for multiple measurements. BMC Bioinformatics 2020;21:336.
114. Gillan CM, Rutledge RB. Smartphones and the neuroscience of mental health. Annu Rev Neurosci (in press).
115. Vivian-Griffiths T, Baker E, Schmidt KM et al. Predictive modeling of schizophrenia from genomic data: comparison of polygenic risk score with kernel support vector machines approach. Am J Med Genet B Neuropsychiatr Genet 2019;180:80-5.
116. Kourou K, Exarchos TP, Exarchos KP et al. Machine learning applications in cancer prognosis and prediction. Comput Struct Biotechnol J 2015;13:8-17.
117. Chabon JJ, Hamilton EG, Kurtz DM et al. Integrating genomic features for non-invasive early lung cancer detection. Nature 2020;580:245-51.
118. Iniesta R, Campbell D, Venturini C et al. Gene variants at loci related to blood pressure account for variation in response to antihypertensive drugs between black and white individuals: genomic precision medicine may dispense with ethnicity. Hypertension 2019;74:614-22.
119. Tansey KE, Guipponi M, Hu X et al. Contribution of common genetic variants to antidepressant response. Biol Psychiatry 2013;73:679-82.
120. Lawrie SM. Clinical risk prediction in schizophrenia. Lancet Psychiatry 2014; 1:406-8.
121. Zheutlin AB, Ross DA. Polygenic risk scores: what are they good for? Biol Psychiatry 2018;83:e51-3.
122. Amare AT, Schubert KO, Tekola-Ayele F et al. Association of the polygenic scores for personality traits and response to selective serotonin reuptake inhibitors in patients with major depressive disorder. Front Psychiatry 2018;9:65.
123. García-González J, Tansey KE, Hauser J et al. Pharmacogenetics of antidepressant response: a polygenic approach. Prog Neuropsychopharmacol Biol Psychiatry 2017;75:128-34.
124. Ward J, Graham N, Strawbridge R et al. Polygenic risk scores for major depressive disorder and neuroticism as predictors of antidepressant response: meta-analysis of three treatment cohorts. bioRxiv 2018:295717.
125. Andersson E, Crowley JJ, Lindefors N et al. Genetics of response to cognitive behavior therapy in adults with major depression: a preliminary report. Mol Psychiatry 2019;24:484-90.
126. Li QS, Tian C, Seabrook GR et al. Analysis of 23andMe antidepressant efficacy survey data: implication of circadian rhythm and neuroplasticity in bupropion response. Transl Psychiatry 2016;6:e889.
127. Wigmore EM, Hafferty JD, Hall LS et al. Genome-wide association study of antidepressant treatment resistance in a population-based cohort using health service prescription data and meta-analysis with GENDEP. Pharmacogenomics J 2020;20:329-41.
128. Baune BT, Soda T, Sullivan PF et al. The Genomics of Electroconvulsive Therapy International Consortium (GenECT-ic). Lancet Psychiatry 2019;6:e23.
129. Lin E, Lin C-H, Lane H-Y. Precision psychiatry applications with pharmacogenomics: artificial intelligence and machine learning approaches. Int J Mol Sci 2020;21:969.
130. Kautzky A, Baldinger P, Souery D et al. The combined effect of genetic polymorphisms and clinical parameters on treatment outcome in treatment-resistant depression. Eur Neuropsychopharmacol 2015;25:441-53.
131. Maciukiewicz M, Marshe VS, Hauschild A-C et al. GWAS-based machine learning approach to predict duloxetine response in major depressive disorder. J Psychiatr Res 2018;99:62-8.
132. Athreya AP, Neavin D, Carrillo-Roa T et al. Pharmacogenomics-driven prediction of antidepressant treatment outcomes: a machine-learning approach with multi-trial replication. Clin Pharmacol Ther 2019;106:855-65.
133. Lin E, Kuo P-H, Liu Y-L et al. A deep learning approach for predicting antidepressant response in major depression using clinical and genetic biomarkers. Front Psychiatry 2018;9:290.
134. Lee Y, Ragguett RM, Mansur RB et al. Applications of machine learning algorithms to predict therapeutic outcomes in depression: a meta-analysis and systematic review. J Affect Disord 2018;241:519-32.
135. Pisanu C, Squassina A. Treatment-resistant schizophrenia: insights from genetic studies and machine learning approaches. Front Pharmacol 2019; 10:1-7.
136. Zubenko GS, Sommer BR, Cohen BM. On the marketing and use of pharmacogenetic tests for psychiatric treatment. JAMA Psychiatry 2018;75:769.
137. Zeier Z, Carpenter LL, Kalin NH et al. Clinical implementation of pharmacogenetic decision support tools for antidepressant drug prescribing. Am J Psychiatry 2018;175:873-86.
138. Enneking V, Leehr EJ, Dannlowski U et al. Brain structural effects of treatments for depression and biomarkers of response: a systematic review of neuroimaging studies. Psychol Med 2020;50:187-209.
139. Colvonen PJ, Glassman LH, Crocker LD et al. Pretreatment biomarkers predicting PTSD psychotherapy outcomes: a systematic review. Neurosci Biobehav Rev 2017;75:140-56.
140. Fonseka TM, MacQueen GM, Kennedy SH. Neuroimaging biomarkers as predictors of treatment outcome in major depressive disorder. J Affect Disord 2018;233:21-35.
141. Fu CH, Steiner H, Costafreda SG. Predictive neural biomarkers of clinical response in depression: a meta-analysis of functional and structural neuroimaging studies of pharmacological and psychological therapies. Neurobiol Dis 2013;52:75-83.
142. Lueken U, Zierhut KC, Hahn T et al. Neurobiological markers predicting treatment response in anxiety disorders: a systematic review and implications for clinical application. Neurosci Biobehav Rev 2016;66:143-62.
143. Molent C, Olivo D, Wolf RC et al. Functional neuroimaging in treatment resistant schizophrenia: a systematic review. Neurosci Biobehav Rev 2019; 104:178-90.
144. Tarcijonas G, Sarpal DK. Neuroimaging markers of antipsychotic treatment response in schizophrenia: an overview of magnetic resonance imaging studies. Neurobiol Dis 2019;131:104209.
145. Phillips ML, Chase HW, Sheline YI et al. Identifying predictors, moderators, and mediators of antidepressant response in major depressive disorder: neuroimaging approaches. Am J Psychiatry 2015;172:124-38.
146. Lueken U, Hahn T. Functional neuroimaging of psychotherapeutic processes in anxiety and depression: from mechanisms to predictions. Curr Opin Psychiatry 2016;29:25-31.
147. Khodayari-Rostamabad A, Hasey GM, Maccrimmon DJ et al. A pilot study to determine whether machine learning methodologies using pre-treatment electroencephalography can predict the symptomatic response to clozapine therapy. Clin Neurophysiol 2010;121:1998-2006.
148. Erguzel TT, Ozekes S, Gultekin S et al. Neural network based response prediction of rTMS in major depressive disorder using QEEG cordance. Psychiatry Investig 2015;12:61-5.
149. Mumtaz W, Xia L, Mohd Yasin MA et al. A wavelet-based technique to predict treatment outcome for major depressive disorder. PLoS One 2017; 12:e0171409.
150. Jaworska N, de la Salle S, Ibrahim MH et al. Leveraging machine learning approaches for predicting antidepressant treatment response using electroencephalography (EEG) and clinical data. Front Psychiatry 2018;9:768.
151. Khodayari-Rostamabad A, Reilly JP, Hasey GM et al. Using pre-treatment electroencephalography data to predict response to transcranial magnetic stimulation therapy for major depression. Annu Int Conf IEEE Eng Med Biol Soc 2011;2011:6418-21.
152. Bailey NW, Hoy KE, Rogasch NC et al. Differentiating responders and non-responders to rTMS treatment for depression after one week using resting EEG connectivity measures. J Affect Disord 2019;242:68-79.
153. Hasanzadeh F, Mohebbi M, Rostami R. Prediction of rTMS treatment response in major depressive disorder using machine learning techniques and nonlinear features of EEG signal. J Affect Disord 2019;256:132-42.
154. Zandvakili A, Philip NS, Jones SR et al. Use of machine learning in predicting clinical response to transcranial magnetic stimulation in comorbid posttraumatic stress disorder and major depression: a resting state electroencephalography study. J Affect Disord 2019;252:47-54.
155. Al-Kaysi AM, Al-Ani A, Loo CK et al. Predicting tDCS treatment outcomes of patients with major depressive disorder using automated EEG classification. J Affect Disord 2017;208:597-603.
156. Corlier J, Carpenter LL, Wilson AC et al. The relationship between individual alpha peak frequency and clinical outcome with repetitive transcranial magnetic stimulation (rTMS) treatment of major depressive disorder (MDD). Brain Stimul 2019;12:1572-8.

157. Woo CW, Chang LJ, Lindquist MA et al. Building better biomarkers: brain models in translational neuroimaging. Nat Neurosci 2017;20:365-77.
158. Costafreda SG, Khanna A, Mourao-Miranda J et al. Neural correlates of sad faces predict clinical remission to cognitive behavioural therapy in depression. Neuroreport 2009;20:637-41.
159. Goldstein-Piekarski AN, Korgaonkar MS, Green E et al. Human amygdala engagement moderated by early life stress exposure is a biobehavioral target for predicting recovery on antidepressants. Proc Natl Acad Sci USA 2016;113:11955-60.
160. Hahn T, Kircher T, Straube B et al. Predicting treatment response to cognitive behavioral therapy in panic disorder with agoraphobia by integrating local neural information. JAMA Psychiatry 2015;72:68-74.
161. Ball TM, Stein MB, Ramsawh HJ et al. Single-subject anxiety treatment outcome prediction using functional neuroimaging. Neuropsychopharmacology 2014;39:1254-61.
162. Tolmeijer E, Kumari V, Peters E et al. Using fMRI and machine learning to predict symptom improvement following cognitive behavioural therapy for psychosis. Neuroimage Clin 2018;20:1053-61.
163. Crane NA, Jenkins LM, Bhaumik R et al. Multidimensional prediction of treatment response to antidepressants with cognitive control and functional MRI. Brain 2017;140:472-86.
164. Nguyen KP, Fatt CC, Treacher A et al. Predicting response to the antidepressant bupropion using pretreatment fMRI. Predict Intell Med 2019;11843:53-62.
165. van Waarde JA, Scholte HS, van Oudheusden LJ et al. A functional MRI marker may predict the outcome of electroconvulsive therapy in severe and treatment-resistant depression. Mol Psychiatry 2015;20:609-14.
166. Whitfield-Gabrieli S, Ghosh SS, Nieto-Castanon A et al. Brain connectomics predict response to treatment in social anxiety disorder. Mol Psychiatry 2016;21:680-5.
167. Zhutovsky P, Thomas RM, Olff M et al. Individual prediction of psychotherapy outcome in posttraumatic stress disorder using neuroimaging data. Transl Psychiatry 2019;9:326.
168. Yuan M, Qiu C, Meng Y et al. Pre-treatment resting-state functional MR imaging predicts the long-term clinical outcome after short-term paroxetine treatment in post-traumatic stress disorder. Front Psychiatry 2018;9:532.
169. Cao B, Cho RY, Chen D et al. Treatment response prediction and individualized identification of first-episode drug-naïve schizophrenia using brain functional connectivity. Mol Psychiatry 2020;25:906-13.
170. Leaver AM, Wade B, Vasavada M et al. Fronto-temporal connectivity predicts ECT outcome in major depression. Front Psychiatry 2018;9:92.
171. Redlich R, Opel N, Grotegerd D et al. Prediction of individual response to electroconvulsive therapy via machine learning on structural magnetic resonance imaging data. JAMA Psychiatry 2016;73:557-64.
172. Cao B, Luo Q, Fu Y et al. Predicting individual responses to the electroconvulsive therapy with hippocampal subfield volumes in major depression disorder. Sci Rep 2018;8:5434-4.
173. Gong J, Cui LB, Xi YB et al. Predicting response to electroconvulsive therapy combined with antipsychotics in schizophrenia using multi-parametric magnetic resonance imaging. Schizophr Res 2020;216:262-71.
174. Germine L, Strong RW, Singh S et al. Toward dynamic phenotypes and the scalable measurement of human behavior. Neuropsychopharmacology 2021;46:209-16.
175. Etkin A, Patenaude B, Song YJ et al. A cognitive-emotional biomarker for predicting remission with antidepressant medications: a report from the iSPOT-D trial. Neuropsychopharmacology 2015;40:1332-42.
176. Rennie JP, Zhang M, Hawkins E et al. Mapping differential responses to cognitive training using machine learning. Dev Sci 2020;23:e12868.
177. Delgadillo J, Gonzalez Salas Duhne P. Targeted prescription of cognitive–behavioral therapy versus person-centered counseling for depression using a machine learning approach. J Consult Clin Psychol 2020;88:14-24.
178. Schwartz B, Cohen ZD, Rubel JA et al. Personalized treatment selection in routine care: integrating machine learning and statistical algorithms to recommend cognitive behavioral or psychodynamic therapy. Psychother Res 2021;31:33-51.
179. Delgadillo J, Ali S, Fleck K et al. StratCare: a pragmatic, multi-site, single-blind, cluster randomised controlled trial of stratified care for depression. Unpublished manuscript.
180. Lutz W, Deisenhofer AK, Rubel J et al. Prospective evaluation of a clinical decision support system in psychological therapy. Unpublished manuscript.
181. Browning M, Bilderbeck AC, Dias R et al. The clinical effectiveness of using a predictive algorithm to guide antidepressant treatment in primary care (PReDicT): an open-label , randomised controlled trial. Neuropsychopharmacology (in press).
182. Jussim L, Harber KD. Teacher expectations and self-fulfilling prophecies: knowns and unknowns, resolved and unresolved controversies. Personal Soc Psychol Rev 2005;9:131-55.
183. Patel SR, Bakken S, Ruland C. Recent advances in shared decision making for mental health. Curr Opin Psychiatry 2008;21:606-12.
184. Lundberg SM, Lee S-I. A Unified approach to interpreting model predictions. arXiv 2017;1705.07874v2.
185. Chekroud AM. Pragmatic, scalable, computational solutions to reduce the burden of major depression. ProQuest 2018;10907745.
186. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell 2019;1:206-15.
187. Salazar de Pablo G, Studerus E, Vaquerizo-Serrano J et al. Implementing precision psychiatry: a systematic review of individualized prediction models for clinical practice. Schizophr Bull (in press).
188. Jacobs M, Pradier MF, McCoy TH Jr et al. How machine-learning recommendations influence clinician treatment selections: the example of the antidepressant selection. Transl Psychiatry 2021;11:108.
189. Delgadillo J, Appleby S, Booth S et al. The Leeds Risk Index: field-test of a stratified psychological treatment selection algorithm. Psychother Psychosom 2020;89:189-90.
190. Müller VC. Ethics of artificial intelligence and robotics. https://plato.stanford.edu/entries/ethics-ai/.
191. Lo Piano S. Ethical principles in machine learning and artificial intelligence: cases from the field and possible ways forward. Humanit Soc Sci Commun 2020;7:9.
192. Mittelstadt BD, Floridi L. The ethics of big data: current and foreseeable issues in biomedical contexts. Sci Eng Ethics 2016;22:303-41.
193. Starke G, De Clercq E, Borgwardt S et al. Computing schizophrenia: ethical challenges for machine learning in psychiatry. Psychol Med (in press).
194. Trujillo AC, Gregory IM, Ackerman KA. Evolving relationship between humans and machines. IFAC-PapersOnLine 2019;51:366-71.
195. Endsley MR. Toward a theory of situation awareness in dynamic systems. Hum Factors 1995;37:32-64.
196. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med 2019;25:44-56.
197. Haque OS, Waytz A. Dehumanization in medicine: causes, solutions, and functions. Perspect Psychol Sci 2012;7:176-86.
198. Chen N-C, Drouhard M, Kocielnik R et al. Using machine learning to support qualitative coding in social science: shifting the focus to ambiguity. ACM Trans Interact Intell Syst 2018;8:1-20.
199. Stewart M. Towards a global definition of patient centred care: the patient should be the judge of patient centred care. BMJ 2001;322:444-5.
200. Robin R, Polanyi M. Personal knowledge. towards a post-critical philosophy. Philos Phenomenol Res 1960;20:429.
201. Bennett NL. Donald A. Schön, educating the reflective practitioner. San Francisco: Jossey-Bass, 1987.
202. Borrell-Carrió F, Suchman A, Epstein RM. The biopsychosocial model 25 years later: principles, practice, and scientific inquiry. Ann Fam Med 2004; 2:576-82.
203. Faden RR, Kass NE, Goodman SN et al. An ethics framework for a learning health care system: a departure from traditional research ethics and clinical ethics. Hastings Cent Rep 2013;43:S16-27.
204. Boff KR. Revolutions and shifting paradigms in human factors & ergonomics. Appl Ergon 2006;37:391-9.
205. Brangier E, Hammes-Adelé S. Beyond the technology acceptance model: elements to validate the human-technology symbiosis model. In: Robertson MM (eds). Ergonomics and health aspects of work with computers. Berlin: Springer, 2011:13-21.
206. Christophe G, Jean-Arthur M-F, Guillaume D. Comment on Starke et al.: ‘Computing schizophrenia: ethical challenges for machine learning in psychiatry’: from machine learning to student learning: pedagogical challenges for psychiatry. Psychol Med (in press).
207. Gerber A, Derckx P, Döppner DA et al. Conceptualization of the human-machine symbiosis – A literature review. Presented at the 53rd Hawaii International Conference on System Sciences, Maui, January 2020.
208. Floridi L, Taddeo M. What is data ethics? Philos Trans R Soc Math Phys Eng Sci 2016;374:2083.
209. Binns R. Fairness in machine learning: lessons from political philosophy. Proceedings of Machine Learning Research 2018;81:149-59.

210. Obermeyer Z, Powers B, Vogeli C et al. Dissecting racial bias in an algorithm used to manage the health of populations. Science 2019;366:447-53.
211. Yeung K, Lodge M (eds). Algorithmic regulation. Oxford: Oxford University Press, 2019.
212. Brownsword R, Scotford E, Yeung K (eds). The Oxford handbook of law, regulation and technology, Vol. 1. Oxford: Oxford University Press, 2016.
213. Jacucci G, Spagnolli A, Freeman J et al. Symbiotic interaction: a critical definition and comparison to other human-computer paradigms. Presented at the International Workshop on Symbiotic Interaction, Helsinki, October 2014.
214. Krebs P, Prochaska JO, Rossi JS. A meta-analysis of computer-tailored interventions for health behavior change. Prev Med 2010;51:214-21.
215. Lustria MLA, Noar SM, Cortese J et al. A meta-analysis of web-delivered tailored health behavior change interventions. J Health Commun 2013;18:1039-69.
216. Bird S, Kenthapadi K, Kiciman E et al. Fairness-aware machine learning: practical challenges and lessons learned. Presented at the 12th ACM International Conference on Web Search and Data Mining, Melbourne, February 2019.
217. Koutsouleris N, Dwyer DB, Degenhardt F et al. Multimodal machine learning workflows for prediction of psychosis in patients with clinical high-risk syndromes and recent-onset depression. JAMA Psychiatry 2021;78:195-209.

DOI:10.1002/wps.20882